Cognition, 25 (1987) 71–102

Functional parallelism in spoken word-recognition

WILLIAM D. MARSLEN-WILSON*

Max-Planck-Institut für Psycholinguistik, Nijmegen, and MRC Applied Psychology Unit, Cambridge

Abstract

The process of spoken word-recognition breaks down into three basic functions, of access, selection and integration. Access concerns the mapping of the speech input onto the representations of lexical form, selection concerns the discrimination of the best-fitting match to this input, and integration covers the mapping of syntactic and semantic information at the lexical level onto higher levels of processing. This paper describes two versions of a "cohort"-based model of these processes, showing how it evolves from a partially interactive model, where access is strictly autonomous but selection is subject to top-down control, to a fully bottom-up model, where context plays no role in the processes of form-based access and selection. Context operates instead at the interface between higher-level representations and information generated on-line about the syntactic and semantic properties of members of the cohort. The new model retains intact the fundamental characteristics of a cohort-based word-recognition process. It embodies the concepts of multiple access and multiple assessment, allowing a maximally efficient recognition process, based on the principle of the contingency of perceptual choice.

1. Introduction

To understand spoken language is to relate sound to meaning. At the core of this process is the recognition of spoken words, since it is the knowledge representations in the mental lexicon that provide the actual bridge between sounds and meanings, linking the phonological properties of specific word-

*I thank Uli Frauenfelder and Lorraine Tyler for their forbearance as editors, and for their comments on the manuscript. I also thank Tom Bever and two anonymous reviewers for their stimulating criticism of previous drafts. The first version of this paper was written with the support of the Department of Experimental Psychology, University of Cambridge, which I gratefully acknowledge. Reprint requests should be sent to William Marslen-Wilson, MPI für Psycholinguistik, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands.

forms to their syntactic and semantic attributes. This duality of lexical representation enables the word-recognition process to mediate between two radically distinct computational domains—the acoustic-phonetic analysis of the incoming speech signal, and the syntactic and semantic interpretation of the message being communicated. In this paper, I am concerned with the consequences of this duality of representation and of function for the organisation of the word-recognition process as an information-processing system.

The overall process of spoken word-recognition breaks down into three fundamental functions. These I will refer to as the *access*, the *selection*, and the *integration* functions. The first of these, the access function, concerns the relationship of the recognition process to the sensory input. The system must provide the basis for a mapping of the speech signal onto the representations of word-forms in the mental lexicon. Assuming some sort of acoustic-phonetic analysis of the speech input, it is a representation of the input in these terms that is projected onto the mental lexicon.

The integration function, conversely, concerns the relationship of the recognition process to the higher-level representation of the utterance. In order to complete the recognition process, the system must provide the basis for the integration, into this higher level of representation, of the syntactic and semantic information associated with the word that is being recognised.

Finally, and mediating between access and integration, there is the selection function. In addition to accessing word-forms from the sensory input, the system must also discriminate between them, selecting the word-form that best matches the available input.

These three functional requirements have to be realised in some way in any model of spoken word-recognition. They need to be translated into claims about the kinds of processes that subserve these functions, and about the processing relations between them during the recognition of a word. I will begin the discussion here by considering the way that the access and selection functions are realised, and their relationship to the integration function. How far do access, selection, and integration correspond to separate processing stages in the recognition of a spoken word, and to what extent do they operate in computational isolation from one another?

I will develop the argument here in its approximate historical sequence. In Section 2 I will argue that, while the accessing of the mental lexicon is a strictly autonomous, bottom-up process, there seems to be a close computational dependency between the process of selecting the word-form that best matches the sensory input and the process of integrating the syntactic and semantic properties of word-forms with their utterance context. The characteristics of the real-time transfer function of the system suggest that the selection phase of the recognition process cannot depend on bottom-up information alone, and that contextual constraints also affect its outcome. This, as I will show in Section 3, led to the first version of the cohort model: a parallel, interactive model of spoken word-recognition. In Section 4 I will examine the properties and predictions of this early model. In Section 5 I will show how this model now needs to be modified. In particular, I will argue that it needs to incorporate the concept of activation, and I will re-examine the role of top-down interaction in the on-line recognition process, suggesting a model where different information sources are integrated together to give the perceptual output of the system, but where they do not, in the conventional sense, interact. In particular, I argue for the autonomy of form-based selection, as well as for the autonomy of form-based access.

2. The earliness of spoken word-recognition

The crucial constraint on the functional properties of access and selection is the *earliness* of correct selection. This I define as the reliable identification of spoken words, in utterance contexts, *before* sufficient acoustic-phonetic information has become available to allow correct identification on that basis alone. If this can be demonstrated, then it places strong restrictions not only on how the selection process is organised, but also on the ways in which representations are initially accessed from the bottom-up.

To prove early selection, two things must be established. The first is how long it takes to recognise a given word. This reflects the timing with which the selection function is completed. The second is whether the acoustic-phonetic information available at this estimated selection-point is or is not sufficient, by itself, to support correct identification.

The major techniques for establishing the timing of on-line word-recognition—thereby answering the first of these two questions—involve fast reaction-time tasks. Typical examples are the shadowing and the identical monitoring tasks, where the listener responds directly to the words he hears—either by repeating them aloud, or by making a detection response to a word-target. The mean reaction-times in such tasks, measured from word-onset, can be used as a direct estimate of selection-time, subject to a correction factor to allow for the time it takes to execute the response.[1] Typical values obtained in these tasks (for one- and two-syllable content words heard

[1] The use of a correction factor compensates for the fact that a monitoring reaction-time of, for example, 250 ms, does not mean that the word was not identified until 250 ms of it had been heard. There is undoubtedly *some* lag between the internal decision process and the external evidence that this decision has been made. The correction factor reflects this.

in normal utterance contexts) are of the order of 250–275 ms, which, with a correction factor of 50–75 ms, gives a mean selection-time of around 200 ms (e.g., Marslen-Wilson, 1973, 1985; Marslen-Wilson & Tyler, 1975, 1980).

Similar values can be obtained, more indirectly, from reaction-time tasks where the listeners are asked to respond, not to the word itself, but to some property of the word whose accessibility for response depends on first identifying the word in question. Examples of this are the rhyme-monitoring results reported by Marslen-Wilson & Tyler (1975, 1980) and others (e.g., Seidenberg & Tanenhaus, 1979), and at least some research involving the phoneme-monitoring task (e.g., Marslen-Wilson, 1984; Morton & Long, 1976). By subtracting an additional constant from the response-times in these tasks, to take into account the extra phonological matching processes they involve, one again arrives at selection-times for words in context of the order of 200 ms from word-onset.

But these estimates are only half of the equation. It is also necessary to establish whether or not the acoustic-phonetic information available at these selection-points is sufficient for correct selection. For the research described above, this could only be done indirectly, by estimating the average number of phonemes that could be identified within 200 ms of word-onset, and then using that estimate to determine how many words would normally still be consistent with the input. If, as the available measurements suggest, 200 ms would only be enough to specify an initial two phonemes, then there would on average be more than 40 words still compatible with the available input (this estimate is based on the analysis of a 20,000-word phonetic dictionary of American English (Marslen-Wilson, 1984)). The limitation of this indirect inference to early selection is that it cannot take into account possible coarticulatory and prosodic effects. This could lead to an underestimate of the amount of sensory information actually available to the listener after 200 ms.

The second main technique allows a more direct measure of the sufficiency of the acoustic-phonetic input available at the estimated selection-point. This is the gating task, as developed by Grosjean (1980), and exploited by Tyler and others (e.g., Salasoo & Pisoni, 1985; Tyler & Wessels, 1983). Listeners are presented with successively longer fragments of a word, at increments ranging (in different experiments) from 20 to 50 ms, and at each increment they are asked to say what they think the word is, or is going to become. This tells us exactly how much acoustic-phonetic input the listener needs to hear to be able to reliably identify a word under various conditions. In the original study by Grosjean (1980), we find that subjects needed to hear an average of 199 ms of a word when it occurred in sentential context, as opposed to 333 ms for the same acoustic token presented in isolation.

Because of the unusual way the auditory input is presented in the gating task, there has been some criticism of its validity as a reflection of normal word-recognition processes. Since the listener hears the same fragments repeated many times in sequence, this might encourage abnormal response strategies. This objection is met by Cotton and Grosjean (1984) and Salasoo and Pisoni (1985), whose subjects heard only one fragment for any given word, and where the pattern of responses matched very closely the results for the same words when presented as complete sequences to each subject. It is also possible that responses are distorted by the effectively unlimited time—in comparison to normal listening—that listeners have available to think about what the word could be at each presentation. This objection is met by Tyler and Wessels (1985), in an experiment where subjects also heard only one fragment from each word, and where they responded by naming the word as quickly as possible. Mean naming latencies were 478 ms from fragment offset, and the response patterns again closely corresponded to those obtained without time-pressure.

In a recent study (Brown, Marslen-Wilson, & Tyler, unpublished) we have combined reaction-time measures for words heard normally with gating tests for the same words. This provides the most direct evidence presently available for early selection. In the first half of the experiment, subjects monitored pairs of sentences for word targets, with a mean reaction-time for words in normal contexts of 241 ms. This gives an estimated selection-time of 200 ms or less. In the second part of the experiment, the target-words were edited out of the stimulus tapes and presented, as isolated words, to a different set of subjects in a standard gating task. The mean identification-time estimated here was 301 ms, indicating that the words were being responded to in the monitoring task some 100 ms before sufficient acoustic-phonetic information could have accumulated to allow recognition on that basis alone.[2]

Given, then, that we have accurate and reliable estimates of the two variables in our equation, simple arithmetic tells us that content words, heard in utterance contexts, can usually be selected—and, indeed, recognised—earlier than would be possible if just the acoustic-phonetic input was being taken into account. Naturally, as Grosjean and Gee (1987, this issue) point out, some words—especially function words and short, infrequent content words—will often not be recognised early. In fact, under certain conditions of temporary ambiguity, as Grosjean (1985) has documented, "late" selection will occur, where the word is not only not recognised early, but may not even be identified until the word following it has been heard. These observations nonetheless do not change the significance of the fact that a large proportion

[2]There is still the problem here of factoring out the purely acoustic-phonetic effects of removing words from their contexts. We are investigating this in current [illegible]

of words *are* selected early. A theory of lexical access has to be able to explain this, just as it has to deal with late selection as well. Late selection, however, places far weaker constraints on the properties of the recognition process than does early selection.[3]

A different type of objection is methodological in character. It is argued that none of the tasks used to establish early selection are measuring "real" word-recognition. Instead, by forcing subjects to respond unnaturally early, they elicit some form of sophisticated guessing behaviour. Forster (1981), for example, argues that when a subject responds before the end of the word, as in the shadowing task, he must in some way be guessing what the word will be, on the basis of fragmentary bottom-up cues plus knowledge of context.

Such objections, however, have little force. First, because the claim that subjects are responding "unnaturally early" does not have any independent empirical basis. There is no counter evidence, from "more natural" tasks, showing that under these conditions different estimates of recognition-time are obtained—nor is the notion "more natural task" itself easy to defend except in terms of subjective preference. Secondly, to distinguish under these conditions between "*perception* of the target word and *guessing*" (Forster, 1981, p. 490; emphases in original) is to assume, as a theoretical a priori, a particular answer to the fundamental questions at issue.

Forster apparently wants to rule out, as an instance of normal perception, cases where the listener responds before all of the sensory information potentially relevant to that response has become available. But this presupposes a theory of perception where there is a very straightforward dependency between the sensory input and the corresponding percept. The claims that I am trying to develop here allow for the possibility of a less direct causal relationship between the sensory input and the percept (see Marcel, 1983, for a discussion of some related issues). These claims may or may not prove to be correct. But one cannot settle the issue in advance by excluding evidence on the grounds that it conflicts with the theoretical assumptions whose validity one is trying to establish. If one is advancing the view that normal perception *is* just the outcome of the integration of partial bottom-up cues with contextual constraints, then it is not an argument against this view simply to assert that perception under these conditions is not perception.

[3]It should also be clear, contrary to Grosjean and others, that the phenomenon of "late selection", does not constitute a problem for theories, like the cohort model, which emphasise the real-time nature of the word-recognition process. Activation-based versions of the cohort model, as discussed in Section 5, and as modelled, for example, in the McClelland and Elman (1986) TRACE model, function equally well independent of whether the critical sensory information arrives before or after the word boundary (as classically defined).

3. Implications of early selection

Early selection means that the acoustic-phonetic and the contextual constraints on the identity of a word can be integrated together at a point in time when each source of constraint is inadequate, by itself, to uniquely specify the correct candidate. The sensory input can do no more than specify a class of potential candidates, consisting of those entries in the mental lexicon that are compatible with the available input. Similarly, the current utterance and discourse context provides a set of acceptability criteria that also can do no more than delimit a class of potentially appropriate candidates. It is only by intersecting these two sets of constraints that the identity of the correct candidate can be derived at the observed selection-point. It is this that forces a parallel model of access and selection, and that poses intractable difficulties for any model which depends on an autonomous bottom-up selection process to reliably identify the single correct candidate for submission to subsequent processing stages (e.g., Forster, 1976, 1979, 1981).

To see this, consider the major functional requirements that early selection places upon the spoken word-recognition system. These are the requirements of *multiple access*, of *multiple assessment*, and of *real-time efficiency*. They reflect the properties the recognition system needs to have if it is to integrate sensory and contextual constraints to yield mean selection-times of the order of 200 ms.

Multiple access is the accessing of multiple candidates in the original mapping of the acoustic-phonetic input onto lexical representations. The sensory input defines a class of potential word-candidates, and, in principle, all of these need to be made available, via a multiple access process, to the selection phase of spoken word-recognition. The second requirement is the requirement for multiple assessment. If contextual constraints are to affect the selection phase at a point in time when many candidates are compatible with the sensory input, then the system must provide a mechanism whereby each of these candidates can be assessed for their syntactic and semantic appropriateness relative to the current context.

The final, and critical, requirement is for real-time efficiency. The system must be organised to allow these access and assessment activities to take place in real time, such that the correct candidate can be identified—and begin to be integrated into an utterance-level representation—within about 200 ms of word-onset.

These requirements, taken together, cannot be met by a serial process moving through the decision space one item at a time (cf. Fahlman, 1979). They point, instead, to some form of parallel or distributed recognition model (e.g., Hinton & Anderson, 1981). But they do not, however, uniquely deter-

mine the form of such a model. In particular, they do not unambiguously dictate the manner in which the word-recognition process is divided up into distinct processing stages. But they do place strong constraints on the functional properties of the recognition model. The strategy that I have followed, therefore, is to propose a model which rather literally embodies these constraints, and then to use this model as a heuristic starting-point for a detailed investigation of the properties of on-line speech processing. Accordingly, I will begin here by describing the first version of this model and the predictions it makes. In a later section, I will discuss the ways the model now needs to be expanded and modified.

The model in question, labelled an "active direct access model" in Marslen-Wilson and Welsh (1978), but now usually referred to as the "cohort model", evolved out of an analysis of Morton's logogen model (as stated in Morton, 1969) and of the Forster "bin" model (Forster, 1976). As originally stated, it meets the requirements of multiple access and multiple assessment by assuming a distributed, parallel processing system. In this system, each individual entry in the mental lexicon is assumed to correspond to a separate computationally active recognition unit. This unit represents a functional coordination of the acoustic-phonetic and of the syntactic and semantic specifications associated with a given lexical entry.

Given such an array of recognition elements, this leads to the characteristic "cohort" view of the recognition process, with its specific claims about the way this process develops over time. A lexical unit is assumed to become active when the sensory input matches the acoustic-phonetic pattern specified for that unit. The model prohibits top-down activation of these units in normal word-recognition, so that *only* the sensory input can activate a unit. There is no contextually driven pre-selection of candidates, so that words cannot become active as potential percepts without some bottom-up (sensory) input to the structures representing these words.

Early in the word, when only the first 100–150 ms have been heard, then the recognition devices corresponding to all of the words in the listener's mental lexicon that begin with this initial sequence will become active—thereby meeting the requirement for multiple access.[4] This subset of active elements, constituting the *word-initial cohort*, monitors both the continuing sensory input, and the compatibility of the words that the elements represent

[4]The notion of "activity" will be examined more closely in Section 5. What it means here is that each lexical recognition unit, as a computationally independent pattern-matching device, can respond to the presence of a match with the signal. All words that *could* match the input *are* matched by it, and this changes the state of the relevant pattern matching devices, thereby differentiating them from the other devices in the system, which do not match the current input.

with the available structural and interpretative context—which meets the requirement for multiple assessment. A mismatch with either source of constraint causes the elements to drop out of the pool of potential candidates. This means that there will be a sequential reduction over time in the initial set of candidates, until only one candidate is left. At this point, the correct word-candidate can be recognised, and the correct word-sense, with its structural consequences, is incorporated into the message-level representation of the utterance. This is a system that allows for optimal real-time efficiency, since each word will be recognised as soon as the accumulating acoustic-phonetic information permits, given the available contextual constraints.[5]

In terms of the issues raised earlier in this paper, the model treats the initial access phase as a functionally separable aspect of the recognition process. It does not do this by postulating an independent processing component which performs the access function—in the style, for example, of the peripheral access files proposed by Forster and others (e.g., Forster, 1976; Norris, 1981). It assumes, instead, that the processing mechanisms underlying word-recognition can only be engaged by a bottom-up input. It is the speech signal, and only the speech signal, that can activate perceptual structures in the recognition lexicon.[6] This has the effect of making access functionally autonomous, without having to make claims about additional levels and processes.

Once the word-initial cohort has been accessed, and the model has entered into the selection phase, then top-down factors begin to affect its behaviour. It is this that allows the model to account for early selection. When a word is heard as part of a normal utterance, then both sensory and contextual constraints contribute jointly to a process of mapping word senses onto

[5]The sequential cohort recognition process is sometimes treated as if it were equivalent to following a path down a "pronunciation tree". This is a branching structure, starting from a single phoneme (e.g., /t/), and branching at each subsequent phonetic choice point. By following the path to its terminal node one arrives at the correct word—*trespass*, *tress*, *trend*, or whatever. This captures in a limited sense the sequential decision process represented in the cohort model. Where it fails, however, is to capture the treatment of context in the cohort model. In a pronunciation tree, it is only when one reaches the terminal node that one can know what word one is hearing. It is only at this point, therefore, that the syntactic and semantic information associated with this word can be accessed, and made available for interaction with context. But the cohort model—and the evidence on which it is based—require context to be able to operate much earlier in the word, to help select the correct word even before the sensory input could have uniquely identified it. The pronunciation tree is neither an adequate model of human word-recognition nor an accurate depiction of the cohort model.

[6]It is not an argument against this claim to point out that one can often predict what someone is going to say before they say it. There is no doubt that this is true. But to be able to predict what someone will say is (i) not the same as having the percept that they have actually said it, nor (ii) is it evidence that this knowledge can penetrate, top-down, into the mental lexicon, and change the state of the basic recognition devices—and it is this that's at issue here.

higher-level representations. The way this is realised in the model is by allowing the semantic and syntactic appropriateness of word-candidates to directly affect their status in the current cohort, which causes the selection process to converge on a single candidate earlier than it would if only acoustic-phonetic constraints were being taken into consideration.

Even in this rough and ready form—that is, as stated in Marslen-Wilson and Welsh (1978) and Marslen-Wilson and Tyler (1980)—the model serves its heuristic purpose. It makes a number of strong predictions, which not only differentiate it from other models, but also, more importantly, raise novel and testable questions about the temporal microstructure of spoken word-recognition. In the next section of this paper I will summarise the research by myself and others into three of these major predictions: The model's claims about the concept of "recognition-point", about optimal real-time analysis, and about the early activation of multiple semantic codes.

4. Some predictions of the cohort model

4.1. The concept of recognition-point

The unique feature of the cohort model is its ability to make predictions about the precise timing of the selection and integration process for any individual word in the language. Other models have had essentially nothing to say about the recognition process at this level of specificity. The cohort model, in contrast, provides a theoretical basis for predicting the *recognition-point* for any given word. This is the point at which, starting from word-onset, a word can be discriminated from the other members of its word-initial cohort, taking into account both contextual and sensory constraints. For many words—especially monosyllables—this point may only be reached when all of the word has being heard. But for longer words—and for words of any length heard in constraining contexts—the recognition-point can occur well before the end of the word.[7]

[7] In a recent paper, Luce (1986) argues against the notion of recognition-point on the grounds that most common words are monosyllables and that most monosyllables (as he establishes by searching a lexical database) do not become unique until the end of the word or after. There are a number of problems with his argument.

The first is that he does not take into account the role of prosodic structure and of various types of anticipatory coarticulation in the recognition process. These will not only position the recognition-point earlier than a purely phonemic analysis would indicate, but will also reduce the potential problem created by short words that are also the first syllables of longer words. The second is that the claims of the cohort model derive, in the first instance, from observations of word-recognition in context, where even monosyllables are normally recognised before all of them have been heard (see Section 2 above). Thirdly, the important claim of the cohort model is, in any case, not whether the recognition-point falls early or late relative to the word-boundary, but rather that the word is uniquely discriminated as soon as the available constraints (sensory, contextual) make it possible for the system to do so. Wherever the recognition-point falls, that is where the listener should identify the word in question. And for content words heard in utterance context, this will be, more often than not, before all of the word has been heard.

Take, for example, the word "trespass". If this word is heard in isolation, then its recognition-point—the point at which it can be securely identified—is at the /p/, since it is here that it separates from words like "tress" and "trestle". The recognition-point for the same word in context might be at the first /s/, however, if these competitors were syntactically or semantically excluded. Similar predictions can be derived for any word in any context, given a specification of the word-initial cohort for that word, and of the constraints derivable from the context in which it is uttered.

The crucial hypothesis underlying the notion of recognition-point is a claim about the *contingency* of the recognition process. The identification of a word does not depend simply on the information that a given word is present. It also depends on the information that other words are *not* present. The word "trespass", heard in isolation, is only identifiable at the /p/ if the decision process can take into account, in real-time, the status of other potential word-candidates. The calculation of recognition-points directly reflects this. If these predicted recognition-points are experimentally validated, then this rules out all models of spoken word-recognition that do not allow for these dependencies.

4.1.1. Evidence for recognition-points

Paralleling the various types of evidence for early selection summarised in Section 2, the evidence for the psychological validity of recognition-points derives from a mixture of reaction-time and gating tasks. In a first experiment (Marslen-Wilson, 1978; Marslen-Wilson, 1984) response-latencies in a phoneme-monitoring task were found to be closely correlated with recognition-points, both as calculated a priori on the basis of cohort analysis, and as operationally defined in a separate gating task.

In phoneme-monitoring, the subject is asked to monitor spoken materials for a phoneme target defined in advance. There are two major strategies listeners can use to do this (cf. Cutler & Norris, 1979). I exploited here the lexical strategy, where the listener determines that a given phoneme is present by reference to his stored phonological knowledge of the word involved. When this strategy is used, response-latency is related to the timing of word identification, since the phonological representation of the word in memory cannot be consulted until it is known which word is being heard. If cohort theory correctly specifies the timing of word-identification, then there should

be a close dependency between the monitoring response and the distance between the phoneme-target and the recognition-point for that word. In particular, response-latency should decrease the later the target occurs relative to the recognition-point, since there will be less of a delay before the subject can identify the word and access its phonological representation.

I evaluated this question for a set of 60 three-syllable words, which contained phoneme targets varying in position from the end of the first syllable until the end of the third syllable. I had already confirmed that a lexical strategy was being used for these stimuli, since overall response-latencies dropped sharply over serial-positions, compared to a control set of nonsense words where there was no change in latency as a function of position (for further details, see Marslen-Wilson, 1984). The cohort structure of the materials was analysed to determine the recognition-point for each word, and the distances measured between the recognition-points and the monitoring targets. These recognition-points could occur as much as two or three hundred ms before or after the target-phoneme.

A linear regression analysis showed that there was a close relationship between these distances and the monitoring response ($r = +.89$).[8] The variations in distance accounted for over 80% of the variance in the mean latencies for the 60 individual words containing targets. This strong correlation with phoneme-monitoring latency shows that recognition-points derived from cohort analysis have a real status in the immediate, on-line processing of the word. The subjects in this experiment were using a lexical strategy, so that their response-latencies reflected the timing of word-recognition processes, and the cohort model correctly specified the timing of these processes for the words involved.

These results were checked in a follow-up study, which used the gating task to operationally define the recognition-points for the same set of materials. Gating offers a variety of methods for calculating recognition-points, depending on whether or not confidence ratings are taken into account. The most satisfactory results are obtained when confidence ratings are included, since this reduces the distorting effects of various response biases. Gating recognition-points were therefore defined as the point in a word at which 85% of the subjects had correctly identified the word, and where these subjects were at least 85% confident.[9] These operationally derived recognition-points correlated very highly both with the previous set of recognition-points (calculated on an a priori basis) and with the phoneme-monitoring response latencies ($r = +.92$).

The comparison between gating recognition-points and a priori recognition-points is further evidence that the cohort model does provide a basis for correctly determining when a word can be recognised. The point at which a word becomes uniquely identifiable, as established through an analysis of that word's initial cohort, corresponds very well to the point at which listeners will confidently identify a word in the gating task. This has been confirmed for a new set of materials, and, in particular, extended to words heard in utterance contexts, in a recent study by Tyler and Wessels (1983). The gating recognition-points calculated in this study are indeed the points at which a single candidate is left, and this point is not only quite independent of the total length of the word, but also varies in the manner predicted by the theory as a function of the availability of contextual constraints.

4.1.2. Implications of on-line recognition-points

The evidence for the psychological reality of the recognition-points specified by cohort analysis poses severe problems for certain classes of word-recognition model. The recognition-points were calculated on the basis not only of the positive information accumulating over time that a given word was present, but also, and equally important, the information that certain other words were *not* present. There is nothing, for example, about *trespass* by itself that predicts a recognition-point at the /p/—or indeed, anywhere else in the word. It is only in terms of the relationship of *trespass* to its initial cohort that the recognition-point can be computed. This contingency of the recognition response on the state of the ensemble of alternatives is in conflict with the basic decision mechanisms employed both by logogen-based theories and by serial search theories.

The results exclude, first, those recognition-models that depend on a self-terminating serial search, in the manner of Forster's models of access and selection (Forster, 1976, 1979, 1981). In this type of model, word-forms are stored in peripheral access files. These access files are organised into "bins", with the words within any one bin arranged in sequential order according to frequency. Once a bin has been accessed, there follows a serial search through the contents of the bin, terminating as soon as a word-form is encountered which matches the search parameters. The search must be self-terminating, since it is this that gives the model its ability to deal with frequency effects—frequent words are recognised more quickly because they are encountered earlier in the search process. Such a procedure could only take into account the status of competing word-candidates if they were higher in frequency

[8] The correlation is positive because the earlier the recognition-point occurs, relative to the position of the target phoneme (which is also the point from which response-time is measured), the longer the subjects have to wait until they can identify the word, access its phonological representation, and then make their response.

[9] The exact percentage chosen as criterial is not critical. Setting the level at 80 or 90%, for example, gives equivalent results.

There's everything in it for implicit kinds

than the actual best match. This would not predict the correct recognition-points.

It is in general a problem for sequential search models if the outcome of the recognition process needs to reflect the state of the entire ensemble of possibilities, since this makes the process extremely sensitive to the size of this ensemble. In fact, evidence I will cite later shows that the timing of word-recognition processes is not affected by the number of alternatives that need to be considered. Parallel access and selection processes are far better suited to the task of providing information about the status of several word-candidates simultaneously. But this by no means guarantees the suitability of all parallel models.

One type of parallel model that is excluded by the present results (as well as by the data reported in the next section) are the logogen-based models. These models depend on the accumulation of positive evidence within a single recognition device as the basis for recognition. Each device has a decision threshold, and the word that is recognised is the one whose corresponding recognition device (or logogen) crosses the threshold first, without reference to the state of any other recognition devices. The model has no mechanism for allowing the behaviour of one unit to take into account the behaviour of other units in the ensemble. This means that it has no basis for computing the recognition-point for a given word as a function of the timing with which that word emerges from the ensemble of its competitors, and, therefore, cannot explain the effectiveness of cohort-based recognition-points in accounting for response variation in the phoneme-monitoring task.

4.2. Optimal real-time analysis

The evidence for the psychological reality of recognition-points is also evidence for a more general claim about the properties of the word-recognition system. In a distributed model of access and selection, information coming into the system is made simultaneously available to all of the processing entities to which it might be relevant. This makes the system capable, in principle, of extracting the maximum *information-value* from the speech signal, in real-time as it is heard.

The information-value of the signal is defined with respect to the information that it provides, over time, for the discrimination of the correct word-candidate from among the total set of possible words that might be uttered. To use this information in an optimally efficient manner requires an access and selection process that can continuously assess the sensory input against all possible word-candidates. It is only by considering all possible lexical interpretations of the accumulating sensory input that the system can be sure, on the one hand, of not selecting an incorrect candidate, and, on the other, of being able to select the single correct candidate as soon as it becomes uniquely discriminable—that is, at the point where all other candidates become excluded by the sensory input. A series of experiments, using an auditory lexical decision task, show that listeners do have access, in real time, to information about the sensory input that could only have derived from an analysis process with these properties (Marslen-Wilson, 1980, 1984).

These experiments focused on the discrimination of nonwords, rather than on the timing of real-word recognition, because this made it possible to ask a wider range of questions about the processes of access and selection. The nonword stimuli—which the subjects heard mixed in with an equal number of real words—were constructed by analysing the cohort structure of sets of English words. The sequence "trenker", for example, becomes a nonword at the /k/, since there are no words in English beginning with /tren/ which have /k/ as a continuation. The use of this type of material allowed us to ask the following questions.

First, can listeners detect that a sound sequence is a nonword at precisely the point where the sequence diverges from the existing possibilities in English—that is, from the offset of the last phoneme in the nonword sequence that could be part of the beginning of a real word in English? If the selection process does continuously assess the incoming speech against possible word-candidates, then decision-time should be constant relative to critical phoneme offset. It should be independent both of the position of the critical phoneme in the sequence, and of the length of the sequence as a whole.

The results were unambiguous. Decision-time, measured from the offset of the last real word phoneme, was remarkably constant, at around 450 ms.[10] It was unaffected either by variations in the position of the nonword point (from the second to the fifth phoneme in the sequence), or by variations in the length of the nonword sequences (from one to three syllables). It appears that not only is there a continuous lexical assessment of the speech input, but also that this input itself is not organised into processing units any larger than a phoneme.

This latter point was investigated in a subsequent experiment (Marslen-Wilson, 1984), which looked specifically at the role of a larger unit—the syllable—in access and selection. If the speech input is fed to the mental

[10] We can also look at the results in terms of the relationship 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 relationship should approach 1.0, since reaction-time from sequence onset should increase as a linear function of the delay until the sequence becomes a nonword. The outcome is very close to this, with an observed slope of +.90, and with a correlation coefficient of +.97.

lexicon in syllable-sized chunks, then nonword decision-time, which depends on access to the lexicon, should increase the further the critical phoneme is from the end of the syllable. To test this, I used nonword sequences where the critical phoneme was either at the beginning, in the middle, or at the end of a syllable. This variation in position had no effect on decision-time, which remained constant at around 450 ms. The absence of any delay for syllable-internal targets shows that subjects do not need to wait until the end of a syllable to make contact with the lexicon. This is consistent with recent evidence (Cutler, Mehler, Norris, & Segui, 1983) that the syllable does not function as a processing unit in English.

The absence of length effects in these experiments appears to be fatal for standard logogen models. A weak point in this type of model, as I have noted elsewhere (Marslen-Wilson & Welsh, 1978), is its treatment of nonwords. A logogen-based recognition system cannot directly identify a nonword, since recognition depends on the triggering of a logogen, and there can be no logogen for a nonword. The system can only determine that a nonword has occurred if no logogen fires in response to some sensory input. But to know that no logogen will fire, it must wait until all of the relevant input has been heard. In the present experiment, therefore, nonword decision-times should have been closely related to item length. In fact, there was no relationship at all between these two variables.

The predicted effect of length derives directly from the fundamental decision mechanism around which logogen-based recognition models are constructed. The failure of this prediction means that we must reject such mechanisms as the building blocks for models of spoken word-recognition.

The second main question I was able to ask, using nonword stimuli, addressed more directly the claim for a parallel access and selection process. A major diagnostic of a parallel, as opposed to serial system, is its relative insensitivity to set size effects. For a distributed system like the cohort model, it need make no difference to the timing of the word-recognition decision whether two candidates have to be considered or two hundred. In either case, the timing of the selection process reflects the point at which a unique solution emerges. This is purely a matter of cohort structure, and has nothing to do with the number of alternatives per se. For a serial process, however, which moves through the alternatives in the decision space one item at a time, an increase in the number of alternatives must mean an increase in decision-time.

I investigated this in two experiments, in which I varied the size of the "terminal cohort" of sets of nonword sequences. This refers to the number of real words that are compatible with the nonword sequences at the point where they start to become nonwords—that is, at the offset of the last real-word phoneme. To make the nonword decision, all of these words presumably need to be analysed, to determine whether the subsequent speech input is a possible continuation of any of them. In the first experiment, the size of these terminal sets varied from one to 30. In the second, replicating the first, the range was from one to over 70. In neither case did I find an effect of set-size. Decision-time was constant, as predicted by the model, from the offset of the last real-word phoneme in the sequence, irrespective of whether only one real word remained consistent with the input up to this decision point, or of whether more than 70 still remained. This is evidence against any sequential search model of spoken word-recognition, whether self-terminating or not.

4.3. *The early activation of multiple semantic codes*

The two preceding sections focused on the way the cohort model leads one to think about the relationship between the sensory input and the mechanisms of access and selection. Here I consider the role of contextual constraints in the operation of these mechanisms.

The cohort model places severe restrictions on the ways in which contextual variables can affect the access and selection process. In particular, it prohibits the top-down pre-selection of potential word-candidates. It is the sensory input that activates the initial set of candidates, which can then be assessed against context. There is no top-down flow of activation (or inhibition) from higher centers, but, rather, the bottom-up activation of the syntactic and semantic information associated with each of the word-forms that has been accessed.

This has two major consequences. It means, first, that contextual constraints cannot prevent the initial accessing (i.e., the entry into the word-initial cohort) of words that do not fit the context. There is already indirect evidence for this from earlier work on lexical ambiguity (e.g., Seidenberg, Tanenhaus, Leiman, & Bienkowski, 1982; Swinney, 1979). More recently, research by Tyler (1984) and Tyler and Wessels (1983) shows that subjects in the gating task produce a substantial proportion of contextually inappropriate responses at the earlier gates–that is, when they have heard between 50 and 200 ms of the word. These are responses that are compatible with the available sensory input, but which do not fit the semantic and syntactic context in which these fragments occur. The existence of these responses at the early gates is evidence for the priority given by the system to the bottom-up input, and for the inability of context to suppress the initial activation of inappropriate candidates.

The second major consequence is that early in the recognition process

there will be the activation of multiple semantic and syntactic codes.[11] If contextual constraints are to affect the selection process, they can only do so within this framework, if they have access to the syntactic and semantic properties of the potential word-candidates. This information must be made available not only about the word that is actually being heard, but also about the other words that are compatible with the sensory input—that is, the other members of the current cohort.

We have evaluated these two claims by using cross-modal priming tasks to tap the activation of different semantic codes early in the recognition process. In these experiments (Marslen-Wilson, Brown, & Zwitserlood, in preparation; Zwitserlood, 1985), the subjects heard spoken words, and made lexical decision judgements to visual probes that were presented concurrently with these words. Previous research by Swinney and his colleagues (e.g., Onifer & Swinney, 1981; Swinney, 1979) had shown that lexical decisions to visually presented stimuli are facilitated when these words are associatively related to spoken words that are being presented at the same time.

The spoken words in our experiments were drawn from pairs of words such as CAPTAIN and CAPTIVE, which only diverge from each other relatively late in the word—in this case at the onset of the vowel following the /t/-burst. The visual probes, to which the subjects made their lexical decisions, were semantically associated with one or the other member of the pair of spoken words—in this case, for example, the probes might be the words SHIP and GUARD, where SHIP is frequently produced as an associate to CAPTAIN but never to CAPTIVE, and vice versa for GUARD. The critical variable, however, was the timing with which the visual probes were presented, relative to the separation-point in the spoken words. We contrasted two probe positions in particular: an Early position, where the probe appeared just before the separation-point, and a Late position, where it occurred at the end of the word, well after the separation-point.

The cohort model claims that both CAPTAIN and CAPTIVE will be accessed early in the selection process, and that this will make available the semantic codes linked to both of them. If this is correct, then there should be facilitation of the lexical decision for both visual probes when they occur

[11] It is important not to equate the kind of activation being postulated here with the activation effects detected by Swinney (1979) and Seidenberg et al. (1982) in experiments using homophones. In these experiments, subjects hear a complete word-form—like "bug" or "rose"—that has two or more different meanings. Under these conditions, there is a strong activation of both meanings, which appears to persist for as much as a second after word offset. This is not the same as the phenomena predicted here, where the transient match, early in the word, of the incoming signal with a number of different word-forms leads to the transient activation of the semantic and syntactic codes associated with these forms. These effects are only the faint precursors of the activation effects to be expected when there is a full match of the input to a given word-form, as in the homophone experiments.

in the Early position. Decision-time for SHIP and GUARD should, therefore, be affected equally when these probes are presented on or before the /t/ in either CAPTIVE or CAPTAIN. In contrast, when the probes are presented in the Late position, then only the probe related to the actual word should be facilitated. If the word is CAPTAIN, for example, there should be facilitation of SHIP at the end of the word but not of GUARD.

This pattern should hold both for isolated words and for the same words in context. If the initial access, first of word-forms, and then of the syntactic and semantic information associated with these word-forms, is triggered from the bottom-up, and if contextual effects can only operate on this information after it has been accessed in this way, then the presence or absence of contextual constraints should not affect the pattern of activation of semantic codes at the early positions.

In a series of experiments this was exactly what we found. For words in isolation we see facilitation of both probes for the Early locations, but only facilitation of one probe at the Late positions (Marslen-Wilson et al., in preparation; Zwitserlood, 1985). The same pattern holds for words in context (Zwitserlood, 1985). The differential facilitation of probes associated with contextually appropriate as opposed to contextually less appropriate words only begins to appear after about 200 ms. At earlier probe positions, there is evidence for the activation of semantic codes linked to contextually inappropriate words, just as we find for words in isolation.

These results support the fundamental claim of the cohort model that the recognition process is based not only on multiple bottom-up access, but also on multiple contextual assessment (as discussed in Section 3). They also demonstrate that the involvement of contextual variables early in the selection process takes place under highly constrained conditions. No contextual preselection is permitted, and context cannot prevent the accessing and activation of contextually inappropriate word-candidates.

These conclusions distinguish the first version of the cohort model both from standard autonomy models and from standard interactive models. The cohort model differs from autonomous models, because it allows contextual variables to affect the selection process. But it shares with autonomy theories the assumption that initial access is autonomous, in the sense that top-down inputs cannot activate perceptual structures in the recognition lexicon.

This partial "autonomy" distinguishes the cohort model from theories which do permit top-down influences on initial access. One example is the logogen model, where logogens can be activated by inputs from the cognitive system as well as by bottom-up inputs. Another, more topical example, is the interactive activation model put forward by Rumelhart and McClelland

(1981), and recently applied to spoken word-recognition in the form of the TRACE model (Elman & McClelland, 1984). This is an approach that can accommodate many of the phenomena driving the cohort model—and, indeed, this was what it was initially designed to do.

It is not clear, however, whether TRACE (or its predecessor COHORT), with its mixture of excitatory connections between levels and inhibitory connections within levels, can accommodate the pattern of semantic activation described here for members of the same cohort heard in context and isolation. It should, first, predict differential patterns early in recognition for the contextually appropriate word, because of feed-forward from excitatory top-down connections. Secondly, because of the inhibitory connections between units within a level, there should be very little early activation of competing words like CAPTAIN and CAPTIVE. They should mutually inhibit each other until after their separation-point. Neither prediction is consistent with our results.

Finally, it is worth remembering that any evidence which reinforces the claims for multiple contextual assessment also serves to underline the fundamental inability of sequential search models to explain the observed properties of the on-line transfer-function of the recognition system—namely, the convergence of two sets of criteria, sensory and contextual, onto a unique solution within 200 ms of word-onset.

5. Information and decision in the cohort model: Some revisions and extensions

The results summarised in the preceding sections illustrate the value of the cohort approach as a basis for research into spoken word-recognition, and they support the accuracy of the claims it embodies about the functional characteristics of the recognition process. Nonetheless, it is also clear that the internal structure of the model, as originally stated, is over-simplified and inadequate on several counts.

In this final section of the paper I want to discuss some problems with the handling of information and decision in the cohort theory. These problems concern the nature of the information coming into the system, the way that information is represented within the system, and the way in which decisions are taken to exclude or include candidates for selection and recognition.

I will argue, in particular, that the cohort model has to move away from its binary concept of information and decision, where candidates are either in the cohort or out of it, towards a more fluid form of organisation, incorporating the concept of activation. The rationale for this derives, first of all, from some recent evidence for the role of word-frequency in the early stages of access and selection.

5.1. *Activation and word-frequency*

As originally stated, the cohort model made no mention at all of word-frequency. The main reason for this was the absence of compelling evidence that word-frequency was an effective variable in the kinds of on-line analysis processes with which the model is concerned. The older research in this area (e.g., Broadbent, 1967; Howes, 1957; Morton, 1969; Pollack, Rubinstein, & Decker, 1960) showed that word-frequency affects the intelligibility of spoken words heard in noise. But it was never clear whether these were immediate perceptual effects or due to post-perceptual response biases.

More recent research, using reaction-time techniques, was flawed by its failure to take into account the distribution of information over time in spoken words. Unless the high and low frequency words in an experiment are matched for recognition-point, and unless reaction-time is measured with respect to this point, then any measures of response-time to the two different classes of stimuli are difficult to interpret. This is the problem, for example, with the auditory lexical decision data reported by McCusker, Holley-Wilcox, and Hillinger (1979) and by Blosfeld and Bradley (1981). Both studies show faster response times to high frequency as opposed to low frequency monosyllables. But in each case reaction-time was measured from word-onset, with no correction for possible variations in recognition-point.

Two new studies provide better evidence for the role of word-frequency. In a preliminary study I looked at lexical decision latencies for matched pairs such as STREET and STREAK, where the recognition-point for each word is in the word-final stop-consonant.[12] This means that reaction-time can be measured from comparable points in each member of the pair—in this case, from the release of the final stop. For a set of 35 matched pairs, with mean frequencies, respectively, of 130 per million and 3 per million, there was a considerable advantage for the high-frequency words (387 vs. 474 ms).

Evidence of a different sort shows that these frequency effects can be detected early in the selection process. This evidence comes from the research on the early activation of semantic codes (see Section 4.3), where we found that the frequency of the spoken words being heard indirectly affected the amount of priming of the concurrent visual probe.

The effective variable was the *difference* in frequency between the word

[12] This was research carried out under my supervision by R. Sanders and E. Eden in 1983, as part of an undergraduate research project in the Cambridge Department of Experimental Psychology.

being heard and its closest competitor—in this experiment usually the other member of the stimulus pair. For the Early probes, presented before the spoken words had separated from each other, we regularly found more facilitation for the probe related to the more frequent member of the pair, with the size of this effect varying according to the size of the frequency difference between the two words.

The word CAPTAIN, for example, is more frequent than its close competitor CAPTIVE. For visual probes presented in the Early position, just before the /t/, there would be more facilitation of SHIP (the probe related to CAPTAIN) than of GUARD (the probe related to CAPTIVE), irrespective of whether the word actually being heard was CAPTAIN or CAPTIVE (Marslen-Wilson et al., in preparation). But for Late probes, presented at the end of the spoken word, these effects of relative frequency had disappeared, so that only the probe associated with the actual word being heard would be facilitated. Comparable effects were found by Zwitserlood (1985), in a study where the relative frequency of the members of such pairs was systematically varied.

These appear to be genuine perceptual effects, reflecting competition between different candidates early in the selection process. Alternative explanations, in terms of post-perceptual response-bias, can be excluded. If there are any bias effects in the data, they will reflect the properties of the visual probes rather than the spoken words, since it was the visual probes the subjects were actually responding to. They were not being asked to make any judgements about the identity or lexical status of the spoken words, nor, in general, did they seem to be aware that there was a relationship between these words and the visual probes. Furthermore, since the effects hold only for the Early probes, they reflect the state of the system *during* the selection phase, and not after it is completed.

Finally, and most significantly for the activation argument, these effects are transient. The effects of relative difference in frequency have dissipated by the time the Late probes are presented (between 200 to 300 ms later). What we appear to be picking up earlier in the word is a temporary advantage accorded to frequent word-forms, where the size of this advantage reflects the degree of differential activation of word-forms and their closest competitors.

Related transient effects can be seen in some other studies. For example, Blosfeld and Bradley (1981) only found significant frequency effects for monosyllables. In disyllabic words, lexical decision time did not vary according to frequency. This is because lexical decision is a task where the listener needs to wait until the end of the word before making a positive response, to make sure that he is not hearing a nonword. If the end of the word comes

significantly later than the recognition-point, as will usually be the case for disyllables, then the effects of word-frequency at the recognition-point will have dissipated when the time comes for the subject to respond.[13] Finally, in the gating task the effects of frequency appear systematically only at the earliest gates (Tyler, 1984).

On the basis of this, I conclude the following. We can still assume that all word-forms which match a given input will be accessed by that input, and will remain active candidates as long as there is a satisfactory match with the sensory input. However, the response of higher-frequency word-forms appears to be enhanced in some way, such that the level of activation of these elements can rise more rapidly, per unit information, than the activation of less frequent elements (cf. Grosjean & Itzler, 1984).

This means that, early in the word, high-frequency words will be stronger candidates than lower-frequency words, just because their relative level of activation will be higher. This transient advantage is what the priming data reflect. And since the selection process is dependent on the emergence of one candidate from among a range of competitors, this should lead to faster recognition-times for high-frequency words than for low-frequency words, especially for low-frequency words with high-frequency competitors. This is because the activation of high-frequency competitors will take longer to drop below the level of the low-frequency candidate, once the critical sensory information has become available which excludes this high-frequency competitor.

To adopt this kind of account means that the behaviour of the cohort system can no longer be characterised in terms of the simple presence or absence of positive or negative information. Elements are not simply switched on or off as the sensory and contextual information accumulates, until a single candidate is left standing. Instead, the outcome and the timing of the recognition process will reflect the differential levels of activation of successful and unsuccessful candidates, and the rate at which their respective activation levels are rising and falling.

Some recent attempts to model a cohort-like analysis process have, in fact, represented the behaviour of the system in these or very similar terms (e.g., Elman & McClelland, 1984; Marcus, 1981, 1984; McClelland & Elman, 1986; Nusbaum & Slowiaczek, 1983). The results of these simulations show that an activation-based system is capable of exhibiting the main characteristics of a cohort selection process, with the correct candidate emerging from among its

[13] I should note, however, that recent research by Frauenfelder (personal communication) has

competitors as the discriminating acoustic-phonetic information starts to accumulate.

But apart from being strongly suggested by the word-frequency data, the activation concept has advantages in other respects. In particular, it enables us to deal in a more satisfactory manner with a second set of issues raised by the cohort model's treatment of information and decision. These concern the nature of the sensory and contextual input to the decision process, and the way that the matching of these inputs to lexical representations affects this process.

5.2. *Matching processes in access and selection*

In the initial formulation of the cohort model it was assumed that the matching process was conducted on an all-or-none basis. The sensory and the contextual input either did or did not match the specifications for a given candidate. If it did not, then the candidate would be dropped from the cohort.

The trouble with this account is that it makes the successful outcome of the recognition process dependent on an unrealistically perfect match between the specifications of the correct candidate and the properties of the sensory input and the context. I will begin with the problems raised by variability in the bottom-up input.

5.2.1. *Matching the sensory input*

The cohort model emphasises the role of sensory information in determining the scope and characteristics of the access and selection process. It is this that determines the membership of the word-initial cohort, and that has the priority in determining which candidates remain in the cohort and which are dropped. The available evidence suggests that this is the correct view to take (see Section 4.3).

To take this view, however, is to run the risk of making the recognition process too sensitive to noise and variation in the sensory input. If sensory information is the primary determinant of cohort membership, and if the matching process operates on an all-or-none basis, then even a small amount of variability in the sensory signal could lead to problems in recognition, with the correct word-candidate either never making it into the word-initial cohort, or being dropped from it for spurious reasons.

In fact, the human spoken word-recognition system seems to be remarkably indifferent to noise in the signal, so long as the disrupted input occurs in an utterance context. Even when deviations are deliberately introduced into words—as in the mispronunciation detection task (Cole, 1973; Marslen-Wilson & Welsh, 1978)—listeners often fail to notice them. Over 70% of small changes (i.e., changes by a single distinctive feature) are not detected when they occur in words in utterance contexts, even though the same changes are readily detectable in isolated syllables (Cole, 1973).

To accommodate this type of result, the model must find some way of permitting deviant words to enter the cohort. The model can only allow context to compensate for deficiencies in the bottom-up specification of the correct candidate if this candidate nonetheless manages to find its way into the cohort.

There are two aspects to the solution of this problem. The first follows from the activation-based selection process sketched out in the previous section. This is not a decision process that requires all-or-none matching, since to discriminate the correct candidate it is not necessary to systematically reduce the cohort to a single member. Selection does not depend on simple presence or absence in the cohort, but on relative goodness of fit to the sensory input. This makes it in principle possible for candidates that do not fully match the sensory input to participate nonetheless in the recognition process.

The second aspect of the solution involves the model's assumptions about the nature of the input. The system will respond quite differently to deviant or noisy input, depending on the description under which this input is fed into the decision process. The more highly categorised the output of acoustic-phonetic analysis, the greater the problems that will be caused by variability and error (cf., Klatt, 1980). In fact, if the cohort model is going to be able to allow contextual constraint to compensate for bottom-up variability, then the input to the lexicon cannot be anything as abstract as a string of phonemes. Instead, a representation is required which preserves more information about the acoustic-phonetic properties of the input—for example, a representation in terms of a feature matrix.

To see this, consider the consequences of minor disruptions of the signal when we adopt different assumptions about the input. Suppose that the disturbance is such that a word-initial voiced stop—for example, /be/—is misidentified as a voiceless stop (/pe/). If the input to the word-recognition system takes the form of a string of phonemic labels, then this error will have drastic consequences for the membership of the cohort. A match will be established for all words beginning with /pe/, and these will be strongly activated. But the word intended by the speaker, beginning with a /b/, will receive no activation at all.

In contrast, if the input is specified in terms of a set of feature values, then such an error will have much less drastic consequences. A minimal pair like /b/ and /p/ only differ in their specifications along one feature parameter—in this case voicing. Even if a wrong assignment is made on this parameter, the

input will still match the specifications for /b/ words along all of the other parameters. This means much less differentiation in the degree of match and mismatch between the /be/ and the /pe/ sets, so that the word-form intended by the speaker has a much better chance of receiving sufficient activation to be treated as a candidate for selection and recognition. In other words, the system will become more tolerant of minor deviations in the sensory input.

To assume a less highly categorised input to the lexicon does not sacrifice the ability of the system to discriminate among different alternatives. There is no inherent advantage to making phonemic distinctions at a pre-lexical decision stage, and the choice between two phonemes can be made just as well at the lexical level, as part of the choice between two words. In each case, the decision takes into account the same bottom-up information. The advantage of making the decision at the lexical level is that it enables the system to delay committing itself to final decisions about the properties of the sensory input until the other information relevant to this decision—including the lexical status of different alternatives and their contextual roles—can be taken into account (Klatt, 1980).

5.2.2. Matching the context

The evidence that selection is intimately bound up with integration lies at the heart of the argument for a distributed model of spoken word-recognition. But despite this, the way that the first version of the cohort model handles the relationship between selection and contextual constraints is seriously flawed.

Early statements of the model (e.g., Marslen-Wilson & Welsh, 1978) assert that candidates drop out of the pool of word-candidates when they do not fit the specifications of context, in the same way as when they do not fit the accumulating sensory input. This runs into similar problems to the all-or-none assumptions about sensory matching that I have just discussed. For the sensory input, the problem was to explain how mispronounced, or otherwise deviant words could nonetheless still be correctly identified. For context, the problem is to explain how contextually anomalous words can be identified (e.g., Norris, 1981).

Commonsense experience, as well as experimental evidence, tells us that contextually inappropriate words can, in fact, be readily perceived and identified, so long as they are unambiguously specified in the signal. In a recent experiment, for example, we compared monitoring latencies to the same target under conditions where it was either normal with respect to its context, or was anomalous in varying degrees of severity (Brown et al., unpublished). Consistent with earlier results, there was a clear effect of anomaly. Response latency to the word GUITAR increased by 27 ms over normal when it occurred in an implausible context ("John buried the guitar"), and by a further 22 ms when it occurred in a semantically anomalous context ("John drank the guitar"). But equally clearly, these anomalies are not causing a major breakdown of the recognition process. In the semantically anomalous condition, for example, response-latencies remain well below 300 ms, and the error rate is essentially zero. Even for grossly anomalous targets ("John slept the guitar"), where verb sub-categorisation constraints are also violated, response-time is still a relatively rapid 320 ms, and the error-rate remains low.

The relative speed and accuracy of correct selection for contextually inappropriate candidates is a reflection of the principle of bottom-up priority (Marslen-Wilson & Tyler, 1980, 1983). The system is organised so that it cannot override unambiguous bottom-up information. This means that there is a considerable asymmetry in the degree to which context can override bottom-up mismatch as opposed to the ability of bottom-up information to override contextual mismatch. If the sensory input clearly differentiates one candidate from all others, then that is the candidate that will emerge from the perceptual process, irrespective of the degree of contextual anomaly. If contextual variables clearly indicate a given candidate, it will nonetheless not emerge as the choice of the system unless it also fits the bottom-up input (within the limits of variation indicated earlier).

The clear implication of this is that context does not function to exclude candidates from the cohort. There is no all-or-none matching with context, and no all-or-none inclusion or exclusion of candidates on this basis. This parallels the points made earlier (Section 4.3), prohibiting top-down influences upon initial access. It looks as if contextual factors can neither determine which candidates can enter the cohort, nor which candidates must leave it.

If we accept this conclusion, then there are two lines we can follow. One is to maintain an interactive model, but to restrict the kinds of top-down effects that are permitted. Since inhibitory effects are now excluded, context will only have facilitatory effects, perhaps by increasing the level of activation of candidates that fit the current context. Alternatively, we can turn towards a different type of model, where no top-down interactions of any sort are permitted. Different types of information are integrated together on-line to produce the perceptual output of the system, but they do not interact in the conventional sense. I will explore here the possibilities for this second kind of account.

The effects of context, within the general framework I have adopted in this paper, reflect the processing relationship between selection and integration. This is the relationship between, on the one hand, the set of potential word-candidates, triggered from the bottom-up, and, on the other, the

higher-level representation of the current utterance and discourse. This contextual representation provides a structured interpretative framework against which the senses associated with different word-forms can be assessed. In a non-interactive model, this framework does not, itself, operate directly on the activation levels of different candidates. These activation levels are a measure of the relative goodness of fit of the candidates to the bottom-up input, and context does not tamper with this measure.

We can capture, instead, the phenomena of early selection, and of contextual compensation for bottom-up deficiency, by exploiting the capacity of a parallel system for multiple access and multiple assessment. These will lead to a form of on-line competition between the most salient candidates (those most strongly activated by the sensory input) to occupy the available sites in the higher-level representation. Once the appropriate senses associated with a given word-form have been bound to these locations in the representation, then we can say that recognition has taken place.[14]

The speed with which this is accomplished will be the joint function of two variables: the extent to which the bottom-up fit for a given candidate differentiates it from its competitors, and the extent to which the contextual match similarly differentiates it. The facilitatory and compensatory effects of context reflect the tendency of the system to commit itself to a particular structural interpretation even though the sensory input may not have fully differentiated the word-form associated with this interpretation. The reason for this lack of full bottom-up differentiation may be either temporal—not all of the sensory input relevant to the decision has been heard yet, or it may be substantive—the sensory input is simply inadequate by itself to indicate a unique candidate.

On this account, both access and certain aspects of selection are autonomous processes, in the sense that they are driven strictly from the bottom-up. Whether the speech signal is heard in context or in isolation, the basic pattern-matching routines of the system will operate in the same way, providing information about the goodness of fit of the sensory signal to the array of lexical representations of word-forms.

This means that when the signal is heard in isolation, we will get something approximating the commonsense concept of word-recognition—that is, a process of form-based selection culminating in the explicit decision that a given word-form is present. But when the signal is heard in context—and note that normal context is fluent conversational speech—there need be no explicit form-based recognition decision. Selection—viewed as the decision that one particular word-form rather than another has been heard—becomes a by-

[14] It is at this point (Marslen-Wilson & Welsh, 1978) that the output of the system becomes perceptually available.

product of the primary process of mapping word-senses into higher-level representations. The bottom-up access and selection processes provide the essential basis for rapid on-line comprehension processes, but they provide no more than a partial input to an integrative system that is only peripherally concerned with identifying word-forms, and whose primary function is to uncover the meanings that the speaker is trying to communicate.

5.2.3. The new cohort

In the preceding section of this paper I have suggested a number of modifications in the way that the cohort concept should be realised as a processing model. These include the use of the activation concept, the introduction of frequency effects into the early stages of the recognition process, the specification of the bottom-up input in terms of some form of sub-phonemic representation, and the exclusion of top-down contextual influences on the state of the actual lexical recognition units. What do these changes mean for the central concepts of the approach, with its emphasis on the contingency of perceptual choice, and on the processing concept of the word-initial cohort?

By moving away from the concept of all-or-none matching against sensory and contextual criteria, and by adopting an activation metaphor to represent the goodness of fit of a given candidate to the bottom-up input, the model abandons the convenient fiction that the cohort is a discrete, well-demarcated entity in the mental life of the listener. The selection process does not depend on the membership of the cohort reducing to a single candidate. It depends instead on the process of mutual differentiation of levels of activation of different candidates. The operation of the system still reflects the state of the entire ensemble of possibilities, but the state of this ensemble is no longer represented simply in terms of the all-or-none presence or absence of different candidates.

Functionally, however, the cohort still exists. The effective core of salient candidates will be much the same as it would have been under an all-or-none regime. Although very many candidates will be momentarily activated as aspects of their phonological representations transiently match the accumulating input, the preceding and subsequent input will not match, and they will fall back into semi-quiescence. It takes some amount of time and input for candidates to start to participate fully in the selection and integration process. The effect of this is that the set of candidates which must be discriminated among will look very similar to the membership of the word-initial cohort as defined on an all-or-none basis. But by not defining it on this all-or-none basis, the system becomes far better equipped to deal with the intrinsic and constant variability of the speech signal.

Overall, none of the modifications I have suggested change the fundamen-

tal functional characteristics of the cohort-based word-recognition process. It still embodies the concepts of multiple access and multiple assessment, allowing a maximally efficient recognition process, based on the principle of the contingency of real-time perceptual choice.

References

Blosfeld, M.E., & Bradley, D.C. (1981). Visual and auditory word recognition: Effects of frequency and syllabicity. Paper presented at the Third Australian Language and Speech Conference, Melbourne.

Broadbent, D.E. (1967). Word-frequency effect and response-bias. *Psychological Review, 74*, 504–506.

Brown, C.M., Marslen-Wilson, W.D., & Tyler, L.K. (no date). Sensory and contextual factors in spoken word-recognition. Unpublished manuscript, Max-Planck Institute, Nijmegen.

Cole, R.A. (1973). Listening for mispronunciations: A measure of what we hear during speech. *Perception & Psychophysics, 13*, 153–156.

Cotton, S., & Grosjean, F. (1984). The gating paradigm: A comparison of successive and individual presentation formats. *Perception & Psychophysics, 35*, 41–48.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1983). A language-specific comprehension strategy. *Nature, 304*, 159–160.

Cutler, A., & Norris, D. (1979). Monitoring sentence comprehension. In W.E. Cooper & E. Walker (Eds.). *Sentence processing: Psycholinguistic studies presented to Merrill Garrett.* Hillsdale, NJ: Erlbaum.

Elman, J.L., & McClelland, J.L. (1984). Speech perception as a cognitive process: The interactive activation model. In N. Lass (Ed.), *Speech and Language, Vol. 10*. New York: Academic Press.

Fahlman, S.E. (1979). *NETL: A system for representing and using real-world knowledge*. Cambridge, MA: MIT Press.

Forster, K.I. (1976). Accessing the mental lexicon. In R.J. Wales & E. Walker (Eds.) *New approaches to language mechanisms*. Amsterdam: North-Holland.

Forster, K.I. (1979). Levels of processing and the structure of the language processor. In W.E. Cooper & E. Walker (Eds.) *Sentence processing: Psycholinguistic studies presented to Merrill Garrett.* Hillsdale, NJ: Erlbaum.

Forster, K.I. (1981). Priming and the effects of sentence and lexical contexts on naming time: Evidence for autonomous lexical processing. *Quarterly Journal of Experimental Psychology, 33A*, 465–495.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics, 28*, 267–283.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception & Psychophysics, 28*, 299–310.

Grosjean, F., & Gee, J.P. (1987). Another view of spoken word recognition. *Cognition, 25*, this issue.

Grosjean, F., & Itzler, J. (1984). Can semantic constraint reduce the role of word frequency during spoken-word recognition? *Bulletin of the Psychonomic Society, 22*, 180–182.

Hinton, G.E., & Anderson, J.A. (Eds.) (1981). *Parallel models of associative memory*. Hillsdale, NJ: Erlbaum.

Howes, D. (1957). On the relationship between the intelligibility and the frequency of occurrence of English words. *Journal of the Acoustical Society of America, 29*, 296–305.

Klatt, D.H. (1980). Speech perception: A model of acoustic-phonetic analysis and lexical access. In R.A. Cole (Ed.) *Perception and production of fluent speech*. Hillsdale, NJ: Erlbaum.

Luce, P.A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics, 39*, 155–159.

Marcel, A.J. (1983). Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes. *Cognitive Psychology, 15*, 238–300.

Marcus, S.M. (1981). ERIS—context-sensitive coding in speech perception. *Journal of Phonetics, 9*, 197–220.

Marcus, S.M. (1984). Recognizing speech: On the mapping from sound to word. In H. Bouma & D.G. Bouwhuis (Eds.) *Attention and Performance X: Control of language processes*. Hillsdale, NJ: Erlbaum.

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

Marslen-Wilson, W.D. (1978). Sequential decision processes during spoken word recognition. Paper presented to the Psychonomic Society, San Antonio, Texas.

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

Marslen-Wilson, W.D. (1984). Function and process in spoken word-recognition. In H. Bouma and D.G. Bouwhuis (Eds.) *Attention and Performance X: Control of language processes*. Hillsdale, NJ: Erlbaum.

Marslen-Wilson, W.D. (1985). Speech shadowing and speech comprehension. *Speech Communication, 4*, 55–73.

Marslen-Wilson, W., Brown, C.M. & Zwitserlood, P. (in preparation). Spoken word-recognition: early activation of multiple semantic codes. Manuscript in preparation. Max-Planck Institute, Nijmegen.

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

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

Marslen-Wilson, W.D., & Tyler, L.K. (1981). Central processes in speech understanding. *Philosophical Transactions of the Royal Society, Series B, 295*, 317–332.

Marslen-Wilson, W.D., & Welsh, A. (1978). Processing interactions during word-recognition in continuous speech. *Cognitive Psychology, 10*, 29–63.

McClelland, J.L., & Elman, J.L. (1986). The TRACE model of speech perception. In McClelland, J.L., & Rumelhart, D.E. (Eds.) *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, Mass. Bradford Books.

McCusker, L.X., Holley-Wilcox, P., & Hillinger, M.L. (1979). Frequency effects in auditory and visual word recognition. Paper presented to the Southwestern Psychological Association, San Antonio, Texas.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76*, 165–178.

Morton, J., & Long, J. (1976). Effect of word transitional probability on phoneme identification. *Journal of Verbal Learning and Verbal Behavior, 15*, 43–52.

Norris, D. (1981). Autonomous processes in comprehension. *Cognition, 11*, 97–101.

Nusbaum, H.C., & Slowiaczek, L.M. (1983). An activation model of the cohort theory of auditory word recognition. Paper presented at the Society for Mathematical Psychology, Boulder, Colorado.

Onifer, W., & Swinney, D.A. (1981). Accessing lexical ambiguities during sentence comprehension: Effects of frequency of meaning and contextual bias. *Memory & Cognition, 9*, 225–236.

Pollack, I., Rubinstein, H., & Decker, L. (1960). Analysis of correct responses to an unknown message set. *Journal of the Acoustical Society of America, 32*, 454–457.

Rumelhart, D.E., & McClelland, J.L. (1981). An interactive activation model of context effects in letter perception, Part II: The contextual enhancement effect and some tests and extensions of the model. *Psychological Review, 89*, 60–94.

Salasoo, A., & Pisoni, D. (1985). Interaction of knowledge sources in spoken word identification. *Journal of Verbal Learning and Verbal Behavior, 24*, 210–231.

Seidenberg, M.S., & Tanenhaus, M.K. (1979). Orthographic effects on rhyme monitoring. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 546–554.

Seidenberg, M.S., Tanenhaus, M.K., Leiman, J.M., & Bienkowski, M. (1982). Automatic access of the

meanings of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology, 14*, 489–537.

Swinney, D. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 14*, 645–660.

Tyler, L.K. (1984). The structure of the initial cohort: evidence from gating. *Perception & Psychophysics, 36*, 217–222.

Tyler, L.K., & Wessels, J. (1983). Quantifying contextual contributions to word-recognition processes. *Perception & Psychophysics, 34*, 409–420.

Tyler, L.K., & Wessels, J. (1985). Is gating an on-line task? Evidence from naming latency data. *Perception & Psychophysics, 38*, 217–222.

Zwitserlood, P. (1985). Activation of word candidates during spoken word-recognition. Paper presented to Psychonomic Society Meetings, Boston, Mass.

Résumé

La reconnaissance de mots (dans la chaîne parlée) englobe trois fonctions fondamentales: l'accès, la sélection et l'intégration. L'*accès* se réfère à l'appareillement de l'onde sonore avec les représentations de formes lexicales; la *sélection*, désigne la discrimination du meilleur "pareil" (match) lexical avec le stimulus, et *l'intégration* recouvre l'appareillement de l'information syntaxique et sémantique avec les niveaux de traitement supérieures.

Cet article décrit comment deux versions d'un modèle (de type "cohorte") rendent compte de ces processus, en traçant son évolution à partir d'une première version comportant un principe d'interaction partielle où l'accès est strictement autonome mais où la sélection est soumise à des contrôles "de haut en bas" vers une deuxième version (à fonctionnement entièrement "de bas en haut") où le contexte n'intervient plus dans les processus d'accès et de sélection.

Par conséquent, le contexte n'intervient qu'à l'interface entre les représentations supérieures et l'information générée en temps réel sur les propriétés syntaxiques et sémantiques des membres du cohorte. Ce nouveau modèle garde intactes les caractéristiques essentielles d'un processus de reconnaissance de type cohorte. Il intègre les notions d'accès et d'évaluation multiples permettant ainsi un processus de reconnaissance optimal fondé sur le principe de contingence de choix perceptif.