3 *Modelling the lexicon*

In Chapter 2 we looked at the question of how lexis is acquired. In the present chapter we turn our attention to the equally fascinating issue of how the lexis that is acquired is managed; in other words, we shall be considering the structure of the lexical storage system and the ways in which that system is accessed under different conditions. We shall also be looking at lexical processing within two broader theoretical frames of reference – respectively, the modularity hypothesis and connectionism.

The present chapter concerns itself mostly with research which does not have a specifically L2 or bilingual focus. However, in research relating to the L2 mental lexicon the same kinds of organizational and operational issues arise as in L1-focused research, the difference being that in the L2 case they are further complicated by questions having to do with precisely the fact that more than one language comes into the picture. These latter questions – (1) the degree to which the L2 lexicon resembles the L1 lexicon and (2) the degree to which and ways in which the L2 lexicon interacts with the L1 lexicon – will be addressed in Chapter 4. With regard to (1), we have already seen in Chapter 2 that there are some similarities between the challenges posed by, respectively, L1 lexical acquisition and L2 lexical acquisition; and we shall see in Chapter 4 that such similarities extend into the operational sphere. We can therefore take it that most of what is said in the present chapter in respect of L1 lexical processing is also relevant to L2.

The chapter begins with a review of some of the available models of lexical processing and of the research evidence that they seek to account for. It then assesses the plausibility and the relevance to the lexicon of the notion that the mind is modularly organized. Finally, it explores connectionism in a lexical perspective.

Some models of the mental lexicon

A distinction is made by Garman in his (1990: 260ff.) discussion of lexical modelling between direct and indirect models. He compares the processes posited by the indirect type of model to those required to negotiate a dictionary or a library, each of which is internally organized in such a way as to facilitate two-stage access via a search procedure and then a retrieval procedure. Direct models, on the other hand, are predicated on one-stage access, the metaphor used by Garman in this case being that of a word-processing package which allows items stored by name to be accessed simply by the typing in of as many letters as are sufficient to identify the relevant name from among all the names available. We shall begin in this section by looking at two oft-cited and influential representatives of the direct kind of model – Morton's logogen model (see, e.g., Morton, 1964a, 1968, 1969, 1970, 1978, 1979; Morton & Patterson, 1980) and Marslen-Wilson's cohort model (see, e.g., Marslen-Wilson, 1980, 1987, 1989a, 1990, 1993; Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978) – before examining the best-known exemplar of the indirect type of model, Forster's (1976, 1979, 1981, 1989) search model of lexical access. We shall then conclude our brief trawl through lexical models by considering Levelt's (1989) 'blueprint for the speaker', which is actually not, or rather not solely, a model of lexical processing, but which ascribes a central mediating role to the lexicon and has accordingly been widely referred to in the context of discussion of the mental lexicon (see, e.g., Bierwisch & Schreuder, 1992; De Bot, 1992; De Bot & Schreuder, 1993; Gass & Selinker, 1994).

Morton's logogen model

The logogen model began as an attempt to account for Morton's (1961, 1964b) finding that there was a relationship between the distribution of lexical responses in sentence-completion tasks – involving items such as *He asked the way to the* _____ – ('transitional probability') and the time taken to recognize certain items in sentence contexts ('visual duration threshold'). This relationship is summarized by Morton as follows:

> any context which increases the probability of words in a generation situation would be expected to lower their threshold of recognition. (Morton, 1964a; reprint: Oldfield and Marshall, 1968: 152)

It is modelled as shown in Figure 3.1.

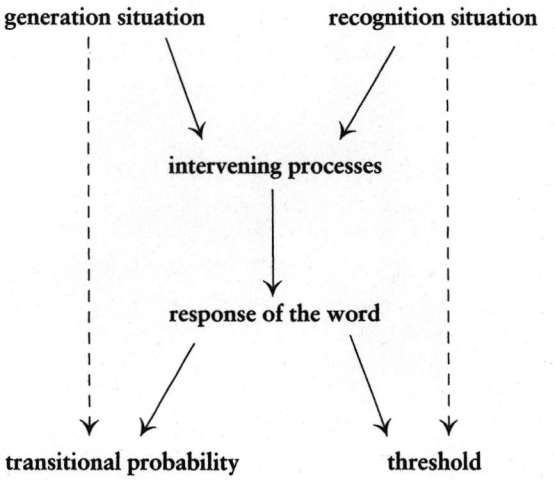

Figure 3.1. Morton's first attempt to model the relationship between transitional probability and visual duration threshold (after Morton, 1964a, Figure 1)

Morton postulates that when a lexical response becomes available there is an 'event' in 'a part of the nervous system' which he initially labels simply as 'neural unit' and to which he in his later writings applies the technical term 'logogen'. He attributes the following properties (P1–P4) to these neural units:

P1 When a unit fires, a particular word is available as a response.
P2 Each unit has a basic, relatively stable, level of activation.
P3 The level of activation can be increased by noise or by outside events.
P4 Each unit has a threshold; when the level of activation exceeds the threshold, the unit fires. (Morton, 1964a, reprint: 148)

Property 3 above refers broadly to context effects, which will be something of a leitmotiv in this chapter. The notion that a prior processing event can facilitate a subsequent processing event is a very familiar one in psycholinguistics and is the basis of the experimental technique of priming, defined by Aitchison as follows:

> A technique used in experimental studies, in which a person is prepared for a subsequent word or utterance. For example, the word *winter* might 'prime' the word *snow*, in that after hearing *winter* a person would be likely to recognize *snow* more quickly in a **lexical decision task** (deciding whether a sequence of sounds or letters is a word or not). (Aitchison, 1992: 72)

Clearly the closest relationship between two items is absolute identity, and, indeed, words prime themselves very effectively. That is to say, if a word is re-presented after an initial presentation, it will be recognized significantly more quickly than if the initial presentation had not taken place.

Morton's model evolved in various ways over subsequent years as more and more experimental and observational evidence was taken account of. In the version current in the late 1960s and early 1970s, there were just three components (see Figure 3.2):

- the logogen system, i.e., a collection of mechanisms – one for each word in a given individual's lexicon – specialized for collecting acoustic evidence (contributed by auditory word analysis), visual evidence (contributed by visual word analysis) and semantic evidence (from the cognitive system) concerning the presence of words to which the logogens correspond;
- the cognitive system, i.e., a collection of semantic information of various kinds, directly connected via a two-way link to the logogen system;
- the response buffer, – i.e. a component responsible for generating spoken or written word production, directly connected to the logogen system via a unidirectional link (logogen system → response buffer).

A basic principle of operation of the model is that any input will be likely to supply evidence to more than one logogen. For example, in the case of the processing of the printed word *cat*, 'the output from the visual analysis might include the attributes <three letter word>, <tall letter at the end>, <initial *c*>, <final *t*>, and so on' (Morton, 1970: 206). Such information is relevant not only to *cat* but to other words too. Accordingly, the attribute <three-letter word>, for example, will be expected to excite not only the logogen for *cat* but the logogens for all three-letter words. Hence the need for the model to incorporate thresholds. Among the fairly widespread excitation of logogens that is set off by a given input, it is necessary that one logogen – on the basis of all the available data – should reach such a level of excitation that it 'fires', in order that the appropriate word should be selected. In fact, Morton's (1970) conception was that there were two such thresholds, one controlling access to the cognitive system and the other controlling access to the response buffer.

One of the frequently observed phenomena addressed by the double-threshold idea and by the proposed architecture of connectivity between the components of the early logogen model was the

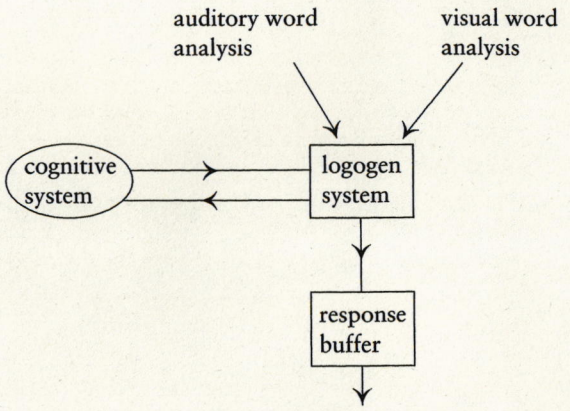

Figure 3.2 The essential components of the early version of the logogen model (based on Morton, 1979: 113, Figure 1, 138, Figure 5)

anticipatory nature of many deviations in reading aloud. When people read aloud they often produce errors that appear to be induced by contextual material which lies ahead of the point they have reached in their vocalizing of the text. The logogen model can explain this in terms of the possibility of words passing the cognitive system threshold before gaining access to the response buffer, the point being that such words can then be fed by the cognitive system directly back into the logogen system to influence (via the direct line between the logogen system and the response buffer) the (mis)reading of items in the response buffer.

Another point arising from the early logogen model has to do with the rate of decay of activation. Morton assumed that once a logogen had 'fired', activation relative to the word in question would have to diminish very rapidly; otherwise there would presumably be interference with the identification of subsequent items. Morton's (1968) suggestion was, in fact, that logogen activation levels after 'firing' returned to something like their original value in about one second. However, there is a further issue to be taken into consideration, namely the above-mentioned question of priming. Where priming effects manifest themselves in very short-term experiments involving lapses of no more than a second or so between initial presentation and re-presentation, they can readily be explained by reference to Morton's hypothesized time-scale for activation decay. But what of the longer-term priming effects which some of Morton's own experiments turned up and which other researchers have found to last for

many hours (see, e.g., Scarborough *et al.*, 1977)? Morton's answer to this point is summarized by M. Harris & Coltheart (1986) as follows:

> It is assumed ... that each time a logogen reaches its threshold, the value of that threshold is lowered; and this value then slowly drifts up towards what it had been, but never quite reaches the previous level. ... long-term priming effects are explained ... by assuming that after threshold has been reached activation dies down rapidly at first ... but does not quite reach the normal resting period: there follows a long period during which there is a slow decay of residual activation – a period measured in hours or even days. (M. Harris & Coltheart, 1986: 140–141)

Morton's original assumption was that the detectability of a given word would be enhanced across the board by any prior encounter with any related stimulus – identical, similar or connected, mental, spoken, written or pictorial. As Garman puts it:

> In Morton's model, evidence about the occurrence of a particular word comes potentially from all modalities, and these inputs are in a 'conspiratorial' relationship with one another ... and they all combine to lower what Morton calls the recognition threshold of the relevant stored forms. (Garman, 1990: 278)

Unfortunately for this view, some experimental evidence casts doubt on the notion of cross-modal priming. Thus, for example, findings from a study by Winnick & Daniel (1970) suggested that whereas reading a printed word aloud facilitated its later recognition in printed form, no such facilitation in this specific respect was brought about by naming a picture of the word's referent or by producing the word in response to a definition. Indeed, some of Morton's own work confirmed the absence of a cross-modal priming effect; for example, Morton's 1978 study failed to demonstrate facilitation of visual word identification by prior exposure to auditory versions of the words in question (cf. also Clarke & Morton, 1983). Accordingly, Morton was led to revise his model in such a way as to allow for separate, independent logogen systems for different types of input. The essential features of the revised version are presented diagrammatically in Figure 3.3.

Another respect in which the revised model altered the earlier conception relates to input-output connectivity. Whereas in the earlier model the connection between input and output was presented as an indirect one – via the logogen system – in the later version direct pathways are envisaged between both the input analysis processes (both auditory and visual) and the response buffer. This is to account for the ability to pronounce visually or auditorily

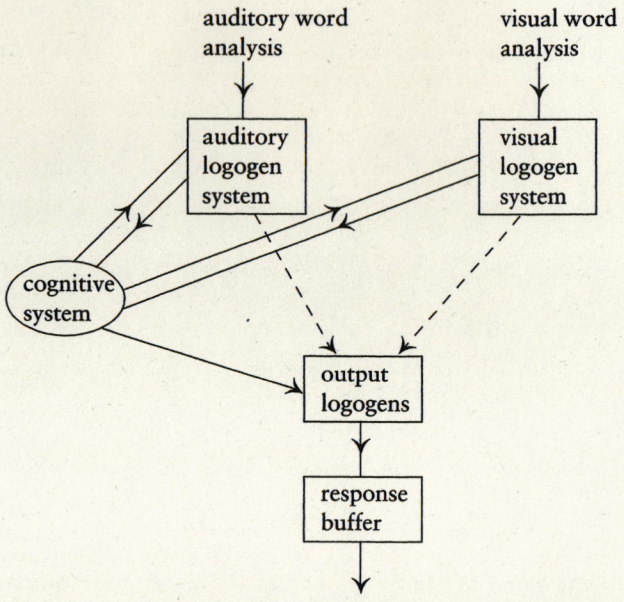

Figure 3.3 The essential components of the revised version of the logogen model (based on Morton & Patterson, 1980: 95, Figure 4.2b)

presented non-words on the basis, in the former case, of knowledge of graphological-phonological correspondences which transcends knowledge of individual words (cf. Campbell, 1983; Gough, 1972), and, in the latter case, of the replicating capacities of phonological working memory (see above, Chapter 2). Non-words clearly cannot trigger the firing of logogens since these latter have to correspond to real lexical items in storage and so, logically, there must be some way in which such non-words can be processed without reference to the logogen system. Moreover, it is also probably the case that in certain circumstances even real words are read aloud or pronounced without being processed at a level other than the purely formal level; who, for example, has not had the occasional experience of mechanically repeating or reading aloud a passage of a language they know without the least scintilla of interest in or comprehension of content being involved?

One should note that, as Morton has been perfectly ready to acknowledge (e.g., 1978), the above representation is still deficient in at least two respects On the one hand, it ought to incorporate a separate pathway for picture recognition and naming, including an

input 'pictogen' system (cf. Seymour, 1973). On the other, the output system is under-differentiated. To be fully complete, the model ought to be equipped with three separate output pathways – one for spoken output, one for written/printed output and one for graphic output (to capture the capacity to draw the referent(s) of lexical input).

Mention of separate output pathways brings us back to the whole question of the componentiality of Morton's model and of the degree of independence of certain of the components. We have seen that Morton's reading of the available experimental evidence persuaded him in revising the model to posit wholly unconnected auditory and visual systems. This particular aspect of the model has attracted criticism from some quarters, on the basis that, whilst the postulation of distinct auditory and visual systems seems sensible, to represent the two systems as entirely independent of each other goes beyond the evidence. For example, Garman (1990) raises the question of lexical-decision responses to items presented visually. He notes that correct *yes* responses can be seen as mediated via the visual system, but that correct *no* responses pose a problem for the model. He sketches a possible solution in terms of an 'access clock' which would, as it were, bring down a guillotine and signal *no* if no visual logogen were triggered within a given time, but again the evidence is awkward:

> ... this suggests that all such responses should be equivalently slow; it is therefore difficult to reconcile with the observed effect [i.e., the effect of slowing down lexical decision-making of homophone non-words (e.g., *sist*, resembling *cyst*)]. Such findings would seem to argue for a link between visual–auditory analysis ... but this is explicitly rejected in the model. (Garman, 1990: 285)

Emmorey & Fromkin (1988: 132), also, cite some psycholinguistic findings which appear to run counter to Morton's later view on the question of the unconnectedness of auditory and visual lexical processing:

- the facilitation of visual (real) word recognition when the word (e.g., *pale*) is preceded by a homophone prime (e.g., *pail*) (Humphreys, Evett & Taylor, 1982);
- faster identification of spoken words (from a list) rhyming with a cue word when the cue word and the rhyming word were orthographically similar (e.g., *glue, clue* versus *grew, clue*) (Seidenberg & Tanenhaus, 1979; Donnenworth-Nolan *et al.*, 1981).

On the other hand, they also cite evidence from a surface dyslexic, Kram (cf. Fromkin, 1985; Newcombe & Marshall, 1985), which supports the notion of the separation of orthographic and phonological representations (with interconnections). Following a brain injury, the patient in question seems to have substantially lost access to orthography. Thus, he pronounces *cape* as /sæpi/ and writes *kap* on hearing the word; he can define the word when he hears it but not when he sees it in print:

> If the orthographic representation is not listed separately from the phonological representation one would have to posit either impairment to orthographic representation of each lexical item, or a complex impairment of the multitude of connections to these representations leaving the pathways to the phonology intact. By positing separate sub-lexicons with interconnecting addresses ... the impairment is more simply explained. (Emmorey & Fromkin, 1988: 133)

The simpler explanation referred to is, of course, that the brain injury has resulted in the disruption of the connection which normally links the phonological sub-lexicon to the orthographic sub-lexicon. On the basis of this kind of evidence, Fromkin (1985) retains componentiality in her own model of the lexicon – which includes a phonological lexicon, an orthographic lexicon and a semantic lexicon – but posits a grapheme–phoneme conversion subsystem as well as a bi-directional link between the phonological and the orthographic components.

Marslen-Wilson's cohort model

A criticism of the logogen model which has not so far been mentioned is Forster's (1976) observation that it is difficult to see how this model can prevent the more frequent item *bright* being more available than the low-frequency target item *blight* in response to the input /blaɪt/, given that higher frequency implies higher levels of activation for the relevant logogen. As Garman (1990: 280, 286) points out, this problem is essentially about the difficulty of making precise statements about notions such as 'threshold' and 'activation level'. Marslen-Wilson's cohort model offers a possible answer to this problem, since it aspires to state exactly for each word where the critical activation level occurs.

The cohort model postulates a set of auditory word detectors which are activated by input from a spoken word and which go into operation as soon as the uttering of the word commences. As soon as the first sounds of the incoming item are processed, all the detectors for words beginning with that acoustic sequence – otherwise known as the relevant word-initial cohort – are fully activated. Each member of this cohort of word candidates then continues to monitor subsequent input, mismatches removing themselves progressively

from the running, until a single word candidate finally tallies with the input:

> Unlike logogens, these elements are assumed to have the ability to respond actively to mismatches in the input signal. Namely, at such point as the input diverges sufficiently from the internal specification for an element then that element will remove itself from the pool of word-candidates ... eventually only a single candidate remains. At this point we may say that the word is recognized. (Marslen-Wilson & Welsh, 1978: 56–57)

What this means is that, in contradistinction to the varying degrees of activation posited by the logogen model, the classic cohort model allows for just two states of activation for a particular item: on (for as long as it forms part of a cohort of word-candidates) or off (when it fails to be selected for the word-initial cohort or is eliminated from the cohort). One should perhaps add, however, that this very simple binary approach to activation levels has been complexified slightly in a more recent version of the model (Marslen-Wilson, 1987), which envisages that, instead of immediately eliminating themselves, non-matching members of a cohort will go into an activation decline in the absence of further bottom-up support.

The cohort model also in principle identifies the precise point – its 'uniqueness point' – at which a word is recognized. This can be illustrated by reference to the word *elephant* (/'elɪfənt/). The word-initial cohort for /'elɪ/ would include words such as *elevate* and *element* (though presumably not *elephantine* or *elephantiasis* because of the absence of primary stress on the first syllable in these words). However, at the point where the /f/ sound occurs the cohort will have only *elephant* and its inflectional variants (*elephants*, *elephant's*, *elephants'*) left, since no other word in English begins with the sequence /'elɪf/. This then is the uniqueness point for *elephant*. The system seems to have the advantage of maximal efficiency; thus, for *elephant* to be identified prior to the occurrence of /f/ would run the risk of occasioning cases of mistaken identity, whilst to wait for more phonemes to be uttered beyond that point would be inefficient insofar as it would increase recognition time to no purpose in terms of gains in accuracy levels. The system also makes possible the definition of a point at which non-words are recognized as such. This is the point at which the sequence of phonemes uttered fails to correspond to any word in the language in question. Thus, for English, the non-word recognition point in *bnoil* will be the occurrence of /n/, since no English word begins with the sequence /bn/, while in the case of *relationshif*, the critical point will coincide with the very last sound /f/, since until this is uttered the possibility of a match still exists.

The experimental evidence in favour of Marslen-Wilson's proposals is quite strong (see, e.g., Marslen-Wilson, 1978, 1984, 1987; Marslen-Wilson & Tyler, 1980; Tyler & Wessels, 1983). For example, it has been shown that the time taken to recognize non-words will be shorter where recognition points come early in words and longer where recognition points come late, even though the decision time is identical if measured from the point at which the critical phoneme is uttered (Marslen-Wilson, 1978). It has also been shown (*ibid.*) that in phoneme-monitoring tasks, where subjects have to check spoken words for the presence of a particular sound and press a button when they hear it, the reaction time from the point at which the target phoneme appears will be shorter when the phoneme occurs late in a word than when it occurs early. Marslen-Wilson's explanation of this latter result proposes that, instead of focusing on listening for the target sound, his subjects were primarily concerned to identify the incoming words and were then searching their phonological representations of the identified words for the presence of the phoneme in question. Accordingly, the time taken to detect the presence of the target phoneme was dependent on the time taken to identify a given word, which in turn depended on the position in the word of its uniqueness/recognition point:

> When a target occurs *late* in a word, it is likely to occur *after* the recognition point. Consequently, the word will often have been identified before the target has even occurred, and so reaction times will be short. In contrast, when a target occurs *early* in a word, it is likely to occur *before* the recognition point. Consequently, such a word can be identified only after the subject has heard phonemes occurring later than the target phoneme, and so reaction times will be long. (M. Harris & Coltheart, 1986: 161–162)

It will be recalled that the original (and abiding) inspiration of the logogen model was Morton's interest in context effects. This interest is very much shared by Marslen-Wilson and his collaborators. The bottom-up aspects of the cohort model which have been discussed so far constitute only one dimension of its representation of lexical processing, the other being everything that might go under the heading of contextual contributions. It is clear that in normal language use we are not usually called upon to process words *in vacuo*; individual lexical items are typically embedded in syntactico-semantico-pragmatically coherent concatenations of other lexical items. The cohort model, like the logogen model, assumes that available contextual information has a facilitatory impact on lexical processing. However, whereas the logogen model seems to suggest that context effects are mediated by a semantic component (the

'cognitive system') separate from, though connected to, the logogen systems, the cohort model posits that each and every entry in the mental lexicon is equipped with inferential procedures:

> each word would have built into its mental representation not simply a listing of syntactic and semantic properties but rather sets of procedures for determining which, if any, of the senses of the word were mappable onto the representation of the utterance up to that point. (Marslen-Wilson & Tyler, 1980: 31)

The way in which semantico-pragmatic information is seen as being used in the cohort model is essentially an 'on-line' view of things. That is to say, the notion that contextual factors pre-select words is rejected (Marslen-Wilson & Tyler, 1980) on the basis that context-driven pre-selection would in fact be a highly inefficient manner of proceeding, given the open-endedness and unpredictability of even normal everyday language use. Reflecting on this point, we might consider the following exchange:

A: Shall we go for a drink down at the Hat and Feathers?
B: No. I feel like seeing a play. Let's go to the Theodore Hotel. The Linthorpe Players are putting on a really hilarious Tom Stoppard play there in the function room. It should be a good laugh.
A: Tom Stoppard at the Theodore, eh? OK. I feel like a lager. They have a really great selection of lagers at the Theodore.

In this exchange, plumping immediately for what might have seemed the contextually most likely word would probably have led A into thinking that B was suggesting a visit to the theatre and would have led B into thinking that A was desirous of a laugh. In the light of this kind of consideration, Marslen-Wilson & Tyler propose that contextual information has no impact on the selection of the word-initial cohort, but that, once the cohort has been established, word candidates which are inconsistent with the context can begin to be deactivated. Thus, the cohort activated on the basis of the /lɑː/ of *lager* (/'lɑgə/ – assuming the interlocutors were speakers of Standard British English) would have included not only *laugh* (/lɑːf/), but also items such as *lamé* (/'lɑmeɪ/), *larva* (/'lɑːvə/) and *lath* (/lɑːθ/). According to Marslen-Wilson & Tyler's proposals, the contextually implausible *lamé*, *larva* and *lath* would have immediately begun to be deactivated, whereas the onset of the deactivation of *laugh*, a contextually highly plausible item, would have had to await the occurrence of /f/ in the input and the recognition of the divergence between this phoneme and the /g/ of the input sequence.

The evidence cited by Marslen-Wilson and his colleagues (e.g., Marslen-Wilson & Welsh, 1978) in favour of a role for context in lexical processing comes not only from their well-known speech-shadowing experiments but also from word-monitoring and rhyme-monitoring studies. In speech-shadowing tasks, participants are required to listen over headphones to a passage of text read aloud and are asked to reproduce it faithfully with as short a time-lag as possible. In some of the studies, words in the original passage were deliberately mispronounced. For example, *tragedy* would be pronounced as *travedy*. Very often subjects replaced such deviant items with the correct versions of the words in question, and in about 50 per cent of cases the corrections effected were in the nature of fluent restorations, that is to say, the substitution of the correct versions of the mispronounced words was not associated with any faltering or hesitation in the flow of the repetition. Fluent restoration is taken to be an indication that a decision regarding the target word has been reached on contextual grounds prior to and irrespective of its formal recognition point. In support of this interpretation one can cite Marslen-Wilson's (1975) finding that fluent restorations were offered markedly more frequently in contexts of normal coherent and cohesive prose than where there was any kind of syntactic or semantic dissonance between the mispronounced item and its linguistic environment. One can also cite Marslen-Wilson & Welsh's (1978) finding that fluent restorations occurred far more often when a word was highly predictable from context than when it was only moderately predictable. Such results are interpreted as follows in terms of the model: the more contextually predictable a word is, the shorter the sequence of sounds required to reduce the cohort to a sole candidate – with the attendant higher probability that the mispronunciation will occur in an unanalysed portion of the word and so will remain undetected.

With regard to word monitoring, this task requires subjects to monitor linguistic material for the presence of a particular target word, pressing a button as soon as they perceive the word in question. Marslen Wilson & Welsh's (1978) subjects were presented auditorily with sentences of two types: (1) normal coherent prose and (2) syntactically licit but semantically anomalous prose. In a third condition, (3), subjects were asked to monitor randomly ordered strings of words. The mean reaction times for the three conditions were as follows:

1 normal coherent prose	273 milliseconds
2 syntactically licit but semantically anomalous prose	331 milliseconds
3 randomly ordered strings of words	358 milliseconds

What is interesting about these results is not only that the decreasing support offered by context in the three conditions corresponds linearly to increasing reaction times but also that the time taken to recognize words in normal prose contexts (273 milliseconds) is nearly 100 milliseconds less than the average time taken to utter the words in the passage in question (369 milliseconds). This clearly demonstrates that in context, words are recognized on the basis of much less than their full form. Moreover, Marslen-Wilson & Welsh estimate that about 75 milliseconds of the 273 constitute the normal, unavoidable, lapse of time between the identification of the target and the pressing of the button, leaving about 200 milliseconds of actual processing time. Now, it turns out that there is evidence to suggest that in the processing of words in isolation, an average of 29 words are still present in the cohort after 200 milliseconds' worth of processing. When one compares this with the single item arrived at after 200 milliseconds' processing in a coherent, meaningful context, one cannot but acknowledge the plausibility of the notion of on-line contextual influence.

The rhyme-monitoring results (Marslen-Wilson, 1980) tend in the same direction. In this case subjects were asked to press a button on hearing a word that rhymed with a particular stimulus word. Again, the material to which subjects were required to attend was presented in three conditions: (1) normal coherent prose and (2) syntactically licit but semantically anomalous prose, (3) randomly ordered strings of words. It was found that reaction times in rhyme monitoring were approximately 140 milliseconds longer than reaction times in word monitoring. This was interpreted as suggesting that subjects identified words first and only then decided about their rhyming possibilities, this latter decision accounting for the additional 140 milliseconds. The results also showed that with normal prose the later the target rhyme cropped up in the context the more speedily it was identified; in the case of semantically anomalous prose, the same effect was discernible but to a much lesser extent, and in the randomly ordered strings condition, the effect was absent. These findings were read as further evidence in favour of a role for contextual constraints, the point being that in normal prose the greater the amount of material preceding the target, the greater the specificity of semantico-pragmatic and syntactic constraints on word choice and the greater the number of word candidates that can be deactivated in the light of these constraints. In the case of the syntactically licit but semantically anomalous prose, meaning-related constraints did not operate but syntactic constraints still did. In the case of the randomly ordered strings of words, there were absolutely no contextual constraints to accelerate cohort reduction.

The question of context effects will recur later in the chapter. To return for the present to the form-based aspects of the cohort model, these have come under critical scrutiny from a number of quarters. Garman (1990) calls into question what others (Matthei & Roeper, 1983: 39ff.) have called the 'beads on a string' view of speech perception that Marslen-Wilson's proposals appear to incorporate:

> the notion of segmental elements ... arriving at the ear over time is certainly oversimplified ... since any 'time slice' through the acoustic signal shows evidence of preceding and succeeding elements. The auditory perception of this signal is therefore not susceptible of discrete judgments of a very precise nature concerning the point at which particular elements 'arrive'. (Garman, 1990: 288)

However, nothing very crucial seems to hang on this objection, which, as Garman acknowledges, 'tends in the direction of recognition points that might actually be in advance of the segmentally defined uniqueness point – by some very small factor (*ibid.*).

Garman goes on (*ibid.*: 288–289) to cite Marcus & Frauenfelder's (1985) suggestion, which they support with numerous references to empirical studies, that speech-sound processing is probabilistic:

> it seems unlikely that such categorical decisions can be made with the noisy and ambiguous signal which is speech. ... Incoming phonetic information cannot always be categorically recognized solely on the basis of the acoustic signal. ... recent data ... supports the idea that phonetic information is evaluated probabilistically rather than categorically during the process of word recognition ... (Marcus & Frauenfelder, 1985: 164)

Marcus & Frauenfelder therefore do not see word recognition as wholly dependent on or exactly contemporaneous with the point at which the item in question diverges by one phoneme from all other items. Rather they claim that subsequent deviation between the target item and other items in the cohort also has to be referred to in arriving at a definitive recognition. They show that, on average – at least in English – over the six phoneme positions following the uniqueness point, deviation between the target item and other candidates increases more or less linearly at a rate of about 0.5 phonemes per position. This means that, if their proposals are correct, the statistical properties of the (English) lexicon would in any case allow words to be recognized very quickly after the strictly defined uniqueness point, which is broadly consistent with the evidence supporting the notion that recognition occurs around the

same time as the occurrence of the uniqueness point (e.g., Marslen-Wilson, 1984; Tyler & Wessels, 1983). This criticism differs from Garman's in positing a recognition point slightly later than instead of slightly earlier than the uniqueness point, but, again, it does no real damage to the model, especially since Marslen-Wilson now takes a fairly flexible line regarding the organization of his model (e.g., 1987, 1989a, 1990, 1993) and, in particular, accepts that input continues to be monitored beyond uniqueness points and that the deactivation of word candidates is reversible.

Finally, there have been some questions raised about the importance attributed to the beginnings of words in the cohort model. Emmorey & Fromkin (1988), while acknowledging that there is a fair amount of evidence in favour of phonological organization by initial segment, also point to some evidence suggesting that ends of words also have some importance in phonological processing. On a somewhat different but related tack, Aitchison notes (1994: 218) that the earliest version of the cohort model 'required undistorted acoustic signals at the beginning of the word' and could not cope with a situation of uncertainty in this position: 'if a wrong decision was made, the wrong cohort would be activated'.

Emmorey & Fromkin (1988) cite the following evidence indicating that words are most easily accessed via their beginnings:

- the fact that subjects in a 'tip of the tongue' state can often access the initial sound or syllable of the word they are looking for even when all else deserts them (R. Brown & McNeill, 1966);
- the fact that patients suffering from anomia (i.e., word-finding problems) are often able to access the word they need if they are given the relevant first segment or syllable (Benson, 1979);
- that fact that in a timed test, subjects have been able to come up with many more words sharing initial segments than having any other portion in common (Baker, 1974).

As for arguments cited in support of a role for final parts of words, these include:

- the greater frequency of misperceptions on ends of words than on medial portions (Browman, 1978);
- the lesser frequency of speech errors on ends of words than on medial portions (Cutler & Fay, 1982);
- the greater accessibility in 'tip of the tongue' states of final compared with medial segments (Brown & McNeill, 1966);
- the greater difficulty of producing words sharing medial vowels than of listing rhyming words (Baker, 1974).

Emmorey & Fromkin are cautious in their interpretation of such findings:

> It may be the case that words are listed by final rhyme structure or final (stressed) syllable ... But these facts can be accounted for by processing strategies separate from the order of listing, or by 'recency effects' found in many memory experiments, i.e. the end of a word is heard more recently and thus might be more easily remembered. (Emmorey & Fromkin, 1988: 128–129)

The implications for the cohort model of the facts regarding the greater memorability/accessibility of ends of words relative to middles of words are, as Emmorey & Fromkin suggest, rather unclear. As for Aitchison's point, she herself recognizes (1994: 218) that it does not hold for the more recent versions of the model, which have become more fluid in their organization.

Forster's search model

We come now to an example of an indirect model, in which access is represented as a serial process involving first a search for a matching element in the relevant mode and then a guided retrieval of the full word. As has already been indicated, it is possible to compare this two-stage process to what happens when we look up a word in a dictionary or look for a book in a library. The first stage of consulting a dictionary (of a language in an alphabetic writing system) is the scanning of head-word forms listed in bold type on the left-hand side of each column until we find the one that matches our target item. We can then go on to check the full entry for pronunciation details, meaning(s), morphosyntactic specifications, stylistic information, etc. In a library we go first to whichever catalogue has a point of departure which coincides with the information we already possess about the book we require (author, title, subject, etc.). Having located the relevant reference in the appropriate catalogue, we can then use the shelf-mark associated with the item to guide us to the actual book on the library shelves.

Of these two analogies, the latter may be the more apposite (see, e.g., Matthei & Roeper, 1983: 188–189), given that we come to the task of lexical access from different starting points (phonological, orthographic, semantic) on different occasions just as we approach the task of finding books in libraries with different kinds of information available to us at different times. In Forster's model, the initial search is carried out with the help of a number of peripheral access files, one organized along phonological lines, one organized according to orthographic properties, one organized on a

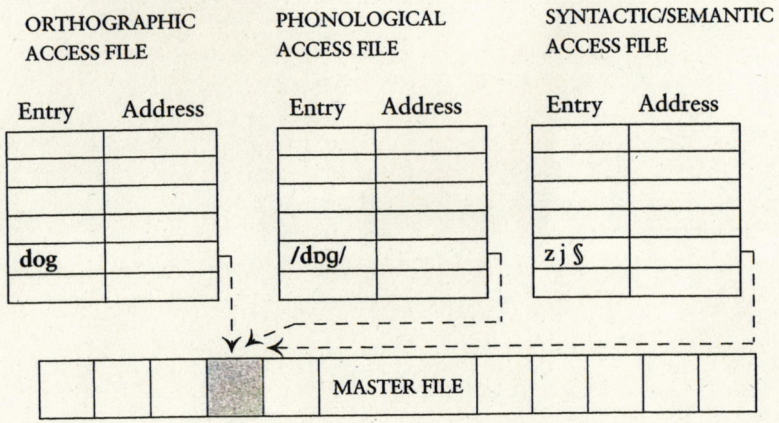

Figure 3.4 The essential components of Forster's search model (based on Forster, 1976: 268, Figure 4)

syntactico-semantic basis, etc. These correspond to the different library catalogues. These peripheral files contain listings of entries in the respective modes and also pointers (corresponding to shelf-marks) to the location of each entry in its complete form in the unitary master file (corresponding to the library shelves). The broad lines of the model are set out diagrammatically in Figure 3.4.

Thus, if one is processing spoken language receptively, one goes first, on this view, to the phonological access file; if one is processing written language receptively, one goes first to the orthographic access file; and if one is producing language on the basis of particular meaning intentions, one goes first to the syntactic/semantic access file. Once the unitary master file has been accessed, it can facilitate any kind of further operations on the word in question – whether these be in the realm of speaking, writing or understanding. It should be noted, however, that the peripheral access files are seen as absolutely autonomous in the sense that the information they contain is represented as strictly limited to the specific modality with which they are respectively concerned and that no connectivity is envisaged between the different access files. In general the model has the advantage of accounting for the intuition 'that an adequate lexicon must permit diversity of access but unity of storage' (Garman, 1990: 267). In other words, it seems to capture the fact that, while we are aware of coming to a given word, say *rain*, via a variety of routes – hearing it, reading it, processing its meaning – we do not usually

consider /reɪn/ (phonological form), *rain* (orthographic form), and 'rain' (meaning) to be three different words but rather think of them as different aspects of or indicators of the same word – which seems to be precisely captured by Forster's representation of lexical relationships. On the other hand, as we have already seen during the course of discussion of the logogen model, to posit total autonomy for phonological and orthographic access respectively goes further in the direction of the separation of processing components than appears to be warranted by the evidence.

The access operation is represented as proceeding as follows. The properties of the stimulus form or meaning cause the search to be concentrated in a particular area or 'bin' of the relevant peripheral access file, but within that 'bin' items are checked serially in order of frequency until a match is found for the specifications of the stimulus. There is some vagueness in relation to the nature of the properties that are critical in the initial guiding of the search and, correspondingly, in relation to the bases on which the different 'bins' are constituted. With regard to phonological access, for example, presumably account needs to be taken (for reasons mentioned earlier) of initial segments, but (again for reasons discussed above) final segments may also play a role, as well, perhaps, as elements such as stress patterns and syllable structure. As for the suggestion that once the appropriate 'bin' has been targeted, subsequent search processes follow a frequency order, this has some intuitive appeal and seems to accord with the available evidence, but, as we shall see a little later, it is not without its problematic aspects. With regard to second-stage operations within the master file, it is envisaged that cross-referencing may occur at this level between words which are closely associated.

Before pursuing the questions raised by the proposals regarding the master file, however, let us return to the issues that have been raised in respect of the access files, namely those of autonomy and the role of frequency. Concerning the lack of provision in the model for connectivity between the peripheral access files, this lack fails to account in full for the evidence in respect of non-words. The fact that on lexical decision tasks it takes longer to reject a phonotactically licit non-word than it does to accept a real word (see, e.g., Gough & Cosky, 1977) can be explained in terms of the real word having an entry in the various files, and thus a *terminus ad quem* for the search processes, as opposed to the absence from the system of entries for non-words, with all that this implies for the necessity of a totally exhaustive (and futile) search. The fact that phonotactically illicit non-words are rejected more rapidly than phonotactically licit words and indeed more rapidly than real words are identified (see *ibid.*) is

also explicable in terms of the model: whereas the last two categories of item cause a search to begin in a particular 'bin' of the phonological access file, the phonotactic illegality of the first category means that no 'bin' is found to correspond to the general properties of the stimulus, which in turn means that no search of entries can actually take place. However, the model provides no explanation for the fact that we are able to read non-words aloud and to attempt orthographic transcriptions of non-words we hear. These possibilities indicate the necessity for at least some system of grapheme–phoneme conversion. In addition, the fact that we are able to replicate pronunciations of non-words argues for a direct non-lexical link between auditory input and articulatory output, since a link via the file system is excluded by the absence of entries in the files corresponding to the non-words in question. Moreover, as we saw in connection with Morton's proposals for complete separation of phonological and orthographic access processes, there are various kinds of evidence which suggest that connections between phonological and orthographic processing exist at a lexical level too.

As regards the role of frequency, we need to ask whether what is being referred to is frequency of occurrence in the input generally, frequency of occurrence in the input attended to or frequency of output. We also need to be aware that frequency is modality-specific, that 'the frequency of the written form of a word may be different from the frequency of its spoken form' (Matthei & Roeper, 1983: 189). A further consideration in this context (cf. Matthei & Roeper, 1983: 184–185) is that word frequency broadly correlates with recency of occurrence in the input and output and that a frequent word is also likely to have been acquired early. How are we to know therefore whether the critical factor determining speed of access is relative frequency of occurrence as such (as not only Forster but also Morton – e.g., 1970 – would claim) or relative recency of occurrence (cf., e.g. J. R. Anderson, 1976; Scarborough, Cortese & Scarborough, 1977)?

So much for the characteristics of the peripheral access files; what now of the master file? This collection of individual (fully specified) lexical items is seen by Forster as having to contain some provision for connections between the items in question. One argument in favour of allowing for such inter-relationships is furnished by the earlier-discussed phenomenon of priming. The priming study cited in the context of Forster's proposals is that of Meyer & Schvaneveldt (1971). This was an experiment based on a lexical decision task in which items were visually presented in pairs. The results revealed that reaction times for the second member of the pair were shorter if

this was semantically related to the first item. For example, the word *nurse* was more rapidly reacted to when preceded by *doctor* than when preceded by *table*. Forster's model tries to account for 'semantic priming' of this kind by positing cross-references in the master file between words that are related in meaning. With regard to the above example, the idea is that the retrieval of the fully specified item *doctor* will cause the item *nurse* to be processed via a direct search path within the master file without the necessity for a return to the relevant peripheral access file in order for this latter item to be dealt with 'from scratch'.

An alternative possibility suggested by Matthei & Roeper (1983: 189–190) and by Emmorey & Fromkin (1988: 143) within the general framework of Forster's model is that there might be two levels of semantic processing, a linguistic or lexical level and a non-linguistic, encyclopedic level:

> Another possibility would be to assume that the master file does not contain very much information about the meanings of words, just a sort of bare-bones specification of meaning. The entries would then be assumed to be linked in some way to another big file of information about the world, how it is structured and how it works. (Matthei & Roeper, 1983: 189–190)

Semantic cross-referencing according to this view would proceed via the general knowledge store. Thus, the master file item *doctor* would trigger reference to a constellation of information about doctors, including the information that they often work in hospitals alongside nurses, which would in turn trigger reference back to the master-file item *nurse*.

Unfortunately, neither Forster's explanation of semantic priming nor the general knowledge store perspective is very satisfactory. To take the latter first, this depends on the possibility of making a distinction between linguistic and non-linguistic or 'pragmatic' meaning. Emmorey & Fromkin are inclined to see this as unproblematic:

> That such a distinction exists seems to be unquestionable, as can be seen by the simple example of the difference between knowing the meaning of the word 'water' and knowing that its chemical structure is H_2O. Obviously one can know the first without knowing the second. (Emmorey & Fromkin, 1988: 143)

Are things really that simple? Let us look more closely at the example given by Emmorey & Fromkin. It is surely possible to see knowing the chemical composition of water in terms merely of having fuller access to the 'lexical' meaning of the word *water*. Such knowledge allows one, for example, to accept as semantically non-anomalous

sentences such as 1 and 2 below in much the same way as one accepts 3 and 4 (whose acceptability would tend to be seen as linguistically based by semanticists of the Emmorey & Fromkin school):

1 Today we shall consider water and other hydrogen compounds.
2 Fish breathe water, just as we breathe air.
3 This is water, and here are some other liquids.
4 We drank some water.

Compare:

1a * Today we shall consider table-salt and other hydrogen compounds.
2a * Fish milk water, just as we milk cows.
3a * This is water, and here are some other solids.
4a * We ate some water.

From another – not incompatible – point of view, it is entirely possible to regard the 'everyday' or 'basic' meaning of *water* as simply a distillation into unconscious automaticity of what one knows 'pragmatically' from one's most frequent experiences with the substance to which the term most often relates.

Because of the relative frequency of uses of *water* which do not allude to the chemistry of its denotatum, one is, of course, able to understand and appropriately use the term in most contexts without any chemical knowledge, but is this qualitatively different from being able to deal with a polysemous word in many contexts without knowing more than one of its meanings? To stay with the example of *water*, unless one knows that this item can in certain contexts be applied to brine, perfumed alcohol, and amniotic fluid, one will make little sense of the following:

5 Water, water, everywhere,/ Nor any drop to drink.
6 He reeked of Cologne water.
7 When her waters broke, she knew the time had come to make a phone call.

However, not having these meanings of *water* at one's disposal will not prevent one from getting by without difficulty in the majority of situations where the word crops up.

Emmorey & Fromkin's statement that the distinction between linguistic meaning and pragmatic meaning is an obvious one is also undermined by the fact that it is a matter about which theoretical linguists have doubts and differences (cf. Maclaran, 1983). Within the Chomskyan school, for example, whereas some followers of Chomsky have taken essentially the Emmorey and Fromkin line (see, e.g., N. Smith & Wilson, 1979), Chomsky himself has consistently expressed worries over the possibility of making the distinction in question:

> It is not clear at all that it is possible to distinguish sharply between the contribution of grammar to the determination of meaning, and the contribution of so-called 'pragmatic considerations', questions of fact and belief and context of utterance. (Chomsky, 1972: 111)

> Do the 'semantic rules' of natural language that are alleged to give the meanings of words belong to the language faculty strictly speaking, or should they be regarded perhaps as centrally-embedded parts of a conceptual or belief system, or do they subdivide in some way? (Chomsky, 1980a: 62)

With regard to Forster's notion of direct cross-referencing within the master file, this has been called into question by some of Forster's own findings. Forster reports (1976) the results of an experiment involving pairs of words of different levels of frequency. The pairs in question were composed of two high-frequency words, two low-frequency words, one high-frequency word and one low-frequency word (so ordered), or one low-frequency word and one high-frequency word (so ordered). Forster's hypothesis was that in the mixed pairs, where the two items were semantically related, the frequency of the first item would determine speed of access. That is to say, he hypothesized that a high frequency first member of a pair would swiftly find a match in the relevant area of the relevant access file and trigger the retrieval of the fully specified item in the master file, and that direct cross-referencing within the master file would mean that the related low-frequency second member of the pair item would also be rapidly found despite its low frequency, because reference would not need to be made back to the access files where frequency was a factor. In the case of mixed pairs beginning with low-frequency items, the processing of both items would be slow because of the initial slowness of the search through the access file. Alas for this elegant hypothesis, the results of Forster's experiment show low-frequency items slowing down processing wherever they occur. It is by no means clear how such results are to be interpreted in the terms of the model.

Commenting on these and other similarly unclear results, Garman raises some fundamental questions about the soundness of the concept of cross-references in the master file:

> do they effectively provide a separate search mechanism, and, if so, does this wastefully duplicate the function of the semantic access file? If there is no duplication, then what are the conditions under which one or the other

search will be carried out? Are there sound-structure cross-references in the master file, and how far might these duplicate the operation of the phonological access file? (Garman, 1990: 270–271)

Such questions in turn lead Garman to put under close scrutiny the whole idea of a distinction between access files and master file and thus the very notion of two-stage lexical processing.

Levelt's 'blueprint'

So far we have been looking at models which are explicitly focused on the lexicon. Levelt's model, which is the subject of the present section, falls into a rather different category insofar as it seeks to address all aspects of language processing. However, as has already been indicated, its lexical dimension is particularly highlighted by its creator, who has continued to evince a special interest in lexical processing (see, e.g., Levelt, 1993a, 1993b). The work in which the model is elaborated (Levelt, 1989) bears the title *Speaking: from intention to articulation*, and, true to this title, the primary perspective of the model is a productive one, although receptive aspects of processing are not entirely ignored.

The model is represented diagrammatically as shown in Figure 3.5. There are in Levelt's conception two categories of component, declarative and procedural. The former – represented in the diagram by the curvilinear elements – deal in 'knowledge that', knowledge as facts – whereas the latter – represented in the diagram by the rectilinear elements – deal in 'knowledge how', knowledge of the steps to be taken in order to achieve particular goals (cf. J. R. Anderson, 1983; Ryle, 1949). Declarative knowledge required for language processing, according to Levelt, includes general information about the world (encyclopedia), information about the specifics of particular situations (situational knowledge), and information about stylistic appropriacy relative to specific sets of circumstances (discourse model). Also located under the rubric of declarative knowledge is lexical knowledge, both semantico-grammatical (lemmas) and morphophonological (forms). As far as the procedural components are concerned, these include:

- the Conceptualizer, responsible for message generation, microplanning and monitoring;
- the Formulator, responsible for giving the pre-verbal message a surface syntactic and phonological shape;
- the Articulator, responsible for executing as overt speech the phonetic plan emerging from the Formulator;

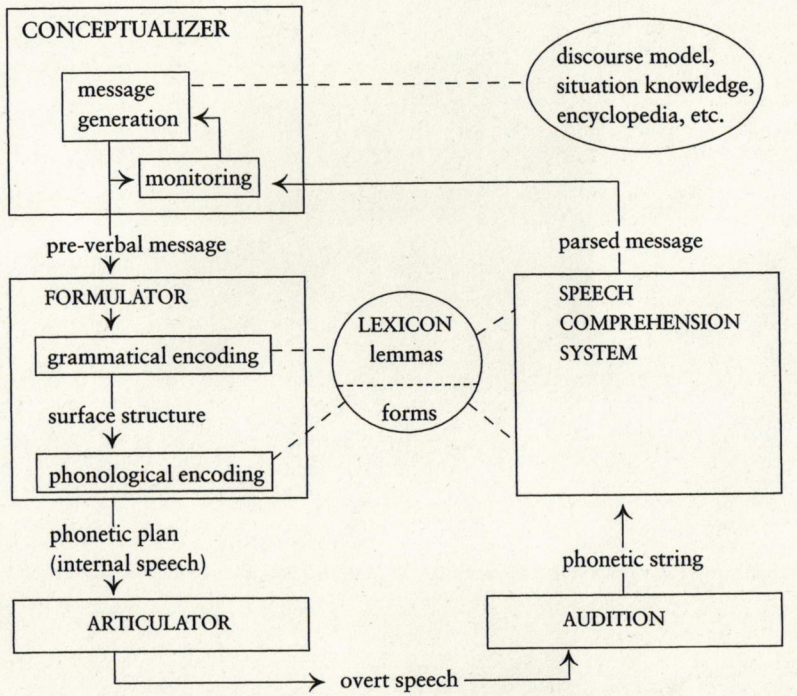

Figure 3.5 Levelt's blueprint for the speaker (based on Levelt, 1989: 9)

- the Audition component, responsible for analysing the speech signal into sound segments;
- the Speech Comprehension System, responsible for making semantico-grammatical sense of phonetic strings received.

To return to the lexical component, this, as has been mentioned, contains on the one hand lemmas and on the other hand forms (alternatively labelled lexemes in Levelt's terminology). A word's lemma is that which specifies its basic meaning, its syntactic category, its conceptual argument structure, its grammatical profile (e.g., in the case of a verb, whether or not it takes a direct object and whether or not it can take a dependent clause (relations to COMP), and its 'diacritic parameters' of variation (tense, aspect, mood, etc.)). The lemma also includes a 'lexical pointer' to the precise place in the lexicon where morphological and phonological information about the word in question is located. The following is Levelt's (1989: 191) outline of the lemma for *give*:

give: conceptual specification:
CAUSE (X, (GOposs (Y, (FROM/TO (X,Z))))) [i.e., X cause Y to pass from X's possession to Z's possession]
conceptual arguments: (X, Y, Z)
syntactic category: V
grammatical functions: (SUBJ [subject], DO [direct object], IO [indirect object]) relations to COMP: none [i.e., does not introduce dependent clauses beginning with complementizers such as *that*, *whether*, *if*, etc.]
lexical pointer: 713 [an 'address' chosen at random and arbitrarily coded]
diacritic parameters: tense
aspect
mood
person
number
pitch accent

As far as the lexical forms are concerned, these specify the precise morphological information that is necessary in order for phonological encoding to be able to take place – prior to the operation of the Articulator.

The part played by the lexicon in speech production is seen by Levelt as absolutely central; for him the whole set of formulation processes is lexically driven:

> This means that grammatical and phonological encodings are mediated by lexical entries. The preverbal message triggers lexical items into activity. The syntactic, morphological, and phonological properties of an activated lexical item trigger, in turn, the grammatical, morphological and phonological encoding procedures underlying the generation of an utterance. (Levelt, 1989: 181)

He refers to his assumption that the lexicon is the mediator between conceptualization and grammatical and phonological formulation as the lexical hypothesis:

> The lexical hypothesis entails, in particular, that nothing in the speaker's message will *by itself* trigger a syntactic form, such as a passive or a dative construction. There must be mediating lexical items, triggered by the message, which by their grammatical properties and their order of activation cause the Grammatical Encoder to generate a particular syntactic structure. (*ibid.*)

This view of the lexicon as mediator sits well with the evidence discussed in Chapter 1 of the interpenetration between lexis and grammar. On the other hand, the separation in the model of lexical meaning from encyclopedic knowledge is, as we have seen in the context of the discussion of Forster's model, rather more problematic.

A further question arises in relation to the representation of lexical knowledge as purely declarative. It is, after all, a commonplace among linguists that the lexicon contains word-formation or lexical-redundancy rules (see, e.g., Radford, 1981, Chapter 4; Cruse, 1986: 50), which make possible the generation of a potentially infinite number of new lexical forms. Since lexical creativity based on such possibilities involves a process and a goal, the psychological correlates of lexical-redundancy rules must surely be classed as procedural knowledge. Equally, from a receptive point of view, the attempt any reader or hearer will typically make to assign meaning and function to novel word forms cannot be a matter of the Speech Comprehension System accessing static lexical facts, but must, one would have thought, involve lexicon-internal consultation and cross-referencing processes – which again implies procedural knowledge. Indeed, the whole range of evidence discussed in earlier sections on context effects and priming seems to point in the direction of the lexicon being a dynamic rather than a static entity.

A further issue concerning the lexical dimension of Levelt's model is raised by Bierwisch & Schreuder (1992) and De Bot & Schreuder (1993), who see the necessity for an intermediary module ('Vbl') between the Conceptualizer and the Formulator. Their reasoning in favour of their proposal is that 'the conceptual structure presents the Formulator with fragments that exceed the size of one lemma's semantic representation', and that therefore there has to be a unit 'responsible for cutting up the fragment in chunks that can be matched with the semantic information associated with the different lemmas in the mental lexicon' (De Bot & Schreuder, 1993: 193). A further consideration they raise in this connection has to do with the fact that an individual may have more than one language and therefore more than one lexicon at his/her disposal and that different languages lexicalize the world differently. If, therefore, they argue, we assume (as Levelt does) that the pre-verbal message is language-neutral, 'then we are forced to assume that the Vbl function is sensitive to different lexicalization patterns and somehow "knows" which lexicalization pattern to choose' (*ibid.*: 195).

The relationship between the Conceptualizer and the Formulator is interesting from another point of view – namely that of the question of the degree of autonomy attributed to the various components of the model. Towell & Hawkins summarize the situation in this regard as follows:

> The production process is thought of as composed of relatively autonomous stages as specified by the boxes in the diagram. Processing has to be both

incremental and parallel to allow for the speed at which it must take place. Together this means that different parts of the message may be being processed at the same time (parallel), different parts of the message may be at different stages of the production process (incremental) and that these will not interfere with one another. In order to ensure that the message can nonetheless be delivered in the right order despite this flexibility there are buffer areas between the units which can delay delivery until the order is correct. (Towell & Hawkins, 1994: 168)

According to Levelt, the Conceptualizer learns what kinds of message can be coped with by the forms of a given language during the language acquisition process, so that during acquisition there has to be feedback between the Formulator and the Conceptualizer. Once the relevant ground-rules relative to the presentation of information to the Formulator have been established, however, such feedback is no longer required:

it is no longer necessary for the Conceptualizer to ask the Formulator at each occasion what it likes as input. In short, the systems have become autonomous. (Levelt, 1989: 105)

Such a view is reminiscent of Bever's (1981) suggestion that the 'psychogrammar' develops in interaction with perception and production processes but that once it has been established it becomes 'decoupled' from such processes and thus unavailable for further development. One objection to this kind of approach is that it is difficult to isolate a precise point at which the 'decoupling' can plausibly take place:

Most evidence suggests that ... the acquisition of pragmatic rules and lexis continues well into adulthood – being bounded perhaps only by death – and that even morphosyntactic and phonological development may persist well beyond puberty. This implies that communication between the processing systems and the psychogrammar has to remain open throughout an individual's life ... (Singleton, 1989: 220–221)

The same kind of argument can – *mutatis mutandis* – be applied to Levelt's notion that the Conceptualizer becomes 'decoupled' from the Formulator. If one accepts the view that first language acquisition continues into and through adulthood, and if one takes into account the fact that individuals also learn other languages at various stages in their life, the logical conclusion in Levelt's terms is that the autonomy of the Conceptualizer and the Formulator with respect to each other is not absolute even as far as the mature native speaker is concerned.

Levelt's model has in common with the earlier models discussed the fact that it is not only concerned with what the mental lexicon

can plausibly be held to contain, but also with how the various postulated elements are to be seen as relating to each other. Perhaps the principal general point to emerge from the discussion of these models is that one should not be too quick to equate distinctiveness with unconnectedness. We have seen in a number of cases that a decision to represent particular components of the lexicon as operating in complete isolation from each other has been vulnerable to the criticism of being too strong in the light of empirical evidence. This particular cautionary message is of relevance not only to discussion in the remaining sections of this chapter, but also to discussion in later chapters of the relationship between the L1 mental lexicon and the L2 mental lexicon.

Modularity and the mental lexicon

We turn now to a view of mind which takes the notion of disconnectedness of different components of mentation very far indeed. In the present section we shall consider the view that the entire language faculty is a fully autonomous module, its operations impermeable to information from other sources, and we shall explore how this view relates to the lexical dimension of language processing both theoretically and empirically (for further discussion see Singleton, 1993a, 1998).

The tradition which recent approaches to modular processing claim as their pedigree is that of 'faculty psychology', whose origins are customarily traced back to the work of Franz Josef Gall (1758–1828), a German anatomist who developed the view that each intellectual and behavioural attribute was controlled by a particular location in the brain. He opposed this 'vertical' account of the nature of mind, in which the character of mentation was seen as dependent on the subject matter involved, to the prevailing 'horizontal' account of the mind, which represented mental operations as transcending content domains.

The current version of the modularity hypothesis is summarized by Garfield as follows:

The mind is not a seamless, unitary whole whose functions merge continuously into one another; rather, it comprises – perhaps in addition to some relatively seamless, general-purpose structures – a number of distinct, specialized, structurally idiosyncratic modules that communicate with other cognitive structures in only very limited ways. (Garfield, 1987b: 1)

The kinds of systems that are hypothesized to be modular within this perspective include input systems such as certain components of the

perceptual and the language-reception systems, and output systems such as aspects of motor control and language production. This view of the mind has two influential advocates in the persons of Noam Chomsky and Jerry Fodor, between whom, however, some differences exist: whereas Chomsky discusses modularity essentially in relation to language acquisition (see, e.g. Chomsky, 1980a, 1980b, 1988), Fodor's concerns are largely processing-oriented (see, e.g., Fodor, 1983, 1989).

The content of the language module: differing views

An important question that immediately arises is that of the actual content of the language faculty or module. Chomsky takes its central component to be 'grammatical competence', but does not come to any very firm conclusions about the precise boundaries of such competence. He questions, for example, whether the organization of sound belongs properly to the system of language rather than to other systems (1980a: 61), and, as we saw in the previous section, has long professed agnosticism about whether one can distinguish linguistic meaning from non-linguistic meaning. Fodor, for his part, seems to entertain no doubts about the intramodularity of phonetic/phonological processing. With regard to semantic processing, he takes a similar line to Chomsky's, although this does not prevent him from seeing the 'shallower' aspects of lexical processing as intramodular (see, e.g., Fodor, 1983: 64ff.; 1987a: 55ff.; 1989: 5ff.; Carston, 1988: 51ff.). Other modularists who have insisted that 'linguistic' meaning is clearly separable from other varieties of meaning (see, e.g., Emmorey & Fromkin, 1988; N. Smith & Wilson, 1979) have been content to regard the former as being represented and processed within the language module.

Neither Chomsky nor Fodor claims that the language module has absolutely no connection with other cognitive operations, nor that every aspect of cognition is modularly organized. (Fodor dismisses this latter notion as 'modularity theory gone mad' (Fodor, 1987b: 27).) It is obvious that the normal use of language requires an interface between language and other aspects of cognition, and Fodor and Chomsky both hold that this interface is provided by some kind of 'central', that is, general, non-modular, system which interconnects the modules and enriches their output with a range of experience accumulated from the previous operation of the various modules. The modularist position posits only, in relation to language, that there is a dimension of language-related cognition which is subserved solely by the language module, and that what happens within this particular dimension is impervious to 'central' knowledge and processes. On this view, the contribution of 'central' elements is a stage or level of language-related cognition which is separate from the strictly linguistic responsibilities of the language module.

It should be noted too that both Chomsky (e.g., 1981: 33) and Fodor (e.g., 1989: 11) admit to some reservations about the general empirical foundations of the modularity hypothesis. Both are obliged by the current state of the evidence to regard the question of whether or not language is subserved by a separate faculty as an empirical one. A very useful contribution to the debate in this connection from the modularist viewpoint would have been some neurolinguistic indications of a specific physiological correlate of an autonomous language module. However, such indications as are forthcoming in this area are not generally seen as offering unambiguous support for the modularity hypothesis (see, e.g., Jacobs & Schumann, 1992; see also Singleton's 1998 discussion of Linebarger, 1989).

Language processing in a Fodorian perspective

Despite such empirical uncertainties, the modular perspective on language and mind remains a powerful paradigm in linguistics and psycholinguistics. As has been indicated, the processing aspect of modularity has been the main focus of Fodor's writings on the topic, and it is also true to say that Fodor's version of the modularity hypothesis has been more influential than any other among psycholinguists working on processing issues. Accordingly, it is appropriate in the present context to pay particular attention to what Fodor has to say on the question of modularity in relation to language processing.

The main features of Fodor's characterization of the language module are as follows:

- Domain specificity: the notion that the language module is uniquely dedicated to a unique subject matter.
- Mandatory processing: the idea that we cannot hear utterances in a language we know without hearing such utterances as sentences.
- Inaccessibility to consciousness: the claim that most genuinely linguistic processes lie in the realm of the unconscious.
- Speed: the assumption that language processing is an inherently rapid process as compared with problem-solving activities such as chess.
- Informational encapsulation: the view that language-processing mechanisms are, as it were, blinkered with regard to data other

than the specifically linguistic data on which they are designed to operate – a view that for Fodor is the very cornerstone of his entire modular edifice, as well as the most controversial of his claims.

- Shallowness of intramodular processing: the suggestion that intramodular language processing is an essentially formal matter, with no semantic analysis taking place 'inside' the items being processed.
- Neural hardwiring: the claim that the language module has its own particular neural architecture.
- Particular breakdown patterns: the interpretation of agnosias and aphasias as 'patterned failures of functioning' which cannot be explained in terms of 'decrements in global, horizontal capacities like memory, attention or problem-solving' (Fodor, 1983: 99).
- Specific developmental features: the reading of the research evidence on ontogenetic sequencing of language acquisition as indicating that much of language development is 'endogenously determined' (*ibid.*: 100).

As has been noted above, the most controversial aspect of the Fodorian conception of modularity is the notion that modules are 'informationally encapsulated' – the notion that, with regard to language processing, for example, general knowledge, contextual information, etc. have no role in intramodular linguistic 'computations'. In arguing for the informational encapsulation of modules, Fodor often refers to what he calls the 'teleological argument', claiming that modules are informationally encapsulated because they need to be in order to operate as efficiently as they do. One of the examples from visual perception he uses (1989: 11) is the case of someone spotting a 'yellow stripey thing' in New York and having to come to a rapid conclusion about whether it is a tiger. He argues that in such circumstances a perceptual system that was permeable to contextual expectations would not function rapidly enough to avoid disaster and that therefore modular processing needs to be 'as much like a reflex as possible' (*ibid.*).

Against this line of reasoning one can cite instances of people not believing and therefore not reacting appropriately to the evidence of their senses. Thus, in relation to language, one can point to what typically happens in situations where, for one reason or another, the expectation is that language *x* is being spoken but where, in fact, language *y* is being used. In such circumstances, comprehension tends to be blocked, even where both languages are familiar to the individual in question. For example, the following experiences were recently retailed to me by a native speaker of Finnish:

> My sister, while studying in France, was once addressed on the street in Finnish. Only after several attempts by the speaker did she understand her own native language, the point being that she was expecting French. I have had a very similar experience trying to make Finnish out of something that was easy enough to understand when I realized it was English. (Service: personal communication)

Another body of evidence which seems to run counter to Fodor's point of view is that which emerges from the observation of the effects of deep hypnosis. With appropriate suggestion, a hypnotized subject may perceive and interact with objects and persons which are not present – or even totally fictitious – and may fail to perceive objects and persons which are present (see, e.g., Orne & Hammer, 1974: 136). Phenomena of this kind surely suggest that all perceptual systems are penetrable by higher-level information. Even reflexive responses may, apparently, be affected by hypnosis. Chertok, for instance, reports (1989: 63–64) cases where hypnosis sufficed to anaesthetize patients undergoing surgical operations, and even to arrest salivation and bleeding. If it is true that something as fast and as automatic as a physiological reflex can be influenced by externally implanted information or pseudo-information, then one surely has to question the credibility of the notion of informational encapsulation in language processing.

In any case, as Fodor acknowledges, such a notion appears to conflict with a large number of psycholinguistic findings, notably the findings of experiments involving reduced-redundancy procedures such as cloze.[1] It is a well-established fact that in cloze tasks the more predictable the target items in relation to the blanks (the higher their 'cloze value'), the better the performance of subjects attempting to fill the blanks will be. This looks like strong evidence of 'cognitive penetration' – evidence of the mechanisms involved in such tasks having access to subjects' expectations. To attempt to deal with evidence of this kind, Fodor (1983) deploys two lines of argument. His first is to question whether the mechanisms involved in the 'highly attentional' process of reconstructing degraded linguistic stimuli are the same as those which mediate 'automatic and fluent' processes. He cites in this connection Fischler & Bloom's (1980) finding that the recognition of test items where no degradation of the stimuli was involved was only marginally affected by

[1] 'In the cloze procedure words are deleted from a text after allowing a few sentences of introduction. The deletion rate is mechanically set, usually between every 5th and 11th word. Candidates have to fill each gap by supplying the word they think has been deleted.' (C. Weir, 1988: 49)

cloze value – and not at all affected by cloze value at high rates of presentation.

Fodor's second line of attack is to suggest that mechanisms internal to the language module may 'mimic' effects of 'cognitive penetration'. In support of this suggestion he refers to an experiment of Swinney's (1979) in which subjects listened to stimulus sentences along the lines of 'Because he was afraid of electronic surveillance, the spy carefully searched the room for bugs' – each containing an ambiguous word such as *bug* – and at the same time made lexical decisions about letter strings presented visually immediately after the occurrence of the ambiguous items (i.e., decided whether the strings in question constituted words or non-words). Swinney found a facilitation effect in relation to lexical decisions on strings forming words with meanings related to the meanings of the ambiguous words determined by their sentential contexts. Thus, the presentation of *bug* in the above sentence would facilitate a decision as to whether or not *microphone* was a word. However, what Swinney also found was that decisions on strings with meanings related to meanings of the ambiguous items which were not suggested contextually were also facilitated. Thus, the presentation of *bug* in the above context would also facilitate a decision on *insect*. To Fodor this finding indicates that what looks like general contextual effects in language processing may in fact be a matter of interlexical excitation. He hypothesizes that the mental lexicon is a sort of connected graph, with lexical items at the nodes and with paths from each item to several others, and that accessing an item in the lexicon consists in exciting the corresponding node, which also occasions the excitation of pathways that lead from that node:

> when excitation spreads through a portion of the lexical network, response thresholds for the excited nodes are correspondingly lowered. Accessing a given lexical item will thus decrease the response times for items to which it is connected. (Fodor, 1983: 80)

Fodor's conception of intramodular excitation of connected lexical forms relates to what he has to say about the relative shallowness of intramodular language processing. Citing evidence from his own work (Fodor *et al.*, 1980) that the recovery of the semantic definition of lexical items is not a prerequisite for processing syntax, he posits that the language module's operations are confined to the processing of 'linguistic and maybe ... logical form' (*ibid.*: 90). This brings us directly to the question of the mental lexicon in relation to the modularity hypothesis.

Modularity and lexical processing

The advantage, from Fodor's point of view, of confining his conception of the language module to that of a non-semantic processor is that it does not confront him with the intractable problem, discussed above, of where to draw the line between linguistic and non-linguistic meaning. However, on the one hand, his postulation of task-induced non-standard processing is a two-edged sword, and, on the other, it is not clear that what he says about the excitation of lexical nodes succeeds in circumventing the semantic/pragmatic issue.

Regarding the non-standard processing argument, if it is legitimate for Fodor to invoke such an argument in relation to modularity-challenging results elicited by cloze procedures, it must be legitimate for others to invoke it to account for modularity-friendly findings from other experiments. Indeed, it seems odd that Fodor should wish to claim that the restoration of degraded linguistic stimuli – by no means unknown in the ordinary use of language – may trigger non-standard processing, whereas he accepts as self-evidently indicative of normal processing the results of Swinney's (1979) experiment. After all, this latter involved subjects in consciously deciding whether or not visually presented strings of letters constituted words while at the same time dealing with a series of unconnected sentences presented in a different mode – i.e., aurally. Swinney's procedure strikes one as far more artificial and form-focused than any cloze task, and thus far more likely than cloze to provoke non-standard processing.

As for the explanation of apparent 'cognitive penetration' in terms of the excitation of complexes of lexical nodes, this seems plausible enough as a non-semantic account of what looks like a semantically motivated phenomenon until one stops to consider the nature of the interconnections it presupposes. The evidence is that such interconnections do indeed exist, but that they are (in the proficient language-user) primarily based on semantic relatedness (see, e.g., Aitchison, 1994). Indeed, if the nodal excitation posited by Fodor were not assumed to proceed along pathways linking semantically related items, then the 'mimicking' of contextual-semantic effects of which he writes would remain unaccounted for. The non-semantic process that Fodor posits as an explanation of evidence of context effects turns out, therefore, to be entirely dependent on connections between lexical nodes which derive from the denotative and connotative associations of the lexical items concerned. There is surely some inconsistency, to say the least, between Fodor's non-semantic conception of the language module and his postulation of

lexical activation via meaning-based pathways. Moreover, the meaning-based character of these pathways brings us right back to the question of the nature of meaning.

A third possible explanation of context effects which preserves Fodor's notion of informational encapsulation of intramodular processing is that the effects in question are genuinely contextually induced, but that they are 'postperceptual' – that is, brought about by operations which (in language reception) come into play after the completion of intramodular processing and which take as their input the output of the module. On this view – proposed by Carston (1988) – exhaustive module-internal lexical access would be followed by parallel mappings and context-related choices between accessed items.

Let us not ignore, however, the possibility that what look like on-line context effects may actually *be* online context effects. Fodor himself notes that Marslen-Wilson's (1973) subjects were not only able to repeat linguistic stimuli with a time-lag of just a quarter of a second but also able to understand the words they were repeating. This means that not only formal aspects but also 'cognitive' aspects of lexical processing must be extremely fast as far as the reception of speech is concerned. More recent experiments by Marslen-Wilson have shown that subjects take no longer to relate an incoming utterance to discoursal context – even where pragmatic inferencing is involved – than to process it 'shallowly' (Marslen-Wilson & Tyler, 1987). Also relevant is the way in which subjects involved in speech shadowing (see above) exhibit highly fluent restoration of mispronounced words, these fluent restorations occurring far more frequently during the shadowing of normal prose than when the mispronounced words were anomalous with respect to context (see, e.g., Marslen-Wilson, 1975; Marslen-Wilson & Welsh, 1978). Other experimental findings (see Marslen-Wilson & Tyler, 1980) – already referred to – have shown that in a normal spoken prose context, target words which took on average 369 milliseconds to say could be identified on average within 200 milliseconds – which must mean that contextual information was somehow causing alternative possibilities to be eliminated while the words in question were still being uttered.

Another piece of evidence in favour of taking context effects at face value emerges from an experiment conducted by Foss (1982) in which he examined the influence of two aurally presented priming words on the identification of a target phoneme in a third aurally presented word. Foss discovered substantial priming across intervening words and sentences when coherent, meaningful sentences were used. For example, the recognition of /f/ in fish was primed in the following kind of context: 'The entire group examined the gills and fins. Everyone agreed that this was unlike any other fish caught in recent years.' However, when the words were jumbled into random lists, the priming effect disappeared. The fact that coherent contexts resulted in priming, whereas lists of words did not, surely constitutes counter-evidence to Fodor's notion that context effects are 'mimicked' by the activation of lexical networks in the mind merely through the occurrence of individual lexical forms. Carston's alternative view – that context effects are real but postperceptual – also receives little comfort from Foss's finding that initial phonemes of words (e.g., the /f/ in fish) were primed by previous context, and still less from Marslen-Wilson & Tyler's above-reported (1980) finding that context information took effect before the uttering of target words was complete.

How then to explain Swinney's cross-modal lexical priming results? One possible explanation lies in their very cross-modality, and, in particular, in the fact that reading was involved. The general view among experimental psychologists seems to be that reading processes differ from listening processes in terms of the extent to which they use context. It is indeed a psychological commonplace that whereas in speech perception, context effects are 'readily obtainable', 'in skilled reading ... context effects seem elusive' (A. Ellis & Beattie, 1986: 222). Thus, Fischler & Bloom's (1980) finding – the absence of facilitatory context effects from normal-speed reading – which Fodor cites against the whole concept of 'cognitive penetration', is normally interpreted as an indication of the particularity of the processing of the written signal with regard to use of context. A. Ellis & Beattie (1986: 225–226), for example, suggest that the ready decipherability of the printed word as opposed to the relatively impoverished nature of the speech signal favours 'bottom-up' rather than 'top-down' processing. If it is true that printed stimuli give rise to a greater measure of 'bottom-up' processing, then the fact that Swinney's experiments involved the use of visually presented letter-strings may well have triggered an across-the-board concentration on the characteristics of individual lexical items, with the result that contextually irrelevant meanings as well as relevant meanings were activated. The same argument can be applied to other cross-modal studies whose findings have been cited as pro-modular (e.g., Seidenberg *et al.*, 1982; Tanenhaus & Donnenworth-Nolan, 1984; Tanenhaus *et al.*, 1979).[2]

[2] Experimental evidence cited against an on-line role for contextual information in lexical processing from studies other than those with a cross-modal design tends to be ambiguous. Even modularists accept that such evidence is amenable to non-modular as well as modular interpretations (see, e.g., Tanenhaus *et al.*, 1987: 100–101).

This is not to say that context effects are entirely absent from processing where printed stimuli are involved. Even Swinney found that the contextually predictable meanings of his ambiguous items were more strongly activated than other meanings (see above). As far as tasks involving only reading are concerned, Fischler & Bloom (1980), while failing to find facilitatory context effects, did find that responses to words which were anomalous in context were inhibited relative to responses to contextually predictable words. This result too has been linked to the specifics of the reading process as opposed to the listening process. Harris & Coltheart (1986) note that in auditory word recognition we hear the sounds of any given word sequentially, which allows for the possibility of interaction between contextual information and recognition processes after only a part of the word has been uttered, but that, in contrast, in visual word recognition we have access to the whole word simultaneously, which abolishes any advantage in having a system which uses context to identify words before their production is complete.

> However, there is an advantage in having a system which can check word identification to see if the word which we have identified is consistent with context, and it is this checking procedure which Fischler and Bloom claim is causing the inhibition effects which they have demonstrated. (Harris & Coltheart, 1986: 170)

The claim here, in other words, is that the use of context in normal reading is 'postperceptual' because of the nature of the signal involved.

If lexical processing in reading does differ from lexical processing in listening because of the ready decipherability and instant availability of the signal in the former, it ought to be the case that rendering the written or printed stimulus more difficult to decipher will cause on-line context effects resembling those found in auditory word recognition to become discernible. And, indeed, this is what has been found (see A. Ellis & Beattie, 1986: 224). This brings us back to Fodor's suggestion that context effects in cloze tasks may be the result of the operation of some kind of abnormal back-up system. On the basis of the foregoing, we can probably accept Fodor's claim, but in a completely contrary sense to the one he intended. It appears to be the case that degrading the written/printed signal causes the reader to activate word-recognition processes which are normally reserved for the perception of speech. In other words, presenting readers with a degraded written/printed stimulus seems to rob the instruction-derived skill of reading of its specific signal-related characteristics in respect of lexical processing and to bring it closer to the 'primary' skill of understanding speech in terms of the degree to which context is exploited and relied upon.

Connectionism

Our final port of call in this chapter is the approach to language processing – indeed to all kinds of mental processing – known as connectionism or parallel distributed processing. The term 'connectionism' relates to the fact that this approach takes its inspiration from what is known about neurophysiological activity in the brain:

> During any brain activity, numerous brain cells are active, sending out signals to other neurons. Some signals are 'excitatory' (causing arousal), others are 'inhibitory' (causing suppression). The result is a 'network' of interconnected units. Arousal of any units causes them to be reinforced, whereas inhibition leads to the gradual loss of a connection. Psychologists have recently tried to build computer models which simulate this connectionist viewpoint. (Aitchison, 1992: 31)

> The connectionist account adopts the analogy of brain-style neuronal interactions (i.e. the fact that we have brains which are made up of millions of interconnected neurones which can be viewed as 'on-off' switches) and proposes that our cognitive system works in a very similar way. (Forrester, 1996: 152)

The alternative label, 'parallel distributed processing', refers to the claim made by connectionists that different portions of information are processed independently of one another ('in parallel') on different levels ('distributed').

One way of thinking about the contribution of connectionism to the modelling of language in the mind is in terms of a change of metaphor. Psycholinguists interested in language processing have in recent decades often drawn analogies with the operations of computers, talking about the 'articulatory program', 'programming errors', etc. Connectionists for their part have seen themselves as wishing to 'replace the "computer metaphor" as a model of the mind with the "brain metaphor"' (Rumelhart *et al.*, 1986: 75).[3] This is not a trivial change, however. It moves psycholinguistics in the direction of taking into account 'constraints from studies of the nervous system' (Broeder & Plunkett, 1994: 433), of proposing models which are alignable with 'neurophysiological reality' in much the same way that certain developments in syntactic theory have been motivated by

[3] 'Indeed, computer scientists are now designing machines called *neural networks* that attempt to imitate the brain's vast grid of densely connected neurons (J. A. Anderson & Rosenfeld, 1988; Levine, 1990).' (Wade & Tavris, 1996: 341)

an aspiration to greater 'psychological reality' in the sense of a requirement that 'a grammar provides us with a description of the abstract structure of the linguistic knowledge domain' which 'corresponds to the speaker's *internal* description of that domain' (Bresnan & Kaplan, 1982: xxiii).

Connectionism, modularity and parallel versus serial processing

The connectionist view of mind is usually taken to be antipathetic to the modular view discussed in the previous section. Thus, for instance, Cook & Newson (1996: 31) contrast the theory which 'divides the mind into separate compartments, separate modules, each responsible for some aspect of mental life' with 'cognitive theories that assume the mind is a single system, for example connectionism'. This is not, however, a universal view. For instance, Tanenhaus *et al.* (1987) posit different networks of connections for the parallel but autonomous processing of different types of information, which they see as merely a connectionist translation of the modularity idea and as doing no violence to the essentials of Fodor's theory. How can two such radically divergent views of connectionism co-exist?

We should note in this context that connectionism belongs to a much broader parallel processing perspective which stands in opposition to the serial processing perspective. As Garman (1990: 175) points out, the issue here is not about simultaneity versus sequentiality. Sequences of operations are found within parallel models, where successively presented domains (processing items/problems) obviously have to be dealt with successively; and simultaneity of operations is found in serial models, where different levels of operation may be simultaneously active though working on different domains – e.g., the processing of item x may be beginning at one level while the processing of item y is nearing completion at another. The essential difference between the parallel perspective and the serial perspective is that the former promulgates the notion of the independence of the different processing operations which are triggered by particular events and stimuli, whereas the latter represents processing as serially organized, with each stage dependent on the output of the previous stage.

The notion of independence of different processing operations in parallel models such as the connectionist one is not, however, (*pace* Tanenhaus *et al.*, 1987) in any real sense comparable to the Fodorian idea of informational encapsulation. Independence of processing in parallel models refers to micro-operations, and is not to be identified with a barrier between, for example, 'higher-level' semantic processes and 'lower-level' formal computations. Parallel-processing models are usually interpreted, on the contrary, as making claims about a high degree of top-down/bottom-up interactivity:

> lower-level processes can influence higher ones within the parallel-processing model, whereas in the serial model all higher-level processing [in speech production] is complete, for a given domain of processing, prior to any lower-level activity. By virtue of this, parallel models may be said to allow for interactive, on-line, bottom-up influences during the time course of [productive] language processing. (Garman, 1990: 174–175)

The symbolic paradigm and the connectionist paradigm

There is a further respect in which connectionism has been seen to pose a challenge to the Chomskyan/Fodorian view of language and mind. The Chomskyan/Fodorian view (in common with many others) is based on what is sometimes called the symbolic paradigm, the idea that cognition involves the manipulation of symbols:

> These symbols could refer to external phenomena and so have a semantics. They were enduring entities which could be stored in and retrieved from memory and transformed according to rules. The rules that specified how symbols could be composed (syntax) and how they could be transformed were taken to govern cognitive performance. (Bechtel & Abrahamsen, 1991: 1)

The connectionist paradigm, on the other hand, is 'distinguished from the traditional paradigm by the fact that it does not construe cognition as involving symbol manipulation', but offers 'a radically different conception of the basic processing system of the mind-brain ... inspired by our knowledge of the nervous system' (*ibid.*: 2). It sees knowledge in terms of connection strength rather than rules or patterns:

> In these models, the patterns themselves are not stored. Rather, what is stored is the connection strengths between units that allow these patterns to be recreated. (McClelland, Rumelhart & Hinton, 1986: 31)

Strong anti-symbolist claims have been made on the basis of Rumelhart & McClelland's (1986) simulation of the learning of past-tense forms using a connectionist architecture. They showed that quite a simple network could be trained to supply appropriate past-tense inflections/mutations for 506 English verbs, including 98 irregular verbs, stabilizing at a 91 per cent level of accuracy after 200 cycles of training. Moreover, the errors the network in question made

en route resembled those made by children acquiring English as their first language. Elman (1990a) obtained not dissimilar results in respect of the acquisition of phonological structure. On the basis of their findings, Rumelhart & McClelland claim that language can be learned and processed without any recourse whatsoever to rules:

> The child need not figure out what the rules are, nor even that there are rules. The child need not decide whether a verb is regular or irregular. There is no question as to whether the inflected form should be stored in the lexicon or derived from general principles. (Rummelhart & McClelland, 1986: 267; see also, e.g., Seidenberg, 1995[4])

Some fairly sharp reaction to this kind of claim has come from Chomskyans. For example, Pinker & Prince (1988) point up divergences of detail between the network's output and that of children acquiring the same verbs, calling into question 'whether it is an accurate model of children' (*ibid.*: 81); Fodor & Pylshyn (1988) choose to interpret the connectionist proposals as relating merely to implementational, low-level phenomena; and Cook & Newson (1996) dismiss the fact that Rumelhart & McClelland's network can learn the particularities of English verb forms as of strictly no consequence from a Chomskyan standpoint:

> As these forms are peripheral to UG [Universal Grammar], whether they are learnable or not by such means has no relevance to the claims of the UG model. (Cook & Newson, 1996: 71)

An interestingly different tack is taken by Stevick (1996), who suggests that, in describing linguistic phenomena in terms of statements of 'rules', we may become the dupes of our own metaphors:

> We conclude (or at least we let our figures of speech give the impression) that these statements, the work of our minds and our hands, must be the cause (or must stand for the cause) of what we have observed, and we express this conclusion by saying, 'Rules govern behaviour.' (Stevick, 1996: 71; cf. also Elman, 1990b)

Stevick suggests that the term 'rule' might be better confined to the statements and that the term 'regularity' be deployed to cover the phenomena being described. In sketching an answer to the question of what causes such regularities, he suggests a possible line of demarcation between the two paradigms which, though expressed in less brusque rhetoric, actually chimes rather well with Cook & Newson's view of things insofar as it allows for the possibility of

[4] I am most grateful to Mickey Bonin for supplying me with the Seidenberg reference as well as the Plaut & Shallice and Regier references cited below.

both biologically endowed '*kinds* of regularities', which are assimilable to core UG principles, and learned specificities, which can be compared to the 'periphery' in UG terms:

> The *kinds* of regularities reflect inborn characteristics of the physiological equipment we use for networking. Some of these characteristics are very subtle, as yet undiscovered, and absent from the mechanical and electronic equipment used in connectionist investigations.
>
> The *specifics* of regularities come from what has previously happened to our networks. (*ibid.*)

Taking a somewhat broader approach to the question and taking it beyond the confines of the question of the relationship between connectionism and UG-style nativism, McShane (1991) suggests that an earlier version of the connectionist model (McClelland & Rumelhart, 1981) may provide a pointer to a compromise between connectionist and symbolic principles. The model in question is focused on reading. Letter identification is via an interconnected network of feature units with inputs to letter units:

> The individual letters are represented as units (and therefore as symbols) one level higher in the processing hierarchy. This level also forms an interconnected network along connectionist principles, which activates word units at the next highest level in the hierarchy. (McShane, 1991: 341)

For McShane, this kind of hybrid model combines the advantage of capturing bottom-up activation by sensory data with that of capturing top-down control from higher levels.

Connectionism and spreading/interactive activation

As was indicated at the outset of this chapter, one of the principal elements in connectionism is the concept of spreading/interactive activation, the idea that in language processing a multiplicity of nodes are excited by the arousal of a node to which they are connected. This notion ante-dated connectionism and has an existence outside as well as within the strictly connectionist school. Thus, for example, Dell (1986) relates his 'spreading activation theory of retrieval' not only to connectionist proposals (e.g., Cottrell & Small, 1983; Feldman & Ballard, 1982; Grossberg & Stone, 1986) but also to pre-connectionist 'interactive activation' models (e.g., McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982) and to other spreading-activation theories (e.g., J. R. Anderson, 1983; Dosher, 1982; Rumelhart & Norman, 1982). However, these different approaches to some extent blur into each other, and, indeed, have at

various times involved the same researchers. Rumelhart, for instance, who is now known as a connectionist, has also worn an interactive-activationist hat and a spreading-activationist hat (see references above and in previous paragraphs). All three approaches posit parallel processing; connectionism and interactive-activation models share the assumption that activation not only spreads outwards to more and more nodes – the spreading-activation view[5] – but also moves backwards and forwards between the activated nodes; connectionism differs from the other two approaches in, broadly speaking, making no use of symbols and in postulating not only excitatory but also inhibitory signals.

Aitchison gives the following account of the workings of activation in the interactive-activation conception of lexical retrieval:

> an initial impetus progressively fans out and activates more and more words as it spreads along the various connections ... As the activated links are inspected, those that are relevant get more and more excited, while those that are unwanted fade away... Since the current is flowing to and fro, anything which is particularly strongly activated in the semantics will cause extra activation in the phonology, and vice versa. (Aitchison, 1994: 206)

The connectionist picture would differ from the above on the one hand in envisaging the 'turning off' rather than just the 'fading away' of certain elements under particular impulses and also in representing 'words' as collections of connections rather than as stable, enduring symbols.

Connectionism, the lexicon and the future

We have seen that connectionism is characterized by a particular configuration of features, some of which it shares with other models. The particular models of the lexicon dealt with in the first part of the chapter demonstrate this overlap admirably. Thus, for example, the Morton model, the Marslen-Wilson model and the Levelt model all posit parallel processing. Only the Forster model is wholly free, as it were, from all taint of connectionism. Moreover, no doubt because it provides the promise of such a ready interlocking with the neurophysiological dimension of language processing (see above), the

[5] This is a simplification. The relationship between terminology and model type is not always as neat as is suggested here. For example, Dell's (1986) version of 'spreading activation' allows for 'positive feedback from later to earlier levels', a feature which 'makes processing in the network highly interactive' (Dell, 1986: 288). Indeed, Green (1993: 269) goes so far as to describe Dell's proposals as a 'three-level connectionist network'.

connectionist optique is exercising a growing influence in the sphere of lexical modelling. For instance, Marslen-Wilson's cohort model, which was always a parallel-processing model, has recently moved very much in the direction of becoming an interactive-activation model too (Marslen-Wilson, 1987, 1990). There has also, let it be said, been some movement from the connectionist side, with more recent versions of connectionism moving beyond a strictly formal account and allowing for access to some level of semantic representation (see, e.g., MacWhinney & Leinbach, 1991; Plaut & Shallice, 1994; Regier, 1996).

Almost everywhere one turns in the recent literature on the mental lexicon one finds references to connectionism. Aitchison, for example, cites (1994: 233) an observation from Elman & McClelland (1984) to the effect that computer modelling of lexical processing of the kind engaged in during the 1970s and early 1980s had not been notably successful. Aitchison goes on to suggest that the new-style connectionist modelling 'may be on the right track' in taking the brain as its inspiration:

> The human brain is capable of massive parallel processing: an uncountable number of connections can be made simultaneously. It activates many more connections than are strictly needed for almost every brain process, then suppresses those which are not required. These properties are found in some of the new 'connectionist' computer models with their intricate networks. (*ibid.*)

In an L2/bilingual perspective, Green (1993: 260) notes that the effect of a delay in lexical comprehension owing to competition between plausible L1 and L2 candidates for recognition has been modelled in a single-language framework (i.e., where the competitors are from the same language) using connectionist frameworks; Gass & Selinker (1994: 276) evoke the relevance of connectionism to current theories of how L2 words relate to each other in the learner's mind (see next chapter); and R. Ellis (1994: 407) refers to Schmidt's (1988) contention that connectionism is well-adapted for dealing with variability and 'fuzzy concepts'.

Ellis also quotes Gasser's (1990) assessment of the future role of connectionism, which will serve well as a coda to the present section:

> It is now clear that some form of connectionism will figure in a general model of human linguistic behavior. The only question is whether the role will be a minor one, relegated to low-level pattern-matching tasks and the learning of exceptional behavior, or whether the connectionist account will supersede symbolic accounts, rendering them nothing more than approximations of the actual messy process. (Gasser, 1990: 186)

Concluding summary

In the first part of this chapter we examined the better-known models of lexical processing. We looked at two influential representatives of the direct, or one-stage, type of lexical model – Morton's logogen model and Marslen-Wilson's cohort model; we explored an oft-cited representative of the indirect, or two-stage, type of model – Forster's search model; and we also considered Levelt's 'blueprint for the speaker', which, though not solely a model of the lexicon, has a great deal to say about lexical operations. In the second part we considered the implications for lexical processing of the idea that the mind is modularly organized, focusing, in particular, on Fodor's proposals in this connection. In the third part we explored the lexical dimension of connectionist models of mind.

With regard to the logogen model, it was suggested that, while the notion of separate components for auditory and visual word analysis seems to be empirically supported, the notion that such components might be totally unconnected does not. With reference to the cohort model, some questions were raised about the strict linearity of its earlier versions, given the importance of ends of words in word recognition. It was recognized that the logogen model and the cohort model have in common the fact that one of their focal aims is the explication of context effects in lexical processing. Forster's model, for its part, was found to be wanting in relation to its proposals in respect of context effects, since the cross-referencing it posits to account for such effects requires either a strictly lexicon-internal solution based on the concept of a 'master file', which runs into empirical difficulties, or an appeal to a general knowledge store, which entails the making of a dubious distinction between linguistic and pragmatic meaning. As far as Levelt's 'blueprint' is concerned, it was seen as having the advantage, in envisaging the lexicon as mediator between conceptualization and grammatical and phonological encoding, of being in harmony with current views of the relationship between lexis and grammar; on the other hand, it was criticized for its suggestion that lexical knowledge is separable from encyclopedic knowledge, for its representation of lexical knowledge as static, declarative knowledge, for its incapacity, as it stands, to cope with multilingual processing, and for its postulation of a developmental point from which feedback between the formulation process and the conceptualization process simply ceases.

The modularity hypothesis was accorded its due place in the 'faculty psychology' tradition, and its relationship to Chomskyan thinking about the 'language faculty' was also noted. However, it

was discussed principally in the perspective of Fodor's conception of the language module, and in particular in the perspective of the Fodorian notion of 'informational encapsulation' and the problem posed for this notion by evidence of lexical context effects. An outline was presented of a number of proposals to deal with apparent context effects in ways which leave the idea of informational encapsulation unscathed, but these proposals were found less persuasive than the simple expedient of taking on-line lexical context effects at face value, an approach which has strong empirical support.

Connectionism was presented as – on most interpretations – antipathetic to at least the Fodorian version of the modularity hypothesis and indeed to the entire 'symbolist' conception of linguistic knowledge in which it has its genesis. Some possible ways of reconciling symbolist and connectionist approaches were mentioned, and the relationship between connectionist models, spreading-activation models and interactive-activation models was briefly explored. Similarities between the connectionist representation of lexical processing and the models of Morton, Marslen-Wilson and Levelt were noted, and attention was drawn to the fact that, according to most commentators, the future evolution of models of the mental lexicon, and indeed of psycholinguistic models generally, will be heavily marked by the influence of connectionism.