PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Central Processes in Speech Understanding [and Discussion]

W. D. Marslen-Wilson, L. K. Tyler and R. B. Le Page

Phil. Trans. R. Soc. Lond. B 1981 **295**, 317-332
doi: 10.1098/rstb.1981.0143

Phil. Trans. R. Soc. Lond. B **295**, 317–332 (1981)
Printed in Great Britain

Central processes in speech understanding

By W. D. Marslen-Wilson and L. K. Tyler
Max-Planck-Institut für Psycholinguistik, Berg en Dalseweg 79, 6522 *BC Nijmegen, The Netherlands*

Considering the psychological mechanisms of language as functioning psycholinguistic processes, the fundamental question – from the perspective of speech comprehension – is how these processes operate over time to transform a transient sensory input into a meaningful utterance. The research reported here suggests that these processes are organized to allow optimally effective use of the information carried by the speech signal, as it becomes available over time. This is achieved by means of a central set of mental operations, which constitute the sequence of obligatory and automatic processes that the listener runs through in interpreting a normal utterance in its natural context. These automatic processes function to bring the speech signal into contact with the word-recognition domain as rapidly as possible. At this point its analysis can begin to interact with the structural and interpretative context in which it is occurring, which not only facilitates the word-recognition process, but also means that the interpretation of the message can begin immediately.

Introduction

The concept of a psychological mechanism of language has a straightforward interpretation for an experimental psycholinguist. It presupposes some psychologically real mental structure that is directly responsible for the perception or production of language by humans. The task of the psycholinguist or cognitive psychologist is not to discover how this structure is realized neurophysiologically in the brain, but rather to uncover its properties as a functioning psychological process. This further presupposes that the empirical basis for such an account must be data about how humans actually process language.

Under this process definition of a psycholinguistic mechanism, the various other disciplines concerned with human language, which do not take as their primary basis data about language processing, cannot provide the basic framework for a process theory of such a mechanism. These other disciplines are of course important in the full development of such a theory but the necessary first step is to develop a genuine psychological process framework from the perspective of which one can then interpret the research in other disciplines. This is the starting point for the discussion here of the psycholinguistic mechanisms underlying the perception and comprehension of spoken language.

A fundamental property of these psycholinguistic mechanisms, considered as psychological processes, is that they operate in real time upon an input – the speech signal – that is itself extended in time. Thus a first and basic question to be asked about these mechanisms is how, in fact, they do operate in time. What is the nature of the processes that go on in the mind of the listener, millisecond by millisecond, as he hears and interprets the speech input? All of the research that we shall describe here has been carried out from this perspective, and has chiefly used fast reaction-time tasks to do so.

The reason for using reaction-time tasks is not that reaction times are necessarily a more 'scientific' or 'empirical' source of data about mental processes than any other source, but because they provide the best technique for pinning down the precise temporal properties of speech understanding processes. The listener typically hears a speech input distributed over periods ranging from a few hundred milliseconds to several seconds. Over these relatively long time periods, the listener is performing some set of analyses of the input. To determine *what* types of analysis he performs *when*, one must have access to these processes as they are actually taking place. Information only about the final product of these processes is simply inadequate (see Levelt 1978).

The advantage of reaction-time tasks is that they tap the listener's representation of the input at a specific moment in time. Given the input available to the listener when he makes the response, it is then possible to infer what types of analysis he must have performed upon this input to produce the effects reflected in the response. The closer in time that this response is to the relevant stretches of the input – that is, the faster the reaction time – then the more closely one can specify the properties of the internal processes involved. In fact, the critical experimental evidence comes from situations in which the response time is of the order of 250 ms.

The further discussion will be divided into three sections. The first will discuss the on-line goals of speech understanding; that is, what are the immediate perceptual targets of the system as it processes the incoming speech stream. The second section will outline the ways in which the system is organized to achieve these goals. The third section will comment on some general implications of the view of speech understanding that has been presented.

Perceptual goals of speech understanding

It is clear enough, in general, what the goal of speech understanding is: to map from a speech signal onto some message level, or interpretative representation. This is defined, for present purposes, as the interpretation of an utterance relative to the listener's model of the current discourse and to his knowledge of the world. The important empirical question is how this eventual goal of the system is related to the immediate processing of the input. The first point that we shall establish is that this is not some distant goal, only reached after large chunks of the utterance have been heard. Rather, the message-level interpretation of the input is initiated right from the beginning of the utterance and directly affects all aspects of the system's operation. This is a quite different picture, with very different implications, from the views of language processing that have dominated much psycholinguistic thinking in the recent past. These alternative views have modelled language processing as a series of sequential stages of analysis, in which a message-level interpretation is only reached relatively late in the utterance (Fodor *et al.* 1974; Forster 1979). The on-line evidence that we shall be presenting is inconsistent with this general approach.

We shall begin with some general evidence that the speech input is continuously analysed by the listener with respect to its implications for a message-level interpretation. This evidence comes from an experiment that tracked the availability of different types of processing information over the course of an entire utterance (Marslen-Wilson & Tyler 1975, 1980*a*).

The experiment used three types of prose materials, of which a sample set is given in table 1. The 'normal prose' strings were normal utterances that could be both semantically and syntactically analysed. The second type, 'anomalous' strings, were approximately syntactically normal but had no coherent semantic interpretation. Thirdly, the 'scrambled' strings could be neither syntactically nor semantically analysed. In addition, to be able to observe the effects of the presence or absence of a discourse context, each type of string was either presented in isolation or preceded by a context sentence. Each test sentence also contained a monitoring target word – such as *lead* in the sample set given in table 1. These target words occurred at different serial positions across the test sentences, varying from the first to the ninth position.

The subjects' task was to listen to the test materials for a word target specified in advance, and to respond as rapidly as possible when they could detect it. We assumed here – correctly – that the different types of processing analysis that we were interested in (syntactic, semantic, discourse-based) would interact with the word-recognition processes upon which the detection of the target words depended. Thus, by measuring response time at different points across each type of test material, we could determine the time course with which syntactic and semantic processing information became available, and how this interacted with the availability or not of a discourse context. We could, in other words, begin to determine what types of analysis were being performed when.

Table 1. Sample stimulus set for word monitoring experiment

(The context sentences for each test-sentence are in parentheses. The monitoring target word for each test sentence is underlined.)

normal prose
(The church was broken into last week.)
Some thieves stole most of the lead off the roof.

anomalous prose
(The power was located in great water.)
No buns puzzle some in the lead off the text.

scrambled prose
(In was great power water the located.)
Some the no puzzle buns in lead text the off.

The main results are summarized in figure 1, which shows the relevant components of the reaction-time curves across word positions for each prose type, with and without the context sentence. The upper panel shows the results when a context sentence is present. Here normal prose sentences show a significant advantage of 50–60 ms over anomalous and scrambled strings right from the beginning of the test sentence. This means that the extra processing information that normal prose provides is being developed by the listener right from the beginning of the utterance. The critical point, established by comparing this result with the no-context condition in the lower panel, is that this early advantage of normal prose over the other conditions depends on the presence of a lead-in sentence. When no lead-in sentence is present, the extra facilitation due to the semantic interpretability of normal prose only develops later in the utterance.

This effect of the context sentence means that even the first words of an utterance must be being evaluated with respect to their discourse context. This strongly implies that the listener, at the point in the utterance where these effects occur, is attempting to construct a representation of the utterance that corresponds to a message-level interpretation. This is not to say that the input can always be fully interpretatively resolved as it is heard, but that this is the type of framework in terms of which the listener is trying to establish the on-line analysis of the input. Thus, for example, in the context 'The church was broken into last night', the phrase 'Some thieves...', at the beginning of the following sentence, can be immediately evaluated in terms of its pragmatic plausibility relative to the context set up by the previous sentence.

Another source of evidence comes from a quite different kind of experiment (Marslen-Wilson & Tyler 1980*b*; Tyler & Marslen-Wilson 1981*b*), which explicitly manipulated the type of linkage between an utterance and its discourse context. It shows that even when these linkages, early in the utterance, can only be based upon pragmatic inference, they still have immediate consequences for the on-line processing of the input.

The experiment used stimuli where two context – or 'scene-setting' – sentences were followed by one of three possible continuation fragments (see table 2). Each of these fragments contains an anaphoric device linking the fragment to the preceding discourse. In (1) the device is simply the repetition of the name of an antecedent individual; in (2) it is an unambiguous personal pronoun; while in (3), an example of 'zero' anaphora, there are no explicit lexical cues at all.

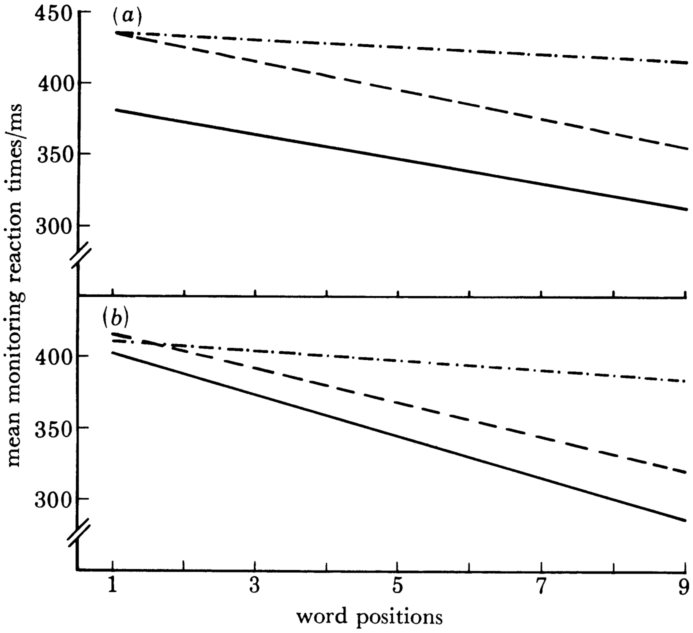

FIGURE 1. Mean word-monitoring latencies across word positions for normal prose (unbroken line), anomalous prose (broken line), and scrambled prose (dotted line). (*a*) Results when the context sentence is present; (*b*) when there is no context. The latencies plotted here combine the results for both Identical and Rhyme monitoring; for further details see Marslen-Wilson & Tyler (1980*a*).

In each case, to interpret the fragment, it is necessary to determine who is the agent of the action denoted by the verb, and to evaluate this with respect to the prior discourse. In (1) and (2) the agent is lexically specified within the fragment ('Philip', 'He'), and can be unambiguously related to possible antecedents in the discourse context just on the basis of this lexical information. However, (3) presents a quite different case. The only way agency can be assigned is on the basis of some inference process that relates the properties of the verb phrase 'Running towards...' to the properties of the potential antecedents in the discourse model. It is necessary to infer, given the scenario set up by the preceding context sentence, who is most likely to be running towards whom.

In the experiment, a visual probe word was presented to the subjects immediately at the end of each fragment, and their task was simply to name the probe as quickly as possible. In this paradigm, naming latency is slower when a word is inconsistent with its prior context. For the examples given in table 2, the probe word could be either *him* or *her*. In each case *her* is clearly a more appropriate continuation than *him*. The critical experimental question was, first, whether on-line preferences between probes could be obtained at all and, secondly, whether the size of such preferences differed across the three types of fragment. All three fragments are functionally equivalent in that they contain enough information so that, given the context, the listener can assign the correct agent for the verb. The question was whether and when the listener could exploit this functional equivalence during real-time speech processing.

Table 2. Sample stimulus set for anaphora experiment

context sentences

As Philip was walking back from the shop he saw an old woman trip and fall flat on her face.
She seemed to be unable to get up again.

continuation fragments

(1) Philip ran towards...
(2) He ran towards...
(3) Running towards...

Table 3. Results of anaphora experiment: mean naming latencies (milliseconds)

type of anaphor	appropriate	inappropriate	difference
repeated name	379	429	50
pronoun	385	434	49
zero	384	420	36

The results of the experiment, summarized in table 3, show naming latencies to appropriate and inappropriate probes following each of the three types of continuation. Latencies to inappropriate probes are slower than to appropriate probes, and the size of the difference does not vary significantly across the three types of continuation. This confirms, first, the immediate integration of the utterance with its context. Secondly, when this linkage can depend only on pragmatic inference, as in the zero anaphor case, this does not significantly impair or slow down the on-line integration process.

The speed of these integration processes is underlined by the outcome of another manipulation in the experiment. This was a variation in the length of the verb phrase in the continuation fragment, so that, in the zero anaphora case, the probe could occur anywhere from the second to the fifth word of the fragment, e.g. after 'Opening...' as opposed to after 'Carefully getting out of...'. This manipulation had no effect on the size of the difference between inappropriate and appropriate probes. This held even for the zero anaphor cases, where one might most expect an advantage of a longer delay from the beginning of the fragment to the appearance of the probe. For zero anaphors, the difference between probe types averaged 36 ms for the shortest verb phrases and 33 ms for the longest verb phrases.

In a related experiment (Tyler & Marslen-Wilson 1977) we found similar effects operating within a single sentence. Here we used structurally ambiguous fragments, such as 'shaking hands', which can either have a reading in which someone's hands are shaking, or a reading in which someone is shaking someone else's hands (e.g. as a greeting). In the experiment, these fragments were preceded by context clauses that biased one reading or the other of the fragment. Thus, for 'shaking hands' the two clauses would be:

(1) As a traditional way of gaining votes, shaking hands...
(2) If you're trying to thread a needle, shaking hands...

The subjects heard either one of the context clauses followed by the ambiguous fragment. As in the experiment just described, an appropriate or inappropriate visual probe was flashed up at the end of the fragment.

The two probes were either *is* or *are*, and whether they were appropriate or not depended on the preceding context. The probe *is* would be more appropriate than *are* following (1), and vice versa following (2). The result was again a significantly faster naming latency to appropriate probes. These on-line preferences can only be explained if we assume that the listener is rapidly evaluating the structural readings of the ambiguous fragments relative both to the meanings of the words involved and to the pragmatic plausibility of each reading in the given context.

In summary, these and a variety of other experiments – including research on children as young as 5 years old (Tyler 1981; Tyler & Marslen-Wilson 1981*a*) – all converge on the same conclusion about the on-line goal of speech understanding: namely, that the goal of the system is to achieve a message-level interpretation of the signal as rapidly and as early as possible. It is now necessary to show how the central processes of speech understanding are organized so as to allow the system to function in this way.

Central processes in speech understanding

We need to begin by severely restricting what is classed as a 'central process'. We define these as the set of automatic and obligatory mental processes that are triggered when a speech input is heard, and which carry the analysis through to its message-level interpretation. These *on-line* processes are not open to conscious awareness or conscious control, and form the basis for the normal, effortless comprehension of an utterance in context by a normal adult listener. Once these automatic processes have run through to completion, or else fail in some way, then the products of these processes become available for all sorts of 'off-line' analyses. These later analyses are idiosyncratic and variable, and not, we believe, central to the normal process of speech understanding.

The primary characteristic of the central processes, as reflected in several experiments, is that they move the analysis of the speech signal as rapidly as possible into a domain where all available sources of information can be brought to bear on its further analysis and interpretation. The essential link here is the word-recognition system. It is only when the signal has been brought into contact with the word-recognition system that its analysis can begin to interact with the structural and interpretative context in which it is occurring. This is because the mental representations of words contain syntactic and semantic information that can be assessed for their compatibility with the context. This makes it possible for contextual constraints to speed up the selection of the correct word candidate, while at the same time allowing the rapid integration of the structural and semantic implications of this word into the developing interpretation of the utterance.

Experiments on spoken word recognition, both in and out of context, show not only that word recognition is a remarkably fast and early interactive process, but also that it appears to be optimally efficient in the use it makes of the speech signal as it accumulates over time. The speed and earliness of word recognition is illustrated in a number of experiments showing that words in normal contexts are typically recognized within about 200 ms of their onset, and usually well before all the word has been heard.

One such experiment used the speech shadowing task, in which the subject repeats back continuous speech as he hears it. Some subjects were able to do this at repetition delays averaging between 250–275 ms, the delay being measured between the onset of a word in the material they are hearing and the onset of the same word in their repetition (Marslen-Wilson 1973). The distribution of latencies in figure 1, showing the performance of some typical fast shadowers, demonstrates the consistency with which they manage to track the input at these short latencies (with error-rates averaging less than 5%). Other experiments show that these fast shadowers understand the material as they repeat it; that they are not just 'echoing' it on the basis of some minimal acoustic–phonetic analysis (Marslen-Wilson 1975, 1976). This type of performance is a direct demonstration of the speed of on-line speech understanding.

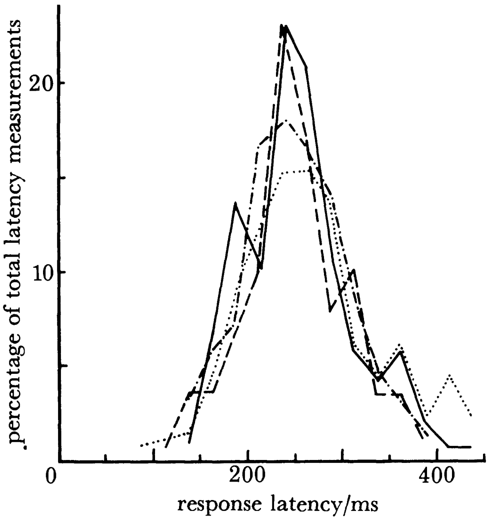

FIGURE 2. Distribution of shadowing latencies for four fast shadowers. The distributions are based on 150 measurements for each subject, taken from two 300-word passages of narrative prose.

A response latency of 250 ms implies, furthermore, that the word has been recognized even earlier than this. We routinely allow 50–75 ms for the time taken to make the actual response – in this case, to say the word. Subtracting this estimate from the overall repetition latencies gives an estimated word-recognition time of 200 ms or less.

The word-monitoring task mentioned earlier provides comparable estimates of recognition time (Marslen-Wilson & Tyler 1980*a*). Mean detection latencies in the fastest monitoring task were 273 ms. Allowing 50–75 ms for response execution, this again gives us recognition within about 200 ms of word onset. Very similar estimates have been obtained in other quite different tasks (Grosjean 1980; Morton & Long 1976). Furthermore, when the total durations of the words are measured, as in the monitoring experiment or in Grosjean's research, these measurements show that words in context are being accurately responded to when little more than half of the word could have been heard. The average duration of the target words in our experiment was 375 ms and in Grosjean's experiment 410 ms.

The critical inference here is that words are being recognized before the listener could have determined what the word was just on the basis of the available acoustic–phonetic information. The initial 200 ms of a word would normally only be enough for the listener to hear the first one or two phonemes of a word: the consonant–vowel or vowel–consonant sequence with which it begins.

It is possible to roughly quantify how insufficient this amount of acoustic–phonetic information would be. The word-monitoring studies mentioned above involved 81 different word-targets. For each word we determined how many words in the language (American English) would be compatible with its first two phonemes. The median number of possibilities that we estimated, ignoring very infrequent words, was 29. There were, for example, about 35 words beginning with the sequence /ve/.

These estimates clearly indicate that there would be, on average, several possible word candidates compatible with the sensory information that would be available when the correct word was successfully recognized. This implies that it must be contextual information that provides the additional necessary constraints to allow the listener to choose the correct word. There is good evidence to support this. In particular, the fast and early recognition times in all the tasks we have mentioned depend on the availability of a normal utterance context. In the shadowing experiment, response latencies increased by 60 ms when subjects were shadowing semantically anomalous prose materials. In the monitoring experiment, as we saw earlier, latencies also increased by about 60 ms for targets in anomalous context, and by a further 35 ms in scrambled strings. In Grosjean's and Morton & Long's experiments, estimated recognition time also increases systematically as the constraints provided by the word's context decrease.

If spoken words can, therefore, be recognized when many words are still compatible with the available sensory input, and if contextual variables do help to select from among this initial set of word candidates, then this has strong implications for how word recognition must be organized. It requires a 'distributed', or parallel, processing system, in which all of the initial set of word candidates at the beginning of a word – which we shall refer to as the *word-initial cohort* – can be simultaneously assessed for their compatibility with the available context.

This leads to a processing system made up of a large array of computationally active recognition devices, with, possibly, each device corresponding to a separate word (or morpheme) in the mental lexicon. These recognition devices are assumed to become active whenever the sensory input matches the acoustic–phonetic pattern specified for each device. Thus, when the initial segments of a word are heard, the recognition devices corresponding to all of the words in the language (known to the listener) that begin with this initial sequence will become active.

This subset of active elements, constituting the word-initial cohort, would be sensitive both to the continuing sensory input and to the compatibility of the words that they represent with the available structural and interpretative context. A mismatch with either source of constraint would cause the elements in question to drop out of the pool of potential word candidates. This means that there will be a sequential reduction over time in the initial set of candidates, until only one candidate is left. We assume that at this point the correct word can be recognized. This type of recognition process not only enables words to be recognized as they are heard, but also as soon as they securely can be.

To give a more concrete example of how such a process would work, consider the analysis of the word *stand*, heard in the context given in table 4. Represented in the table is a rough estimate of the word-initial cohort for the word *stand*; that is, the set of words, likely to be known to the listener, that begin with the sound-sequence /stæ/. Note that, for present purposes, we are ignoring possible co-articulatory effects. The vowel /æ/ in the word-initial cohort for /stæ/ is certainly not acoustically identical in each case. The properties of a vowel differ depending on the phonetic environment in which it occurs, so that, for example, the /æ/ in *stand*, followed by a nasal consonant, is articulated somewhat differently from the /æ/ in a word where the following consonant is not nasalized. If the processing system can use this type of information in word-recognition processes – and it is not clear whether it can or not – then the size of the word-initial cohort would be somewhat smaller. Even so, this type of information could not allow the listener to discriminate in advance between *stand*, *stanza*, and all the other words where /stæ/ is followed by a nasal.

Table 4. Sample word-initial cohort in context

John was trying to get some bottles down from the top shelf. To reach them he had to /stæ.../

stag	stagger	stack
stalactite	stagnate	stand
stamina	stammer	
stance	stamp	
standard	stampede	
standoffish	stab	
static		
statistic		
statue		
stature		
statute		
stanza		

Assuming, then, a word-initial cohort of the range specified in table 4, it is clear that if sensory information only is used to make the word-recognition decision, then the word *stand* could not be identified until all of it had been heard – at which point, for example, it could be discriminated from *standard*. If syntactic constraints could be used in the decision process, then the set of words in the first column on the table could be immediately excluded, since they could not occur at that point in the utterance. This would allow the word *stand* to be discriminated as soon as the consonant following /æ/ could be identified as an /n/, rather than an /m/ (as in the word *stamp*).

Finally, if interpretative constraints also contribute to on-line word recognition, then the word could be identified even earlier. None of the words in the middle column are plausible in the discourse context provided. The two remaining candidates, *stand* and *stack*, are the only members of the cohort for /stæ/ that both match the sensory input and fully satisfy contextual constraints. Thus as soon as the listener can determine that the next sound is a nasal, he can select the word *stand*. This would, in a recognition task, give a reaction time of 300 ms or less.

This kind of on-line integration of the implications of the sensory input with the requirements of the utterance and discourse context is what one is forced to postulate to explain the speed and earliness of spoken-word recognition. A system with these general properties, therefore, allows a word to be recognized at that point, starting from the beginning of the word, at which the word in question becomes uniquely distinguishable from all of the other words in the language, beginning with the same sound sequence, that are also compatible with the available context. Such a system can be said to be *optimally efficient* in the use it makes of the sensory input as it accumulates over time. We have investigated this claim directly, in experiments using isolated words, which provide the clearest test case.

The main experiment here used an auditory lexical decision task, in which subjects heard isolated sound sequences and made non-word decisions (pressed a response key) whenever they thought they heard a sound sequence which did not form a word in English. The important variable here was the point in each non-word sequence at which it became a non-word, and the way in which these 'non-word points' were determined.

The calculation of the non-word point for each stimulus was based on the cohort structure of the language (General British English). For example, one of the non-words was *stadate*. As we can see from the list of words in table 4, this becomes a non-word at the /d/, since there are no words in the word-initial cohort for /stæ/ that have the continuation /d/. The non-word points, determined in this way, varied in position across the stimuli from the second to the fifth phoneme in a sequence. A sample set of stimuli are given in table 5.

Table 5. Sample non-word sequences

(Last 'real-word' phoneme is underlined.)

v̲li:sɪdəns	v̲leesidence
sn̲edɪst	sn̲edist
le̲ʃtɪnɪk	le̲shtinik
stæ̲deɪt	sta̲date
græn̲kɪmənt	gran̲kiment
swəʊl̲aɪt	swol̲l̲ite

If speech processing is indeed optimally efficient in the way we have proposed, then the non-word discrimination decision can begin to be made at exactly that point where the sound-sequence diverges from the existing possibilities in English – that is, from the offset of the last phoneme in the non-word sequence that could be part of the beginning of a real word in English (underlined in the table). This predicts that non-word decision time, relative to critical phoneme offset, should remain the same independent of where this phoneme (the last 'real-word' phoneme) comes in the sequence. This prediction was borne out by the results, since decision time from the offset of the critical phoneme did remain constant, at about 450 ms, across experimental conditions.

Figure 3 gives a summary of the results. It shows the close dependence between overall reaction time (measured from sequence onset) and the delay from sequence onset until the offset of the critical phoneme. In an optimal system, the slope of this graph should approach 1.0 (broken line). The outcome is very close to this; the observed slope is +0.88 ($r = +0.98$). The second set of points in the figure (open circles) belong to a different experimental group, and show a similar slope, but with the curve displaced by about 70 ms (for further details, see Marslen-Wilson 1980).

These results, and related experiments on the recognition of real words (Marslen-Wilson 1980), show that spoken-word processing is based upon a highly efficient real-time analysis of the speech signal. In the non-word detection task, this allows a non-word to be discriminated as soon as it diverges from the existing possibilities in the language. In the normal recognition of real words, this allows a word to be discriminated as soon as it diverges from other possible words in the language. We hypothesize that this optimal processing property – the ability to establish an analysis of the input as soon as the minimum necessary information becomes available – applies to all aspects of the speech understanding process.

This view of human speech processing, as an optimal and interactive process, makes it essential to control and regulate these processes in some way. This can be achieved on the basis of two complementary processing principles. First, that perceptual processing is *data-driven*. This means that the initial set of possible analyses of the input is determined from the bottom up. Thus, in word recognition, the initial set of possible word candidates is strictly determined by the properties of the sensory input. This word-initial cohort delimits the decision space within which further analyses, incorporating contextual factors, can take place.

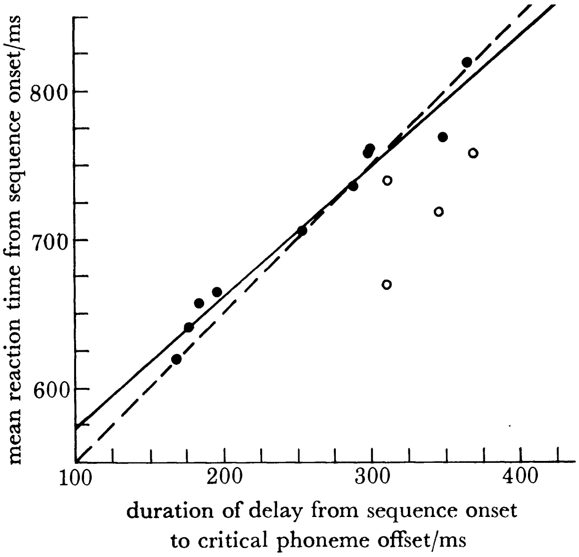

FIGURE 3. Mean reaction time in auditory lexical decision as a function of the delay from the onset of the non-word sequence to the point in the sequence at which it becomes a non-word.

The second principle is that speech processing operations are *obligatory* in character. Given a bottom-up input, the system must run through its characteristic operations on this input. Again, there are clear instances of this in spoken-word recognition. If the speech input can be lexically interpreted, then it apparently must be. Apart from one's own phenomenal experience, the evidence for this comes from several studies that show that, even when subjects are asked to focus their attention on the acoustic–phonetic properties of the input, they do not seem to be able to avoid identifying the words involved (see Marslen-Wilson & Tyler 1980*a*). This implies that the kind of processing operations observable in spoken-word recognition are mediated by automatic processes, which are obligatorily applied to any acoustic–phonetic input. The same property holds for operations throughout the system.

These complementary principles, of bottom-up and obligatory processing, first of all force the on-line analysis process to stay in contact with the sensory input. The crucial link between input and interpretation is the word-recognition process, and its operations are strictly constrained by its sensory input. The scope of the structural and interpretative analysis of the input is determined by the words that are recognized, and the principle of bottom-up priority protects this recognition process from dominance by these contextual sources.

This means that the system cannot allow, in normal first-pass processing, for any direct top-down effects – that is, for direct effects of contextual expectations on lower-level processes. The term 'interactive' must therefore be interpreted here in a different way from that in the standard psychological literature, where it specifically entails these top-down effects. In the present system contextual criteria apply only to the set of candidate analyses specified from the bottom up. Given the simultaneous assessment processes permitted by a distributed processing model, this restriction on the system does not prevent the on-line cooperation of sensory and contextual sources of information. In word recognition, therefore, the word sense that is selected is the one that best maps onto the available utterance and discourse representation, and, as we have seen, this normally appears to be achieved within about 200 ms of word onset. This gives the appearance of top-down interactions with word-recognition processes. But this is illusory, since what is really happening is not so much the recognition of *words*, as the incorporation of the correct word sense, with its structural consequences, into the interpretative representation of the utterance.

The second benefit of adopting these bottom-up processing principles is that they account for the kinds of processing phenomena that are otherwise assigned to autonomous processing modules, where 'autonomous' means that the operations of the module cannot be affected by inputs from components further along in the analysis sequence. It is, for example, often argued that word recognition is an autonomous, modular process because contextual constraints do not prevent contextually inappropriate word senses from being activated. This effect is clear from experiments (Swinney 1979; Tanenhaus *et al.* 1979) showing, for example, that the 'river' meaning of *bank* is momentarily activated even when the word occurs in a very strong biasing context – for example, of someone going to cash a cheque. The existence of such effects is of course exactly required by the principles of obligatory bottom-up processing. Note that the failure of context to suppress the contextually non-preferred reading of a word does *not* mean that the selection of the contextually preferred reading was itself not affected by contextual variables.

The modular approach is forced to account for such effects by drawing an analytic line, as it were, around the phenomenon in question, and assigning it to an independent processing module. The theoretical cost of doing this is that it carries with it strong claims about distinct levels of representation in the processing system. Each new module must presuppose a new level of computational representation. The advantage of the approach we are suggesting is that one is not forced to proliferate components in this way. One may still want to postulate distinct levels of representation in the system, but one is free to show more discretion in doing so.

The third consequence of obligatory bottom-up processing is that it ensures that the analysis of the input will always be developed as far as it can be, and as rapidly as it can be. As soon as the appropriate bottom-up input becomes available to the word-recognition system, then some subset of recognition elements will be activated. This will make information available about the syntactic and semantic properties of these word candidates, which will immediately trigger attempted structural and interpretative analyses. To the extent that a particular input leads to acceptable further analyses, then these analyses must propagate through the system and establish consequences at the interpretative level.

These various implications of bottom-up obligatory processing fit in very well with the requirements of an interactive system with the temporal parameters we have described. The result is a model of human speech understanding in which the speech input can be securely interpreted as early as the properties of the signal permit. Given that the speech signal is necessarily extended in time, it seems desirable that the speech-recognition system should be designed to minimize the delay with which the significance of this signal can begin to be made available to the perceiver.

Implications for psycholinguistic theory

In the previous two sections we have outlined the view of speech understanding that develops when one pays close attention to the real-time properties of the process. What general implications does this have for a psychological process model of language?

The central issue is the role of linguistic structure in the perceptual process. A model of speech understanding will be completely inadequate unless the role of linguistic structure is somehow realized in the model. It is none the less quite unclear how this should be done. The real problem here has been the considerable uncertainty, not to say confusion, about how abstract a theory of linguistic knowledge a theory of grammar really is. Psycholinguists have typically tended to assume that linguistic theories are really not very abstract at all. Historically, this attitude clearly reflects the fact that the chief intellectual impetus for the founding of modern psycholinguistics was the development of modern syntactic theory – that is, of the transformational grammar of the late 1950s and early 1960s (see Fodor *et al.* 1974; Levelt 1974; Marslen-Wilson 1976). But this type of syntactic theory is of course not *per se* a processing theory, and linguists have generally attempted to make this clear.

For a psycholinguist, however, the precise nature of the implications of the syntactic theory for the processing theory has been, and still is, an empirical question of the greatest importance (see Tyler 1980). One can, in fact, distinguish three types of increasingly distant relationships between the syntactic theory and the processing theory. It is only with respect to data about how language processes actually function that the appropriate type of relation can be decided upon. Note, however, that the following remarks are based on the standard conception of a syntactic theory as a formal device that generates the well-formed strings in a language. It is not clear to us how these issues can be related to linguistic theory construed as universal grammar, in the sense of Chomsky (1980).

The closest relationship between a syntactic theory and a process model is one in which the grammar is literally taken as a processing theory, so that the operations carried out in the grammar are taken to be essentially identical to those in the processing mechanism. This was the position taken by the earliest psycholinguists working in a transformationalist environment, and is not a view that is widely held any more (we do not include here the approach recently taken by Bresnan and colleagues (see Bresnan 1978), which differs in important respects from the position stated above).

Psycholinguists instead moved to a more distant relationship between the syntactic theory and the processing model, by postulating a language processing system that contained an autonomous syntactic processor. This processor operated upon the output of the word-recognition system to produce a syntactic structural representation of the input, which could then be semantically interpreted. But, while the processing theory still maintained the linguistic concept of a syntactic level of representation, the connection with the grammar became sufficiently weakened that the grammar no longer specified the actual sequence of syntactic processing operations within the processor. The eventual outcome of this was that the only testable claim that was really being made was that there was a separate syntactic processor, and that, whatever its internal structure might be, it was autonomous. This claim can be directly tested by determining whether the language processing system does indeed behave as if it contained an autonomous processing module operating at the appropriate level of analysis of the input. Considering the on-line evidence from this perspective, it provides no evidence of this sort.

First, one does not find the types of distinctive psychological effects that are normally needed to motivate independent processing stages. Secondly, there is evidence for interactions between processing stages of a type that are expressly prohibited within an autonomous processing system. Thirdly, one does not find the kind of sequential ordering of different types of processing information that is predicted by an autonomous processing theory. There is no evidence that syntactic analyses are performed first, and are only then followed by a semantic and interpretative analysis (Marslen-Wilson & Tyler 1980*a*).

The postulated syntactic processor could of course be modified so that, for example, it did communicate continuously with other processing components, and so became consistent with the observed experimental data. But if one weakens the autonomy of a syntactic processor in this way, then the claim for a syntactic module becomes empirically empty. The processing system would now behave in exactly the same way, in processing experiments, whether there was an independent syntactic processor or not. To the extent that the claim for an autonomous syntactic component is made empirically distinct, then this claim is disconfirmed by the available data.

This means that the on-line processing evidence is completely compatible with a still more distant relationship between the syntactic theory and the processing theory. Exactly what this relation should be cannot be specified at the moment, given the evident incompleteness of the available processing theories. But one possible analogy is suggested by Marr & Nishihara (1978), in discussing the relation between two different algorithms for computing the same abstract function. Marr & Nishihara give as an example the choice between different algorithms for computing the Fourier transform, where one can contrast the fast Fourier transform, which is a sequence of mathematical operations, with the so-called optical method, based on the mechanisms of laser optics. There is clearly no direct relation between the two procedures, except that they produce the same final result.

In this vein, then, the relation between a syntactic theory and a processing theory could be as distant as, for example, that between a mathematical characterization of the transfer function of the vocal tract and the actual physical mechanisms that produce speech. From this perspective, it would be as fruitless an enterprise to look for a syntactic module in the language processing system as it would be to look for a module in the vocal tract within which was explicitly represented the mathematical transfer function of the system. In other words, the concept of an autonomous syntax, as represented in a linguistic theory, could no longer have any direct interpretation within a psychological processing model.

The assignment of this degree of distance to the relation between a syntactic theory and a processing theory may nevertheless be the most fruitful strategy for a psycholinguist to adopt. First, it rescues him (or her) from the ambiguities and paradoxes inherent in adopting the first or second types of relation that we discussed. Secondly, it makes it clear how a syntactic theory can, on the one hand, be a genuine theory of a psychological object, but, on the other hand, not make any obviously verifiable claims about how linguistic knowledge is used in some actual process. Thirdly, and perhaps most significantly, it enables the psycholinguist to start asking potentially more appropriate questions about how linguistic structure is realized in the operations of the processing system.

A final, and related, point concerns the general computational assumptions that constrain the possible organization of the language processing system. The analysis that we have presented of the speech-understanding system, as an optimally effective processing device, is only feasible given particular computational assumptions about the processing system. The phenomena that the model is designed to explain seem to reflect a very rapid real-time solution to a complex set-intersection problem. Given the set of properties defined by the current discourse and utterance context, the listener appears to be able to select, from among the potentially large number of words initially activated, that word whose syntactic and semantic properties best match those of the context set. It is almost certain that this could not be done, within the observed temporal constraints, by a serial machine – let alone one made up of slow elements like neurons (see Fahlman 1979). Only some form of parallel machine could show the required insensitivity to the size of the search space within which the on-line set-intersection problem has to be solved.

If these implications for the general properties of the neural machine are taken seriously, then this places quite different constraints on the way one can think about the psychological mechanisms of language than if one assumes a computational process of the more familiar serial type.

References (Marslen-Wilson & Tyler)

Bresnan, J. 1978 A realistic transformational grammar. In *Linguistic theory and psychological reality* (ed. M. Halle, J. Bresnan & G. A. Miller). Cambridge, Massachusetts: M.I.T. Press.

Chomsky, N. 1980 Rules and representations. *Behav. Brain Sci.* **3**, 1–61.

Fahlman, S. E. 1979 *NETL: a system for representing and using real-world knowledge.* Cambridge, Massachusetts: M.I.T. Press.

Fodor, J. A., Bever, T. G. & Garrett, M. F. 1974 *The psychology of language.* New York: McGraw-Hill.

Forster, K. 1979 Levels of processing and the structure of the language processor. In *Sentence processing: psycholinguistic studies presented to Merrill Garrett* (ed. W. E. Cooper & E. C. T. Walker). Hillsdale, New Jersey: L.E.A.

Grosjean, F. 1980 Spoken word recognition processes and the gating paradigm. *Percept. Psychophys.* **28**, 267–283.

Levelt, W. J. M. 1974 *Formal grammars in linguistics and psycholinguistics.* 3 vols. The Hague: Mouton.

Levelt, W. J. M. 1978 A survey of studies in sentence perception: 1970–1976. In *Studies in the perception of language* (ed. W. J. M. Levelt & G. B. Flores d'Arcais). New York: Wiley.

Marr, D. & Nishihara, H. K. 1978 Visual information processing: artificial intelligence and the sensorium of sight. *Technol. Rev.* **81**, 2–23.

Marslen-Wilson, W. D. 1973 Linguistic structure and speech shadowing at very short latencies. *Nature, Lond.* **244**, 522–523.

Marslen-Wilson, W. D. 1975 Sentence perception as an interactive parallel process. *Science, N.Y.* **189**, 226–228.

Marslen-Wilson, W. D. 1976 Linguistic descriptions and psychological assumptions in the study of sentence perception. In *New approaches to the study of language* (ed. R. J. Wales & E. C. T. Walker). Amsterdam: North-Holland.

Marslen-Wilson, W. D. 1980 Speech understanding as a psychological process. In *Spoken language generation and recognition* (ed. J. C. Simon). Dordrecht: Reidel.

Marslen-Wilson, W. D. & Tyler, L. K. 1975 Processing structure of sentence perception. *Nature, Lond.* **257**, 784–786.

Marslen-Wilson, W. D. & Tyler, L. K. 1980*a* The temporal structure of spoken language understanding. *Cognition* **8**, 1–71.

Marslen-Wilson, W. D. & Tyler, L. K. 1980*b* Towards a psychological basis for a theory of anaphora. In *Papers from the parasession on pronouns and anaphora* (ed. J. Kreiman & A. Ojeda). Chicago: Chicago Linguistic Society.

Morton, J. & Long, T. 1976 Effect of word transitional probability on phoneme identification. *J. verb. Learn. verb. Behav.* **15**, 43–51.

Swinney, D. A. 1979 Lexical access during sentence comprehension: (re)consideration of context effects. *J. verb. Learn. Verb. Behav.* **18**, 545–569.

Tanenhaus, M. K., Leiman, J. M. & Seidenberg, M. S. 1979 Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *J. verb. Learn. verb. Behav.* **18**, 427–440.

Tyler, L. K. 1980 Serial and interactive approaches to on-line sentence processing. *Theor. Ling.* **7**. (In the press.)

Tyler, L. K. 1981 Syntactic and interpretative factors in the development of language comprehension. In *The child's construction of language* (ed. W. Deutsch). London: Academic Press.

Tyler, L. K. & Marslen-Wilson, W. D. 1977 The on-line effects of semantic context on syntactic processing. *J. verb. Learn. verb. Behav.* **16**, 683–692.

Tyler, L. K. & Marslen-Wilson, W. D. 1981*a* Children's processing of spoken language. *J. verb. Learn. verb. Behav.* **20**. (In the press.)

Tyler, L. K. & Marslen-Wilson, W. D. 1981*b* Processing utterances in discourse contexts: on-line resolution of anaphors. (Submitted for publication.)

Discussion

R. B. Le Page (*Department of Language, University of York, U.K.*). Apart from the fact that the precise *phonetic* information in 'the first few phonemes' of utterances of *stand* is unlikely to be identical with that for *any* of the other words cited (i.e. the phonetic value of /æ/ in *stand* is quite different from that in *stamp*, etc.) I should like to know exactly what the authors mean by the subject *hearing* the word. We make our first-year students listen to Creole tapes and write down what they think they can hear. Because some of the words resemble English words they tend to interpret the acoustic data in terms of its resemblance to or difference from their own English. Once they have been introduced to Creole, and its idealized morphology, they can 'learn' to hear, for example, [wã] as a future marker where previously they heard [wi], and cannot subsequently understand how they heard [wi]. All systematic descriptions of language are subjective, and all hearing is perceptually subjective also; the experiments seem, however, to have been carried out on the assumption that the listeners were matching objectively heard 'segments' against some objective phonotactic system of which they were components.

W. D. Marslen-Wilson and L. K. Tyler. We agree with Professor Le Page that the realization of a given phoneme varies according to the context in which it occurs. He may have missed our explicit statement in the lecture that the sample cohort for /stæ/ was for illustrative purposes only, and indeed glosses over possible co-articulatory effects. As for the second point, we regret that Professor Le Page should have been led to believe that our experiments presuppose some form of 'objective' linguistic sense datum. It is of course well known that the interpretation of a given acoustic signal varies according to the framework in terms of which one can interpret it, and a major goal of our research has been precisely to investigate the properties of this interaction between the analysis of the sensory input and the contextual framework within which the input is being lexically interpreted. Our research, then, is carried out on the opposite sort of assumption to the one that Professor Le Page assigns to us, and is quite compatible with his observations on Creole. We also believe, however, that the properties of the signal do nevertheless matter for the listener, and in systematic ways, which is what may have misled Professor Le Page.