5 *Comprehending spoken language: a blueprint of the listener*

Anne Cutler and Charles Clifton, Jr.

5.1 Introduction

The listener can be thought of as a device for conversion of acoustic input into meaning. The purpose of this chapter is to provide an outline, necessarily superficial but we hope not fragmentary, of the course of this conversion process. Language comprehension is the most active area of Psycholinguistics, and while word recognition has been more intensely studied than sentence understanding, and comprehension in the visual mode has attracted more research effort than listening, there is nevertheless a vast body of relevant literature to which a single chapter cannot hope to do true justice.

Figure 5.1, the blueprint of the listener, sketches the account which we will flesh out in the following sections. The process of listening to spoken language begins when an auditory input is presented to the ear. The manner in which auditory information is initially processed—the psychoacoustic 'front-end'—will not form part of our account; the initial processing of the auditory input with which we will be concerned is the speech decoding process. Here the listener first has to separate speech from any other auditory input which might be reaching the ear at the same time, then has to turn it into some more abstract representation in terms of which, for instance, a particular sound can be accorded the same representation when uttered in different contexts, at different rates of speech, or by differing speakers. These operations are discussed in Section 5.2.

The next stage is segmentation of the (continuous) signal into its component parts. However, the computations involved in segmentation do not form a separate stage which must be traversed before, say, word processing can begin. In Fig. 5.1 this is represented by the overlapping of the boxes: segmentation results in large part from the processing operations involved in word recognition and utterance interpretation. Nevertheless there is abundant evidence that listeners also use aspects of the spoken input—explicit segmentation cues, as they are referred to in Fig. 5.1—to determine word and syntactic boundaries. This evidence is described in section 5.3.

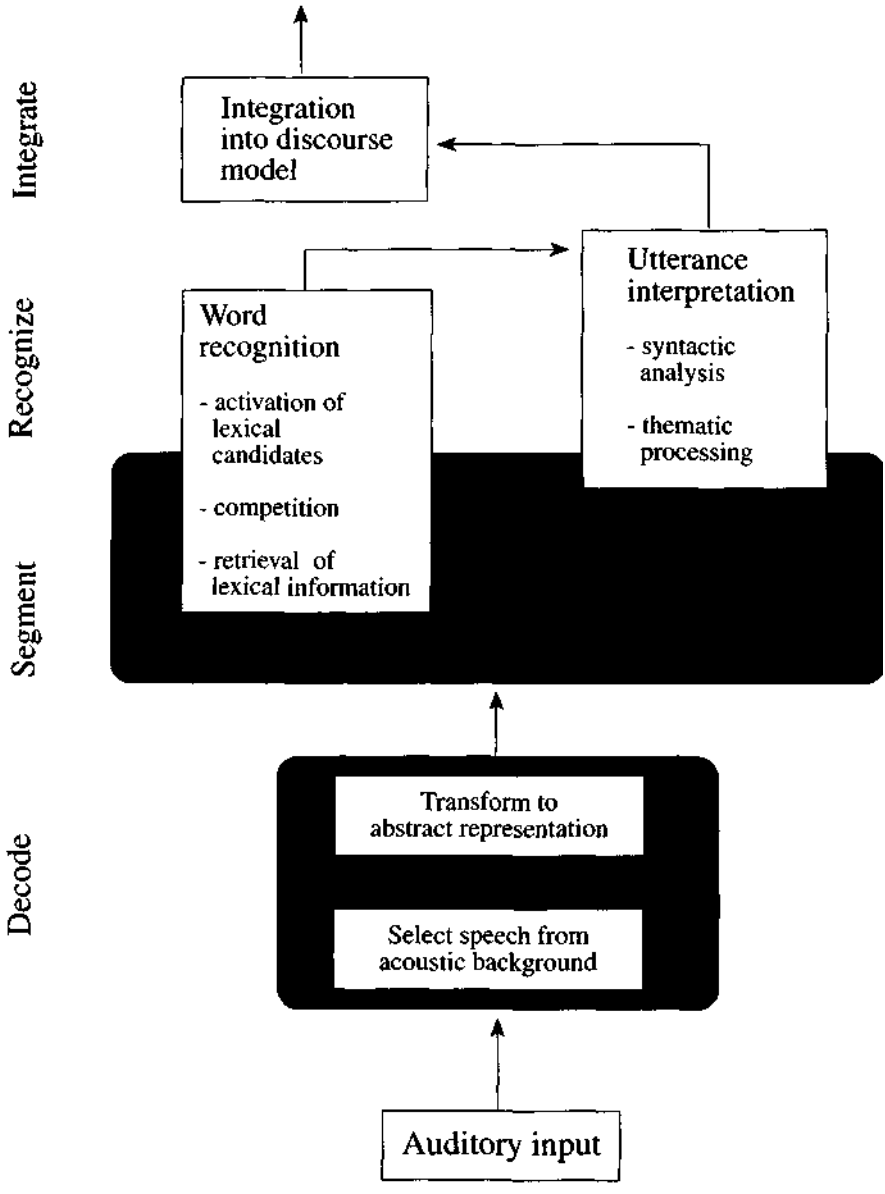

Fig. 5.1 A blueprint of the listener.

The process of lexical activation is the topic of section 5.4. Twenty years of lively research on the recognition of spoken words have led to a new generation of models in which multiple activation of word candidates, with ensuing competition between candidate words, is the core mechanism of recognition. We summarize the evidence regarding what type of information plays a role in initial activation of lexical candidates, and whether both matching and mismatching information are relevant. We also pay particular attention to a problem peculiar to auditory as opposed to visual word recognition, namely the relative weight of segmental versus suprasegmental information.

What information is retrieved from the lexicon is the topic of section 5.5, which discusses evidence concerning word semantics and morphology. Further evidence concerning retrieval of the syntactic relationships in which a word can occur, and the thematic roles it can adopt within a semantic structure, is examined in section 5.6, where we describe how the sequence of words which is the output of the processing so far is syntactically and thematically interpreted. We show how this interpretation process is incremental, is as near as possible to immediate, and is sensitive to a wide variety of lexical, pragmatic, discourse, and knowledge-based factors. The problems that the sequential, one-time-only nature of the auditory input pose for the listener may be solved by specialized characteristics of the human sentence processing system. This section also discusses the ways in which prosodic information can constrain the comprehension process.

As Fig. 5.1 shows, utterance interpretation as described so far is not the end-stage in the listener's process of extracting the speakers message from the auditory input. The utterance must be related to its discourse context in a wide sense, beyond the computation of, for example, thematic relationships. Further, beliefs about the speaker, and about the speaker's knowledge, need to be taken into account, as well as a range of sociolinguistic factors involved in conversational interaction. These again go beyond the scope of our account.

In our final section 5.7, however, we consider another topic of considerable generality, namely the architecture of the entire device which we have called the listener. Among the central issues in language comprehension research are the extent and nature of interaction among distinct components of processing, such as those outlined in Fig. 5.1, as well as whether all aspects of this processing system are based on common underlying principles or whether different aspects of the system call for different operating principles and different representational vocabularies.

5.2 Decoding the signal

Speech, as Alvin Liberman and his colleagues (Liberman *et al.* 1967) so memorably declared, is a code. The listener has the key, and can unravel the code to reveal the message it contains. But the unravelling operation is one of fearsome complexity.

Even this operation cannot begin before the speech signal itself has been identified, of course. Speech is presented as sound waves to the ear of the listener; but it does not command an exclusive acoustic channel. The sound waves reaching the ear carry any

other noise present in the listener's environment just as efficiently as speech-related noise. Thus the listener's first task is to separate speech from other auditory input reaching the ear at the same time.

Picking out a speech signal from background noise exploits the periodic nature of speech signals; noise is aperiodic and a regular structure stands out against it. Perceiving speech against a background of other sounds which, like speech, have a regular structure is less simple. However, the human auditory system can exploit grouping mechanisms which effectively assign acoustic signals to putative sources according to, for example, their frequency characteristics (see Bregman 1990 for a review).

Having isolated the part of the incoming noise which corresponds to the speech signal, the listener can then begin the decoding. The task is now to transform a time-varying input into a representation consisting of discrete elements. Linguists describe speech as a series of phonetic segments; a phonetic segment (phoneme) is simply the smallest unit in terms of which spoken language can be sequentially described. Thus the word *key* consists of the two segments /ki/, and *sea* of the two segments /si/; they differ in the first phoneme. The first phoneme of *key* is the same as the second phoneme of *ski* /ski/ or *school* /skul/ or *axe* /aeks/, the third phoneme *of back* /baek/ or *ask* /ask/, or the last phoneme *of pneumatic* /njumaetik/.

The structure of the phonemes themselves can be further described in terms of linguistic units: distinctive features are based on articulatory factors, and allow us to describe the phoneme /k/, for example, as a velar voiceless stop consonant. That is, the place of articulation for /k/ is velar (the back of the tongue touches the soft palate); its manner of articulation is a stop (it involves a closure of the vocal tract); and it is not voiced (there is no vibration of the vocal folds during the articulation of/k/). It contrasts only in place of articulation with /t/ (alveolar) and /p/ (bilabial); only in manner with no other sound in American or southern British English, but with the velar fricative /x/ or the velar ejective /k'/ in some languages, and only in voicing with /g/, which is the same as *jkj* except that articulation of/g/ involves vocal fold vibration. However, note that the articulatory features are not sequentially present; phonetic segments provide the finest-grained sequential description.

It is these phonetic segments which are present in speech only in an encoded form. Note that the linguistic description of phonetic structure does not imply a claim that this level of description constitutes an explicit representation which listeners have to construct in order to understand speech; such a description is necessary purely to capture the underlying distinctive linguistic contrasts. We will return later to the issue of whether explicit phonemic representations form part of the speech recognition process; for the present discussion, the statement 'recognizing the phoneme /k/' should be taken as equivalent to 'discriminating the word whose phonetic description includes /k/ from words containing contrasting phonemes'—for example *key* from *see, tea, pea,* or *he.*

The number of different acoustic realizations in which a particular phoneme can be manifested is potentially infinite. The acoustic realization is of course partly determined by any background noise against which it is presented. But is also to a

substantial extent determined by the speaker. Different speakers have different voices. Children's vocal tracts are much smaller than those of adults; women's vocal tracts tend to be smaller than those of men. The larger the vocal tract, in general, the lower the fundamental frequency, and thus the pitch, of the voice. Voices also change with age. Further, even a single speaker can have varying voice quality due to fatigue, hoarseness, illness and so on. The amplitude, and hence the perceived loudness, of speech signals, varies with speaker-listener distance, and with vocal effort. Emotional state can affect voice pitch (tension tightens the vocal folds and raises pitch), and can also, of course, affect the amplitude of the voice. Thus there is a very large range, both of amplitude and of frequency, across which the acoustic realization of a given phonetic segment can vary. Finally, the timing of segments is also subject to considerable variation, since rate of speech is another important variable affecting the realization of segments (and one to which listeners are highly sensitive; Miller and Liberman 1979; Miller 1981).

In addition to all these sources of variability affecting the realization of phonetic segments, the segments themselves are not discretely present in the speech waveform. Segments overlap, and vary as a function of the context (surrounding segments) in which they occur. They are coarticulated—that is, speakers do not utter one segment discretely after another, but, as described by Levelt, they articulate words, phrases, utterances as fluent wholes; the smallest articulatory segment for which some degree of invariance could be claimed is, as Levelt (this volume) points out, the syllable. Thus the properties of the signal which are relevant for the perception of one segment flow seamlessly into, and to a great extent overlap with, the properties relevant for adjacent segments. Coarticulation effects can in fact stretch across several segments. For instance, the /s/ segments at the beginning of *strew* versus *street* are uttered differently due to anticipatory coarticulation of the vowel: lip-rounding for /u/, lip-spreading for /i/. Again, listeners are sensitive to these contextual effects, in that experimenter-induced mismatches in coarticulatory information impair processing in a wide range of phoneme and word recognition tasks (Streeter and Nigro 1979; Martin and Bunnell 1981,1982;Whalen 1984,199l;Marslen-Wilson and Warren 1994; McQueen *et ai,* in press).

Variation as a function of context can indeed result in completely different forms of the acoustic information to signal the same phoneme; thus /k/ before /i/ as in *key* is quite different from /k/ before /u/ or *joj* as in *coo* or *caw,* and /k/ in initial position, as in cab */kxb/* or *keep* /kip/, is very different from /k/ in word-final position, as in *back* /baek/ *or peak* /pik/. Moreover, the same acoustic form can signal different phonemes in different phonetic contexts; thus the noise burst appropriate for /k/ in /ka/ will signal /p/ in /pi/, and, more dramatically, the form of /p/ in *speak* is essentially identical to the form of /b/ in *beak.* In other words, there is no one-to-one mapping of acoustic realization to phonetic identity.

A spectrogram is a visual representation of an auditory signal. It displays frequency (on the vertical axis) against time (on the horizontal axis), with greater energy represented by, for instance, darker shading. Figure 5.2 presents, in its top panel, a

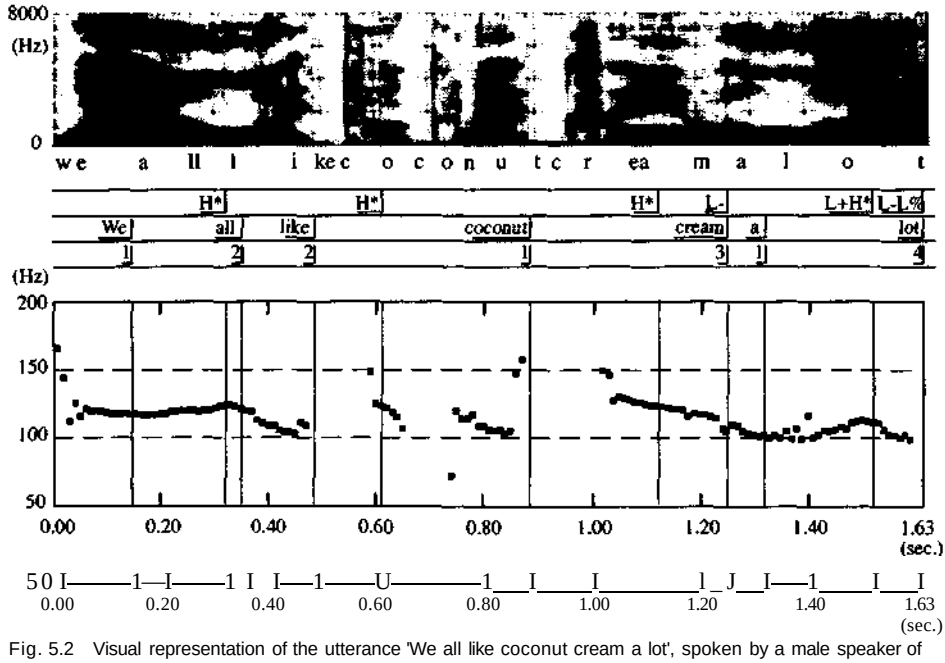

Fig. 5.2 Visual representation of the utterance 'We all like coconut cream a lot', spoken by a male speaker of American English. The top panel is a spectogram. which shows frequency (from 0 to 8000 Hz) on the vertical axis against time (0 to 163 s) on the horizontal axis. The lower panel traces the pitch contour (fundamental frequency in Hz) of the utterance. The middle panel is a prosodic transcription using theToBI (Tones and Break Indices) system.

spectrogram of the utterance *We all like coconut cream a lot*, spoken by a male speaker of American English. At various points in the spectrogram clear horizontal striations can be seen, indicating concentration of energy in particular frequency bands. These frequency bands are the formants: the resonant frequencies of the speaker's vocal tract. The more steady-state portions of the speech signal, in which the formant structure clearly appears, are the vowels. Sometimes, as can be seen in Fig. 5.2, the vowels remain steady; these are the monophthongs (single vowels), as in *nut* and *lot.* Other vowels are double (diphthongs), and show clear movement—for example in *like.* Semivowels, as their name suggests, are sounds which are between consonants and vowels; they show formant structure moving into the vowel that follows them. At the beginning of Fig. 5.2 the semivowel /w/ can be seen, and the speaker has also inserted a semivowel /j/ between *we* and *all* (the movement from about 0.12 to 0.18 seconds).

This insertion—effectively converting the string *we all* into homophony with *we yawl*—shows that sounds can occur in the signal even though they are not part of the representation which the speaker encodes. No semi-vowel would have been inserted if the utterance had begun *Kids all....* Similarly, sounds can be omitted or assimilated (the /k/'s at the end of *like* and the beginning of *coconut* have effectively become one sound), or can surface as other sounds (the final consonant of *coconut* and the initial consonant of *cream* have similarly merged into, effectively, one long /k/). These are effects of the coarticulation processes described above.

And, of course, it is very clear that the different phonemes in terms of which the utterance may be described are not discretely represented in the signal. Likewise, although the utterance shown in Fig. 5.2 consists of seven words, it is not easy to see in the visual representation where one word ends and the next begins. There is a clear gap in the signal at about 0.5-0.55, the boundary between *like* and *coconut.* But this gap appears almost identical to the gap at 0.66-0.71, and that latter gap falls in the middle of *coconut.* The gaps in both cases are stop closures—they are caused by the fact that the manner of articulation of /k/ involves a brief closure of the vocal tract; when the vocal tract is closed, no sound emerges, and there is a brief gap in the signal. This gap is acoustic evidence for a stop; it has no relevance to the segmentation of the signal into lexical units. At the boundary between *we* and *all* (approximately 0.12), and *all* and *like* (approximately 0.32), the speech signal is unbroken, and one phonemically determined pattern flows smoothly into the next; the same is true at the boundary of *cream* and *a* (approximately 1.22), and *a* and *lot* (approximately 1.33). Speech reaches the listener as a continuous stream, and, as we shall see in Section 5.3 below, this has important consequences for the operations involved in comprehension.

Finally, we return to the question of what units are explicitly represented during listening. Some decades ago, psycholinguists expended much research effort on investigating this question. The phoneme, as the smallest unit in terms of which the phonological form of utterances can be sequentially described, naturally exercises an intuitive appeal as a candidate 'unit of perception' in terms of which access to stored lexical representations could be achieved. Foss and Gernsbacher (1983), Marslen-Wilson and Welsh (1978), Pisoni and Luce (1987) are examples of models incorporating a phonemic level of representation in word recognition. However, the lack of one-to-one mapping between acoustic form and the phoneme led some researchers to reject the phoneme as a candidate perceptual unit. Thus Mehler *et al.* (1981) and Segui (1984) proposed obligatory intermediate representations in the form of syllables, and other models exist in which the intermediate representations are in the form of stress units (a stressed syllable plus optionally one or more unstressed syllables; Grosjean and Gee 1987), demisyllables (i.e. a vowel plus a preceding syllabic onset or a following syllabic coda; Fujimura and Lovins 1978; Samuel 1989), or diphones (i.e. effectively the stretch of speech from the midpoint of one phoneme to the midpoint of the next, capturing thus all the information relevant to contextual effects on phonemic realization, cf. Klatt 1979; Marcus 1981).

A radically different solution to the problems caused by lack of invariance in the acoustic signal was to postulate that the listener could in some manner reconstruct the true invariant underlying any given phoneme, namely the speaker's intended phonetic gestures. Although the /k/ sounds in *key* and *caw* are acoustically different, they both involve a burst of air produced with the back of the tongue in a given relationship to the velum (soft palate). The earliest version of such an approach, the Motor Theory of Speech Perception (Liberman *et al.* 1967) proposed invariant motor commands underlying the articulatory gestures corresponding to phonetic units, but in a later form of the theory (Liberman and Mattingly 1985) the invariant feature was proposed

to be the more abstract intentional structures controlling articulatory movements. The heterogeneity and interdependence of gestural cues to a single phonetic unit however pose serious problems for the model (Klatt 1989).

More recently, there have been several related proposals which cast doubt on the necessity of phonemic representations in spoken-language recognition, or, in some cases, on the necessity for any intermediate representations at all. These proposals suggest that listeners are sensitive to various distinctive features of sounds (Elman and McClelland 1986); that there are no discrete units, but a continuous uptake of information relevant to the stored phonetic form of a word (Lahiri and Marslen-Wilson 1991; Marslen-Wilson and Warren 1994); or, most radically, that there are no single stored forms of words, but only a complete collection of memory traces of every earlier occurrence of words, against which incoming forms are compared for a process of recognition-by-analogy (Goldinger 1998). As a result, the theory of input representations for lexical access is currently once more a very active area.

5.3 Finding the constituent parts

The way in which spoken language differs most strikingly from written language, as we just saw, is that only in written text is clear information to be found (e.g. in spaces inserted between words) about the constituent units of which the text is composed. For a spoken message to be understood, however, the perceiver must find and recognize these discrete parts, because these are the jointly known and agreed building blocks of which the new message has been constructed by the speaker and can be reconstructed by the listener. In speech, words are not separated by discontinuities in the signal; they are uttered in a continuous stream, and coarticulation and other phonological assimilations may cross word boundaries. Likewise there is no necessary localized punctuation in speech to signal syntactic boundaries. Thus the listener's task involves computations with no counterpart for the reader (at least, for the reader of a text like this one).

As we foreshadowed in the introduction, the decisions which listeners eventually make regarding the constituent parts of incoming utterances may result to as great an extent from the word recognition and utterance interpretation processes themselves as from exploitation of explicit cues in the utterance form. Nevertheless, there is now a solid body of evidence that listeners can use aspects of the spoken form to determine word and syntactic boundaries. Prosodic structure—at one or another level—is closely involved in both. Listeners use the procedures which are summarized in this section, exploiting information included in the initial decoded representation of the utterance, to constrain aspects of both the lexical activation process (see section 5.4) and the interpretation process (see section 5.6).

To discover the constituent parts (such as words) of which continuous speech signals are composed, it would clearly be helpful if listeners were able to rely on explicit procedures which would enable them to locate where word boundaries are most likely to occur. Indeed, experimental evidence exists for explicit segmentation (i) into

syllables: listeners detect target strings such as *ba* or *bal* more rapidly when the strings correspond exactly to a syllable of a heard word than when they constitute more or less than a syllable (Mehler *et al.* 1981; Zwitserlood *et al.* 1993); and (ii) at stress unit boundaries: recognition of real words embedded in nonsense bisyllables is inhibited if the word spans a boundary between two strong syllables (i.e. two syllables containing full vowels), but not if it spans a boundary between a strong and a weak syllable, since only the former is a stress unit boundary (Cutler and Norris 1988).

A striking outcome of the explicit segmentation research is that segmentation units appear to differ across languages. The evidence for syllables reported above comes from French and Dutch. Evidence of syllabic segmentation has also been observed in Spanish (Bradley *et al.* 1993) and Catalan (Sebastian-Galles *et al.* 1992). Other tasks confirm the robustness of syllabic segmentation in French (Segui *et al.* 1981; Kolinsky 1992; Peretz *et al.* 1996). However, target detection does not show effects of syllabic segmentation in English (Cutler *et al.* 1986) or in Japanese (Otake *et al.* 1993). Cutler and Norris' (1988) observation that segmentation in English is stress-based, by contrast, is supported by patterns of word boundary misperceptions; for example, *a must to avoid* (in which only the second and last syllables are strong) is perceived as *a muscular boy* (Cutler and Butterfield 1992). Support also comes from evidence of activation of monosyllabic words embedded as strong syllables in longer words (e.g. *bone* in *trombone;* Shillcock 1990, Vroomen and de Gelder 1997).

This apparent asymmetry turned out to be in fact evidence of a deeper symmetry. Languages have many levels of structure, and one of these is rhythmic regularity. Yet rhythm is not the same for every language—there are several potential phonological levels at which regularity can be defined. (Such differences can be easily observed in the variation across poetic conventions used in different languages.) As it happens, the basic unit of language rhythm in French is the syllable, whereas the rhythm of English is stress-based. The most obvious reflection of this is in timing; syllables in French tend not to contract or expand, whereas in English unstressed syllables can be considerably compressed, and stressed syllables expanded, to maintain a perceived regularity in the occurrence of stress beats.

Given this parallelism, the evidence of stress-based segmentation in English and syllabic segmentation in French led to the hypothesis that the segmentation of continuous speech involved a universal strategy which exploited the rhythmic structure of speech input; apparent language-specificity in processing was simply due to different implementations of the rhythmic procedure for different language rhythms. Japanese offered a test case for the rhythmic hypothesis because it has a different kind of rhythm than the languages which had previously been tested. In Japanese, the unit of rhythmic regularity is the mora, a subsyllabic unit which can be a vowel, a vowel plus an onset, or a syllable coda. Thus the Japanese name *Honda* consists of three morae: *ho-n-da;* Japanese poetic forms are defined in terms of morae (seventeen morae in a haiku, for example).

Otake *et al.* (1993), and Otake *et al.* (1996a), using the fragment detection methodology that had been used by Mehler *et al.* (1981) and Cutler *et al.* (1986), found that

Japanese listeners indeed segmented speech most easily at mora boundaries. The targets were, for example, *ta* or *tan* in words such as *tanishi* or *tanshi,* each of which consists of three morae: *ta-ni-shi, ta-n-shi.* The target *ta* (which corresponds to the first mora of each word) was detected equally rapidly and equally accurately in both types of word. The target *tan* was hardly ever detected in *tanishi,* in which, in terms of mora structure, it is simply not present. Phonemes can constitute a mora by themselves: a nasal consonant in syllable coda position is moraic, and a vowel not preceded by a consonant is moraic; and Japanese listeners detect consonants and vowels significantly faster if they constitute a mora than if they do not. Thus /n/ is detected faster in *kanko* than in *kanojo,* and /o/ faster in *aoki* than in *tokage* (Cutler and Otake 1994; Otake *et al.* 19965).

The rhythm of a language is a part of prosodic structure. This means that it represents a level of organization above the segmental level; it may be expressed in supra-segmental structure—timing, for instance—but it may also have segmental expression. In English this is so: the stress-based rhythm of English is defined in terms of the pattern of strong and weak syllables, and, as we saw above, strong syllables are defined as those containing full vowels, whereas weak syllables contain reduced vowels. English listeners are thus using segmental information—vowel quality—to drive their hypotheses about plausible word-boundary locations; yet the comparison with how such hypotheses are generated in other languages reveals that this behaviour arises because of the role of vowel quality in encoding rhythmic structure. Across languages, listeners exploit rhythm for lexical segmentation.

This is not the only heuristic which listeners can call upon in finding probable word boundary locations. There are language-specific effects such as the exploitation of vowel harmony information in Finnish (Suomi *et al.* 1997), and other general effects such as those based on phoneme sequencing constraints (McQueen 1998)—the sequence /mr/ cannot occur within a syllable, therefore there must be a syllable boundary between the two phonemes. Further, Norris *et al.* (1997) showed that recognition of real words embedded in nonsense strings is inhibited if the remainder of the string, once the word has been extracted, could not possibly itself be a word. English listeners in their study were presented with words like *egg,* embedded in nonsense strings like *fegg* and *maffegg.* In *fegg,* the added context /f/ is not a possible word of English—there are no English lexical items consisting of a single consonant. In contrast, the added context *maff* in *maffegg,* although it is actually not a word of English, might conceivably have been one—*mat, muff*and *gaff are* all English words. The listeners were faster and more accurate in detecting real words embedded in possible-word than in impossible-word contexts; in other words, they appeared to be able to rule out any candidate segmentation which would postulate, elsewhere in the input, a residue which was unparseable into words. Sections 5.4 and 5.5 will discuss in further detail how these effects feed into the process of recognizing words.

Words are the known constituent units of utterances, but there are higher levels of grouping of the words within any utterance as well. Section 5.6 will deal with the processing of syntactic and semantic structure. Cutler *et al.* (1997), reviewing the

literature on the processing of prosodic structure, conclude that cues in the pitch contour of the utterance which signal a break, or cues in relative word prominence which signal an accent, can have an effect upon syntactic processing, by leading listeners to prefer potential analyses consistent with the prosodic information provided. But listeners cannot simply rely on the prosody of utterances to provide them with the necessary syntactic information or discourse-structure information—for the very good reason that prosodic structure is not directly isomorphic with these higher levels of utterance structure. Nevertheless, placement of sentence accent, or marking of a syntactic boundary via pitch movement, can result in marked effects in the speech signal, which in turn can be exploited by listeners even at the phonetic processing and word-recognition levels. For instance, the pitch movement associated with a boundary may be followed by a pitch reset, which will have the effect of ruling out coarticulation and hence make a boundary between words clearer, or the utterance of a word which is accented will be clearer and less likely to contain variant forms of the word's constituent phonemes. These effects are reviewed by Cutler *et al.* (1997).

5.4 Activating lexical representations

The recognition of spoken words differs from word reading not only in the lack of clear segmentation of the input into its constituent units, but also, and most clearly, in the temporal aspect of the input. Words do not arrive at the peripheral input stage all at once—they are presented over time, the beginning arrives first, the end arrives last. As Section 5.3 showed, listeners are adept at exploiting their knowledge of language phonology to circumvent the potential problems caused by the continuity and variability of speech signals; thus the task of recognizing spoken words might seem to be a matter merely of matching the incoming sequence to the stored word forms in the listener's mental lexicon.

Unfortunately there is another problem for the listener, and that is simply the size of the listener's vocabulary compared with the size of the set of phonetic components from which it is constructed. A listener's vocabulary contains tens of thousands of words (although of course the relevant measure here is not words as they are orthographically defined, in texts such as this one by spaces between the printed word forms; the relevant measure is sound-meaning mappings in the mental lexicon, and these will exist in comparable numbers for speakers of uninfected languages like Chinese. morphologically simple languages like English, or highly agglutinating languages like Turkish).

The words, however, are built up out of a repertoire of on average only 30-40 phonemes (Maddieson 1984). It requires only simple mathematics to realize that words are not highly distinctive. Any spoken word tends to resemble other words, and may have other words embedded within it (thus *steak* contains possible pronunciations of *stay* and *take* and *ache,* it resembles *state* and *snake* and *slack,* it occurs embedded within possible pronunciations of *mistake* or *first acre,* and so on). Computations of the amount of embedding in the vocabulary by Frauenfelder (1991; for Dutch) and

McQueen and Cutler (1992; for English) have shown that a majority of polysyllabic words have shorter words embedded within them. Moreover, these embedded words are most likely to appear at the onsets of their matrix words; Luce (1986) computed that, when frequency is taken into account, more than one-third of short words in English could not be reliably identified until after their offset (and experimental studies of the perception of incrementally presented words by Grosjean (1985) and Bard *et al.* (1988) have confirmed that this does form an actual problem for listeners). *Stay* could become *steak, steak* could become *stokehold,* and so on. So how do listeners know when to recognize *steak* and when not?

5.4.1 Concurrent activation and competition

The solution to this problem is a fundamental notion which is now accepted by nearly all researchers in the field of spoken-word recognition; candidate words compatible with (portions of) an incoming speech signal are simultaneously activated and actively compete for recognition. Concurrent activation has been a feature of all models of spoken-word recognition since Marslen-Wilson and Welsh's (1978) cohort model. Competition was first proposed in the TRACE model of McClelland and Elman (1986), and in the same form—competition via lateral inhibition between competitors—forms the central mechanism of the Shortlist model (Norris 1994). In other forms it is also found in the other main models currently available, such as the Neighbourhood Activation Model (Luce *et al.* 1990) and the re-revised cohort model (Gaskell and Marslen-Wilson 1997).

There is substantial evidence of activation of words embedded within other words (Shillcock 1990; Cluff and Luce 1990), and of simultaneous activation of partially overlapping words (Goldinger *et al.* 1989; Zwitserlood 1989; Marslen-Wilson 1990; Goldinger *et al.* 1992; Gow and Gordon 1995; Wallace *et al.* 1995). Although such evidence is consistent with the competition notion, it does not entail it. Inhibition of recognition as a function of the existence of competitors provides direct evidence. Taft (1986) observed that non-words which form part of real words are hard to reject. Priming studies by Goldinger *et al.* (1989) and Goldinger *et al.* (1992) suggested that recognition may be inhibited when words are preceded by similar-sounding words, the inhibition being presumably due to competition between the preceding word and the target. Direct evidence of competition between word candidates comes from a study by McQueen *et al.* (1994), who found that word-spotting latencies were significantly longer in nonsense strings which activated competing words; that is, *mess* was harder to find in *domess* (which could partially activate *domestic)* than in *nemess* (which activates no competitor). Similarly in Dutch *zee* (sea) is harder to spot in *muzee* (which can be continued to form *museum)* than in *luzee* (Donselaar *et al.* 1998). Norris *et al.* (1995) and Vroomen and de Gelder (1995) showed further that the more competing words may be activated, the more the recognition of embedded words will be inhibited.

As this last result emphasizes, analysis of patterns of competition depends crucially on precise knowledge of vocabulary structure. Studies of lexical structure have been revolutionized in recent years by the availability of computerized dictionaries; it is now

easy to analyse the composition of the vocabulary in many languages, and arguments based on analyses of lexical databases have played an important role in theorizing about spoken-word recognition for the past decade (e.g. Marcus and Frauenfelder 1985; Luce 1986; Cutler and Carter 1987). It should be noted, however, that substantial corpora of spoken language, and the estimates of spoken-word frequency which could be derived from them, are still lacking; such spoken-word frequency counts as exist to date (e.g. Howes 1966; Brown 1984) are, for practical reasons, small in scale compared to written frequency counts.

Competition between candidate words which are not aligned in the signal provides a potential mechanism to achieve segmentation of the speech stream into individual words. Thus although the recognition *of first acre* may involve competition from *stay, steak,* and *take,* this will eventually be overcome by joint inhibition *fromjirst* and *acre.* However, competition can also co-exist with explicit segmentation procedures of the type described above in section 5.3. When inter-word competition and stress-based segmentation are compared in the same experiment, independent evidence appears for both (McQueen *et al.* 1994). In section 5.3, we further described a prelexical effect in which listeners display sensitivity to the viability of stretches of speech as possible word candidates. When this Possible Word Constraint is incorporated in the Shortlist model, the model accurately simulates not only the experimental findings which motivated the constraint, but also a range of other experimental demonstrations of competition and explicit segmentation effects (Norris *et al.* 1997). In other words, such prelexical segmentation effects can be thought of as exercising constraints on the activation and competition process.

The way in which segmental information contributes to word-candidate activation has been the focus of much recent experimental attention. Connine *et al.* (1997) found that phoneme-monitoring for phonemes occurring at the end of non-word targets is faster the more similar the non-word is to a real word: /1/ was detected faster at the end of *gabinet* (which resembles *cabinet)* than at the end of *shuffinet* (which is less close to any existing word). This suggests that even partial information for a word, as present for instance in a non-word which resembles that word, will activate lexical information.

Marslen-Wilson and Warren (1994), following up the work of Streeter and Nigra (1979) and Whalen (1984, 1991) mentioned in section 5.2 above, examined the differential effects of subphonemic mismatch in words and non-words. They constructed three experimental versions of matched pairs of words and non-words like *job* and *smob,* by cross-splicing different initial consonant-vowel sequences (CVs) onto the final consonant of each item. The CV could either be from another token of the same word/non-word, from another word *(jog* or *smog),* or from another non-word *(jod* or *smod).* Marslen-Wilson and Warren performed lexical decision and phonetic decision experiments; in both of these tasks listeners were sensitive to a mismatch between CV and final consonant (e.g. a token *of job* in which the *jo-* had been taken from yog and therefore contained formant transitions into a velar place of articulation in the later part of the vowel). However, the effect of a mismatch on non-words was much greater

when the CV came from a word than from another non-word, whereas for words, whether the CV came from another word or from a non-word had very little effect. McQueen *et al.* (in press) report similar studies manipulating the conditions under which these competition effects could be made to come and go.

All the results described in this section show that activation of lexical representations is a continuous process, based on whatever information is available. Even partial information (in partial words, for instance, or in non-words which in part overlap with real words) suffices to produce partial activation. Thus *domess* briefly activates *domestic; smob* briefly activates *mob* and *smog.* Activation of a lexical representation does not obligatorily require full presentation of the corresponding word form; the competition process, and its concomitant constraints, can so efficiently result in victory for words which are fully present in the signal, that concurrent activation of partially present words, or of words embedded within other words, is simply a low-cost by-product of the efficiency with which the earliest hints of a word's presence can be translated into activation.

5.4.2 Segmental versus suprasegmental information

The above studies showed clear evidence for continuous activation of potential candidate words based on the incoming segmental information. The discussion in section 5.2 above showed that there is controversy as to whether candidate-word activation proceeds via an explicit representation of phonetic segments; segmental information is encoded in the signal, and fully unravelling the code may not be necessary. However, the signal also contains suprasegmental information—variations in fundamental frequency, amplitude, and duration of the constituent parts, and this information may also be relevant to word identity.

Because Fig. 5.2 represents an utterance in English, it does not manifest suprasegmental contrasts with great clarity. The word *coconut* has a stressed first syllable and a weak second syllable, with a reduced vowel; orthography notwithstanding, the vowels in the two syllables are quite different. The first syllable is also nearly twice as long (from about 0.52 to 0.67) as the second (from 0.67 to 0.75, approximately). Other languages, however, offer more obvious contrasts. In the tone language Cantonese, for instance, a single CV syllable such as [si] can be realized with six different tones, and all possible realizations exist, with different meanings—some with multiple meanings, in fact. With tone 1 (high level) [si] means 'poem', with tone 2 (high rising) it means 'history', with tone 6 (low level) it means 'time', and so on. Tone distinctions are realized in the fundamental frequency contour of an utterance (F_0 height and F_0 movements), although tone and syllable duration do covary (Kong 1987; Kratochvil 1971), and tones may be distinguished by the timing of their movement within a syllable (Shen and Lin 1991). In Japanese, fundamental frequency distinctions also play a role in distinguishing between words; thus *ame* with a high-low (HL) pitch accent pattern means 'rain', *ame* with a LH pattern 'candy'.

Although stress is part of the acoustic realization of every polysyllabic word in English, there are remarkably few pairs of English words which are distinguished only

by differences in suprasegmental structure: *FOREgoing* versus *forGOing*, *TRVSTy* versus *trustEE*, and a handful more (upper case here signifies stress). There are many more pairs like *SUBject/subJECT* or *REcord/reCORD*, which differ in segmental as well as in suprasegmental structure. The vowels in the latter word pairs, especially in the first syllables, are quite clearly different, just as are the first two vowels in *coconut* in Fig. 5.2. Stress in English is in fact expressed as much in the segmental structure of words (stressed syllables must have full vowels, while reduced vowels must be unstressed) as in the suprasegmental structure. Correspondingly, the segmental (vowel quality) distinctions involved in stress contrasts seem far more crucial to English listeners than the suprasegmental distinctions; cross-splicing vowels with different stress patterns produces unacceptable results only if vowel quality is changed (Fear *et al.* 1995). Studies of'elliptic speech'—speech containing some systematic segmental distortion—showed that the manipulation which most inhibited word recognition was changing full vowels to reduced and vice versa (Bond 1981). Slowiaczek (1990) found that mis-stressing *without* resulting change in vowel quality had no significant effect on the identification of noise-masked words; and Small *et al.* (1988) and Taft (1984) found that such mis-stressing also had no effect on detection time for following phonemes. On the other hand, Bond and Small (1983) found that mis-stressed words *with* vowel changes were not restored to correct stress when listeners repeated a heard text at speed (indicating that subjects perceived the mis-stressed form and may not at all have accessed the intended word).

If, as this evidence combines to suggest, English listeners do not use suprasegmental information in activating word candidates, then pairs like *FOREgoing* and/orGO/wg, distinguished only in suprasegmental structure, will be functionally homophonous: both *FOREgoing andforGOing* should be activated whenever either of them is heard. Indeed Cutler (1986) showed that listeners did not distinguish between these two word forms in initially achieving access to the lexicon.

The situation is quite different, however, in other languages. In tone languages. tonal information may be crucial for determining word identity. A categorization experiment by Fox and Unkefer (1985), using a continuum varying from one tone of Mandarin to another, confirms that listeners use tone to distinguish words: the crossover point at which listeners switched from reporting one tone to reporting the other shifted as a function of whether the CV syllable upon which the tone was realized formed a real word when combined only with one tone or only with the other tone (in comparison to control conditions in which both tones, or neither tone, formed a real word in combination with the CV). The lexical effect appeared only when the listeners were Mandarin speakers; English listeners showed no such shift, and on the control continua the two subject groups did not differ. Lexical priming studies in Cantonese also suggest that the role of a syllable's tone in word recognition is analogous to the role of the vowel (Chen and Cutler 1997; Cutler and Chen 1995); in auditory lexical decision, overlap between a prime word and the target word in tone or in vowel exercised parallel effects. On the other hand, there is evidence from a variety of experiments on the processing of Chinese languages that the processing of tonal information may be

more error-prone than the processing of segmental information (Tsang and Hoosain 1979; Taft and Chen 1992; Cutler and Chen 1997). This suggests that suprasegmental information does constrain word activation in Chinese languages, but the effect of suprasegmental information may be weaker than that of segmental information.

Pitch accent in Japanese words can also be processed efficiently to constrain word activation at an early point in presentation of a word. Cutler and Otake (1999) presented Japanese listeners with single syllables edited out of bisyllabic words differing in accent pattern; listeners were able to determine, with great accuracy, whether the syllable came from a word in which it had high or low pitch accent. Interestingly, their scores were significantly more accurate for initial (80% correct) than for final syllables (68%). This suggests that pitch accent information is realized most clearly in just the position where it would be of most use for listeners in on-line spoken-word recognition. They then tested this suggestion in a gating experiment using pairs of Japanese words such as *nimotsu/nimono,* beginning with the same CVCV sequence but with the accent pattern of this initial CVCV being HL in one word and LH in the other. Fragments of the word extending no further than the first vowel («/-) were sufficient to produce word guesses which correctly reproduced the initial accent patterns of the actually spoken words with a probability significantly above chance. Thus Japanese listeners can exploit pitch-accent information effectively at an early stage in the presentation of a word, and use it to constrain selection of lexical candidates. The strong dependence of pitch accent realization on dialect in Japanese, however, suggests that again, segmental information may be accorded priority in constraining word activation.

Thus both tonal information and pitch accent information are used by listeners in word activation, even though the evidence from English showed that stress information was not exploited in this way. Even in other stress languages, however, the situation turns out to differ from English. In Dutch, for example, mis-stressing a word can prevent lexical activation. The competition effect described above, in which *zee* (sea) is harder to spot in *muzee* (which can be continued to form *museum)* than in *luzee* holds only *it muzee* is, like *museum,* stressed on the second syllables. *If muzee* and *luzee* are stressed on the initial syllable then there is no longer a significant difference between them in detection time for *zee,* suggesting that there was in this case no competition from *museum* because it simply was not activated by input lacking the correct stress pattern (Donselaar *et al.* 1998). In Dutch, at least, there may be on-line directive use of stress information in lexical access, and this in turn suggests that the failure to find similar evidence in English may arise from the peculiar redundancy of purely prosodic cues to stress in English; stress information can nearly always be derived from segmental structure.

5.5 Retrieving lexical information

Once a word form has triumphed in competition over its rivals, what information does it bring with it from the lexicon for integration into the representation which the

listener is forming of the utterance as a whole? Psycholinguistic research has lavished attention on some aspects of this question, while almost ignoring others. The experimental evidence on morphological and semantic information in the lexicon is summarized here; section 5.6.1 below discusses the syntactic and thematic information which may be made available by lexical retrieval.

5.5.1 Morphological structure

The stored forms of words in morphologically simple languages like English include considerable morphological detail; this conclusion can be drawn from the substantial literature investigating the role of morphological structure in word recognition. Recent models of the lexical representation of morphology have fallen into two general classes. On the one hand are models in which the stored representations consist of stems with the affixes with which they may combine; in such a model *count* would be stored as the head of an entry, and would be furnished with the prefixes *dis-, mis-, vis-, ac-,* and the suffixes *-s, -ed, -er, -able* etc. (see e.g. Caramazza *et al.* 1988; Marslen-Wilson *et al.* 1994). Contrasted with these are models in which full forms are separately represented but are linked with related forms (so that *count* and *counts* and *discount* and *counter* and *unaccountability* would all be stored forms, but linked to a common node; Schriefers *et al.* 1991; Baayen *et al.* 1997). McQueen and Cutler (1998), in a review of this literature, conclude that the evidence supports the latter type of model, with the additional specification that the links between morphological relatives are strong and that the stored word-forms do contain information about the morphological structure and relationships.

These relationships between morphologically complex words in English encode different types of linguistic connection. Thus inflection, for example of tense on verbs or number on nouns, as in *discount-ed* and *viscount-s,* contrasts with derivation, for example the addition of affixes especially to change word class, as in *account, account-able, accountabil-ity.* Yet it appears not to be linguistic relationships which determine the relative closeness of connections in the language user's lexicon. Instead, McQueen and Cutler (1998) conclude that the stored relationships are principally based on such factors as frequency of occurrence *(counts* is a more frequent form than *countering)* and semantic transparency *{count* and *counting* are more clearly related to one another than *discount* and *counter).*

The evidence thus suggests that in languages like English the recognition of any morphologically complex word will not involve obligatory decomposition of the word into its constituent morphemes, but that the full form will be activated by the incoming speech signal and will participate in the competition process as a whole. Importantly, however, the result of the recognition of a spoken word will be a form which brings with it information (such as word class and the fact of marking for tense, number, etc.) which can constrain the computation of the higher-level structure in which the word participates.

An important caveat must always be added, however, to the discussion of this body of research. English is in this instance not necessarily representative of the world's

languages. Thus the model of access and retrieval which holds for English (and similar languages) does not necessarily hold for languages with different structure. Even within morphologically similar languages, Orsolini and Marslen-Wilson (1997) have proposed, different processing principles may be warranted. It is certainly conceivable that word recognition in Turkish and Finnish (languages with rich combinatorial morphology) might require affixes to be computed and accessed as separate entities, while word recognition in Chinese (a language with little affixal morphology) might provide little information which constrains the syntactic computation. There is not yet sufficient evidence to fill out the blueprint such that it would cover listeners in all languages.

5.5.2 Semantics

The meaning of a word is presumably above all what must be retrieved from the lexicon if the listener is to evaluate correctly the role that the word plays in the speaker's utterance. In fact research on precisely what becomes available in word meaning retrieval has more often been based on written input (see the chapter by Perfetti, this volume) than on spoken input. Much of the research described in section 5.4 above involves tasks which to a greater or lesser extent use meaning activation as a dependent variable (various forms of lexical decision; cross-modal priming). But while it seems relatively straightforward to envisage the meaning associated with the lexical representation of *egg* or *smog,* not all referential relations are so simple.

One of the semantic issues which has sparked as much attention in spoken-word recognition as in written-word recognition is the role of lexical ambiguity. The word *week,* for example, refers to a period of seven days. But exactly the same sequence of sounds encodes an adjective, meaning 'lacking in strength'. As in the reading domain, the principal questions concern whether both meanings are retrieved when an English-speaking listener hears [wik]; whether it matters that the two words differ in form class; and whether meaning retrieval depends upon degree of fit to the context of the rest of the utterance.

Studies with the cross-modal priming task have produced evidence for momentary simultaneous activation of all senses of an ambiguous spoken word, irrespective of relative frequency or contextual probability. Thus Swinney's (1979) original studies with this task showed that words such as *bug* facilitated the recognition of words related to both their senses, even when prior context was consistent with only one of the senses (a few syllables later, however, only the contextually appropriate sense remained active). Facilitation of words related to both senses occurred even when one reading of the ambiguous word was more likely simply because it had a higher frequency (thus *scale* primed both *weight* and *fish;* Onifer and Swinney 1981), and it occurred even when one reading was more likely because it had the word class required by the syntactic context (thus *week/weak* primed both *month* and *strong;* Lucas 1987). Tanenhaus and his colleagues (Tanenhaus *et al.* 1979; Seidenberg *et al.* 1982; Tanenhaus and Donnenwerth-Nolan 1984) also produced evidence for multiple-sense activation with a very similar task, and further support appeared from other kinds of

listening experiments. Thus Lackner and Garrett (1972) presented listeners with two competing messages, and required them to attend to one and to paraphrase it. Speech in the unattended channel (which subjects could not report), resolved ambiguities in the attended utterances; subjects' paraphrases reflected either sense, depending on the available disambiguation, again suggesting availability of all senses. And the task of naming the colour of a visually presented word, which becomes harder if the word's meaning is activated, was also used to show that both meanings of a spoken ambiguous word were available to exercise this interference (Conrad 1974; Oden and Spira 1983).

Later experiments (Tabossi 1988a; Tabossi *et al.* 1987) found however that strongly constraining contexts could lead to only one sense being activated if that particular sense was highly dominant (e.g. the weight sense of *scale* in a sentence about weighing). But again, these contexts effectively primed the relevant sense via occurrence of a related word—contexts which forced one sense but did not prime it (e.g. *On the table stood a scale)* produced facilitation for all senses. The current picture is therefore that all meanings of an ambiguous word are potentially available, but that contextually inappropriate meanings may in many circumstances have no chance to play a role in the recognition process.

The same picture can potentially be constructed for the various senses in which even an unambiguous word can be interpreted. Tabossi (19886) found that sentence contexts could constrain activation of different aspects of an unambiguous word's meaning; *hard* was primed after *The strong blow didn't crack the diamond,* but not after *The jeweller polished the diamond.* But other studies showed that all attributes may be momentarily activated when a word is heard, irrespective of their relative dominance and of their contextual appropriateness (Whitney *et al.* 1985). Shortly after word offset, however, attributes which are dominant and/or contextually appropriate are still active, but contextually inappropriate non-dominant attributes are not. Greenspan (1986) found that central properties of unambiguous words (e.g. that ice is cold) are activated irrespective of contextual appropriateness, but peripheral properties (e.g. that ice is slippery) may only be activated when appropriate. We will return to the issue of the relation between context and word meaning in section 5.7 below.

5.6 Interpreting the sequence

The comprehension process does not end with identifying words and their meanings. Determining what message a sequence of words conveys involves far more than simply adding together the meanings of the words. The sentence that contains them must be divided into its component parts, the relations between these parts must be determined and interpreted semantically, and the reference of the parts, their relation to ongoing discourse, and the truth or communicative force of the whole sentence or discourse must be determined. This process is guided by a language user's knowledge of the structure of his or her language, together with specific structural information made available by the particular words in a sentence. All this holds true whether reading or listening is involved.

We will first review what we take to be strong candidates for phenomena and processes common to reading and listening, focusing on data from reading experiments (see Perfetti, this volume, for more thorough coverage). We will then turn to phenomena and processes that may be specific to interpreting a heard sentence.

5.6.1 Processes common to listening and reading

The past three decades of study of sentence and text comprehension allow some strong conclusions. Readers and listeners often arrive at a semantic interpretation of a sentence in an apparently-incremental and nearly-immediate fashion. They do not wait for the end of a clause or sentence, but instead (to a first approximation) their understanding of a sentence seems to keep up with words as they are heard or as the eyes land on them. The understanding that they arrive at honours grammatical knowledge, even when it forces an unexpected or implausible interpretation.

While it is now clear that grammatical information must be used in sentence comprehension, researchers disagree about just what grammatical knowledge is used, at least in the initial stages of analysing or 'parsing' a sentence. Some researchers argue that a grammatical structure must be built first, in order to support semantic interpretation, and propose that only relatively global grammatical information (e.g. about possible phrase structure configurations or templates and about the part of speech of an individual word) is used to build such a structure (Frazier 1979; Frazier 1987; Frazier 1989; Frazier and Rayner 1982). Other 'lexicalist' theorists place similar emphasis on the creation of grammatical structures but suggest that a richer set of information about the usage of individual lexical items guides their construction (Abney 1989; Konieczny *et al.* 1997; MacDonald *et al.* 1994; Tanenhaus *et al.* 1990; Tanenhaus *et al.* 1993). This richer information can include both grammatical information (about, e.g. the possible argument structures assigned by a verb) and extra-grammatical information (about, e.g. the relative frequency of usage in different constructions, or the plausibility of the different constructions). Theorists also differ in their opinion about whether a single analysis is built and interpreted at a time, or whether multiple analyses are built and allowed to compete with one another.

Some of these theoretical approaches have led to the identification of important new phenomena of sentence comprehension. For instance, working in the phrase-structure parsing tradition, Frazier and Rayner (1982) claimed that a preposition phrase (PP) in the configuration V-NP-PP (e.g. *John hit the girl with the wart)* is initially taken as a complement of the verb (V) rather than a modifier of the noun phrase (NP). The example sentence is therefore read relatively slowly because it violates this initial preference, which Frazier and Rayner claimed reflects a preference for the simplest, most-quickly-constructed, syntactic analysis.

More recent work has made it clear that detailed lexical properties of verbs and referential properties of noun phrases (as well as syntactic simplicity) affect comprehension very quickly (cf. MacDonald *et al.* 1994; Tanenhaus *et al.* 1993). This research was stimulated by changes in linguistic theory over the past two decades that

accommodate substantial parts of syntactic functioning in the lexicon (including such approaches as Lexical Functional Grammar, Bresnan 1982; Head-driven Phrase Structure Grammar, Pollard and Sag 1994; and Pustejovsky's 1995 lexicalist approach). Psycholinguists have focused most on the argument structures and thematic structures made available by lexical items, usually verbs. The verb *cook,* for example, would allow argument structures with only an agent (the intransitive reading), or with both agent and theme (transitive reading). This information would become available upon retrieval of the word from the lexicon.

Marslen-Wilson, Tyler, and colleagues (e.g. Marslen-Wilson *et al.* 1988; Tyler 1989; Jennings *et al.* 1997) have provided evidence from listening experiments that verb-subcategorization information is available early in the process of sentence comprehension. They observed processing difficulty for sentences with subcategorization violations (e.g. *He slept the guitar,* compared for instance with the merely implausible *He buried the guitar;* the violation occurs because sleep cannot take a direct object). Subcategorization violations also caused greater difficulty than violations of selection restrictions (e.g. *He drank the guitar, drink* may take a direct object, but it must be something which can be drunk).

Spivey-Knowlton and Sedivy (1995) examined the effects of more detailed lexical information. They found that the advantage of a V complement interpretation (as observed by Frazier and Rayner 1982) seems to hold true only for action verbs. For perception and 'psych' verbs followed by an indefinite NP (e.g. *The salesman glanced at a customer with suspicion I ripped jeans),* modification of the NP is the preferred interpretation.

Properties of discourse as well as properties of lexical items also play a quick role in sentence comprehension. As one example, Trueswell and Tanenhaus (1991) showed that the classic *The horse raced past the barn fell* garden-path sentence (Bever 1970) no longer caused readers measurable difficulty when the temporal relations introduced by a discourse blocked the preferred main clause reading. Trueswell and Tanenhaus's subjects read sentences like *The student spotted by the proctor will receive a warning.* Normally, these sentences would be expected to be difficult, since a reader would initially take *The student* as the subject of *spotted.* However, if the discourse in which the sentence appeared specified a future time *{A proctor will come up and notice a student cheating),* this preference seemed to be replaced by a full willingness to take *spotted* as beginning a relative clause. The past tense interpretation of *spotted* was inappropriate for the future context, while the passive participle interpretation was acceptable.

5.6.2 Auditory sentence comprehension

One goal of a psycholinguistic theorist is to arrive at a model of a language user that explains how he or she can use the wide range of information provided by language in the course of understanding text or speech. Considering how people understand spoken as well as written language might seem simply to make the theorist's (and the

listener's) task harder. More different types of information must be accounted for. But in fact, considering what might be special about the listener's task provides some new insights into what the language user's skills really are. Recent research, using ways of looking at how auditory language is processed, has turned up very informative phenomena about language comprehension. For an extensive review of this research, see Cutler *et al.* 1997; for concentrated presentations of a selection of recent research, see the special issues of the journals *Language and Cognitive Processes* (volume 11, 1996, numbers 1 and 2) and *Journal of Psycholinguistic Research* (volume 25, 1996, number 2) devoted to prosody and sentence processing. In the present brief survey, we will consider ways in which the auditory modality might be expected to present additional challenges to the listener as well as ways in which the auditory modality might carry additional useful information.

5.6.2.1 Added challenges to the listener

We assume, in the absence of clear evidence to the contrary, that the architecture of the system that interprets auditory sentences is the same as that of the system that interprets written sentences. It is true, though, that auditory presentation sets this system some extra challenges. One challenge has already been described extensively; in listening, the words are not physically set apart from one another as they are in reading. It is clear that a listener has some means of identifying candidate words in the speech stream (just as it is clear that a reader *can* read words printed without spaces between them, albeit at a generally substantial cost in reading time; Rayner and Pollatsek 1996). However, the uncertainties of segmenting the word stream might be expected to interact in interesting ways with the uncertainties of interpretation that have been identified in research on reading.

Another challenge comes from the evanescent nature of speech. A listener cannot listen back to what he or she has just heard in the way a reader can make a regressive eye movement. Some researchers have suggested that this difference may play a major role in sentence comprehension. Watt and Murray (1996) claim that since 'auditory input is fleeting and not readily available for 'reinspection'' (p. 293), a listener may delay structural commitments until the end of a constituent. A reader, who can look back to recover from an erroneous early commitment, can afford to make such commitments. There are some reasons to discount this claim. First, readers look back rather infrequently, about 10 to 15 per cent of the time (Rayner and Pollatsek 1989). Second, several researchers have reported garden-path effects in listening experiments, indicating that listeners do sometimes make an erroneous early commitment (Carroll and Slowiaczek 1987; Pynte and Prieur 1996; Speer *et al.* 1996).

It is possible to take the opposite perspective and view listening as more basic and somehow 'simpler' than reading. For most people, reading is, after all, developmentally parasitic on listening. Some researchers have even suggested that some reading phenomena can be understood by claiming that skilled readers create (perhaps via implicit subvocalization) an auditory representation of what they are reading. Creating an auditory representation may facilitate some aspects of comprehension

(Slowiaczek and Clifton 1980); creating the right auditory representation may block miscomprehension (Bader 1994).

This perspective is encouraged by the observation that humans are adapted through evolution to process auditory language, not written language. One must assume that our brains are well-tuned to extract information from an auditory signal, and that our language is adapted to the capacities of our auditory system. Exactly what the relevant capacities are, however, is far from understood, especially at the levels of parsing and sentence interpretation. One reasonable candidate is the existence of auditory sensory memory, which may be able to span a period of time on the order of one or a few seconds (Cowan 1984). Contrary to the suggestion discussed above, heard language may persist for a longer period of time than read language, permitting more effective revision of analysis. Another candidate is the facilitating effects of auditory structuring on short-term memory; imposing a rhythm on the items in a list to be remembered can facilitate their memory (Glanzer 1976; Ryan 1969). Beyond carrying the information needed to recognize words, the auditory signal is richly structured in its melody and rhythm, its prosody. This structuring can certainly affect memory for language (Speer *et al.* 1993), and could serve as a source of information that might guide the parsing and interpretation of utterances.

5.6.2.2 Prosody in auditory sentence comprehension

The prosody of an utterance plays many roles. It can help in resolving lexical and syntactic ambiguities. It can signal the importance, novelty, and contrastive value of phrases and relate newly-heard information to the prior discourse. It can signal the attitude and affect of a speaker toward his or her topic. We will review selected recent research on some of these topics. Before doing so, however, we will turn to the topic of how one might describe the prosody of an utterance.

We will treat prosody as the structure that underlies the melody and rhythm of a sentence. Much recent work aimed at examining how the auditory signal can convey information has assumed an explicit analysis of prosody, an analysis that developed out of work done by Pierrehumbert (1980) (cf. also Beckman and Pierrehumbert 1986; Beckman and Ayers 1993; Ladd 1996; Selkirk 1984). Pierrehumbert devised an elegant description of English prosody. In her scheme, an utterance is viewed as a shallow hierarchy of prosodic elements. For present purposes, the elementary prosodic unit is the phonological (or intermediate) phrase, a string of speech that must end with a phrase accent (high, H-, or low, L-), and must contain at least one pitch accent (which can be high or low, H* or L*, or bitonal, e.g. L + H*). One or more phonological (or intermediate) phrases constitute an intonation phrase, which must end with a boundary tone (high, H%, or low, L%). An utterance can contain one or more intonational phrases. The end of an intonational phrase is signalled by pausing, lengthening, and segmental variation in addition to the presence of a phrase accent and a boundary tone, where the combination of phrase accent and boundary tone can appear in any of several forms, such as a 'continuation rise' or the normal 'declarative' contour. An intermediate phrase is typically associated with a smaller amount of pausing and lengthening than

an intonational phrase, and ends with a phrase accent but not a boundary tone. A pitch accent is associated with the stressed syllable of any word that receives focus-marking. The accent can be high or low, or moving, and generally falls on each word that is not treated as 'given' or predictable from context.

In our opinion, some explicit scheme for describing prosody must replace the vague, intuitive, and theoretically unmotivated descriptions psychologists have often used in the past. One such explicit scheme for coding the prosody of English sentences has developed out of the theoretical position sketched above. The scheme, called ToBI for 'Tones and Break Indices', is one that a researcher can learn with a reasonable amount of effort, since it is documented by a full training manual with examples (Beckman and Ayres 1993; Silverman *et al.* 1992; cf. Shattuck-Hufnagel and Turk 1996, for a brief introduction).

To see an application of ToBI analysis, consider Fig. 5.2 above. This acoustic representation of a sentence includes a pitch trace as well as an annotation of the pitch accents, phrase accents, boundary tones, and break indices (measures of the magnitude of a prosodic boundary) for the sentence *We all like coconut cream a lot.* This sentence contains just one intonational phrase and two phonological (intermediate) phrases. It has one maximal break (break index 4) at the end of the intonational phrase that ends the whole utterance, one substantial break at the end of the intermediate phrase within the sentence (break index 3), one less marked break (break index 2) after *all,* and a word-level break (break index 1) after each other word. The intonational phrase ends with a L% boundary tone preceded by a L— phrase accent and a L + H* pitch accent on *lot.* One acoustic reflection of the L + H* pitch accent can be seen in the pitch track at the bottom of the figure; the pitch of *lot* begins relatively low, but rises before falling again to the phrase accent and boundary tone. The remaining three pitch accents (on the stressed syllables of *all, coconut,* and *cream)* are simple H* accents, which are reflected in relatively high values of the pitch track.

Doing a ToBI analysis is not an automatic procedure. The elements of an analysis have neither invariant acoustic signals nor invariant syntactic associations that would unambiguously signal their identity. Nonetheless, training in the ToBI system does permit researchers to provide a rich, informative, and consistent description of the materials they are studying.

Once prosody has been described, psycholinguists can ask how it functions in language comprehension. Prosody can convey a speaker's attitude and emotion, it can help integrate a sentence into the proceeding discourse, and it can disambiguate otherwise ambiguous sentences. Consider the last function first. Some of the ambiguities that affect reading can disappear in listening. A typical student's first response to seeing the 'late closure' garden path sentence *Because John ran a mile seemed short* (Frazier and Rayner 1982) is that the possible misinterpretation would be blocked by speaking the sentence (or by putting a comma after *ran* in its written version). There are experimental demonstrations that speakers can provide cues that resolve such ambiguities as *The old men and women stayed home* (Lehiste 1973; Lehiste *et al.* 1976; were the women who stayed home old?). It is interesting that speakers may provide

markedly more adequate cues when they are given clear reasons to do so—for example if the contrast they are supposed to disambiguate is made clear to them (Allbritton *et al.* 1996; Lehiste 1973; Wales and Toner 1979).

This observation means that a speaker has some options in what prosody to assign to an utterance, and reflects the important point that there is not a one-to-one mapping between syntax and prosody (Selkirk 1984, 1995; cf. Shattuck-Hufnagel and Turk 1996, for a review). A given syntactic structure can have multiple acceptable prosodic realizations, and a given prosody can be ambiguous between two or more syntactic structures. One can legitimately convey the same message by saying *The woman sent the gift to her daughter, The woman ' sent the gift to her daughter,* and *The woman sent the gift * to her daughter* (intonational phrase breaks marked by '"). Not all possibilities are legitimate, though. Selkirk (1984) notes that sentences like *The woman gave " the gift to her daughter* violate what she calls the "Sense Unit Condition'. Conversely, one can convey either the message that the cop or the robber had a gun with the utterance *The cop shot the robber with a gun* (as well as several of its prosodic variants). While not all ambiguities can be eliminated prosodically, we can still legitimately ask what kinds of ambiguities can be resolved by what prosodic information, and we can ask how the processor uses this information.

One common goal of early work on prosody was to map out what sorts of ambiguities could be resolved in spoken language (e.g. Wales and Toner 1979). Success in reaching this goal was limited. It is not too much of a caricature to say that the basic conclusion was, if you want to get across a weird interpretation, say the sentence in a weird way. A more enduring suggestion of the early work is that some ambiguities of how the string of words could be broken up into phrases ('bracketing ambiguities', as *old men and women)* could be disambiguated prosodically, but alternative syntactic category membership of the words or phrases ('labeling ambiguities', as *visiting relatives can be a nuisance)* could not (Lehiste 1973).

This early work suffered from the lack of an adequate and explicit way of describing prosody, and it suffered from limitations of the then current syntactic analyses with their heavy emphasis on a distinction between deep and surface structure. However, it did point to important effects of the presence of prosodic boundaries at potential syntactic boundaries. It established the existence of acoustic correlates of major syntactic boundaries (e.g. lengthening and greater $F_{(1}$ movement; Cooper and Paccia-Cooper 1980; Cooper and Sorenson 1981), and demonstrated that listeners can make use of these cues. In fact, some researchers interpreted the apparent limitation of prosodic disambiguation to bracketing ambiguities to suggest that prosodic boundaries provide the only prosodic information that is used in disambiguating ambiguous sentences (e.g. Lehiste 1973; Nespor and Vogel 1986). Price *et al. (1991)* present a particularly strong argument for this position, suggesting that only major intonational phrase breaks (in the ToBI system intonational phrase boundaries, as opposed to intermediate phrase boundaries) will successfully disambiguate strings like *Mary knows many languages you know.*

More recent work suggests that this claim is too strong. Speer *et al.* (1996) (cf. Kjelgaard and Speer, in press) studied sentences like (1).

(1) a. Whenever the guard checks' the door * it's locked.

b. Whenever the guard checks * the door * is locked.

They found that placing either an intonational phrase boundary or a less salient phonological phrase boundary at one of the points marked by a ' effectively disambiguated the sentence. These sentences are sometimes referred to as 'late closure' ambiguities, because of Frazier's (1979) analysis of the preference for (1 a) in terms of her late closure strategy. The ambiguous NP, *the door,* is preferentially taken as the object of the first, subordinate clause verb, *checks.* Speer *et* a/.'s (1996) work shows that placing either kind of boundary in the appropriate position (after *the door* for (la), before for (lb)) affects parsing preferences, when compared to placing the boundary in the other position.

Schafer *et al.* (1996) provided evidence that at least one kind of syntactic ambiguity can be disambiguated by placement of a pitch accent without changing the prosodic phrasing. They studied sentences like (2), in which the relative clause *that we bought yesterday* could legitimately modify either the first (2a) or the second noun (2b). They found that puttinga H* pitch accent (indicated by uppercase letters) on one of these two nouns made it more likely to be chosen as the host for the modifying relative clause.

(2) a. We already have to repair the TIRE of the bicycle that we bought yesterday.

b. We already have to repair the tire of the BICYCLE that we bought yesterday.

Given that at least some aspects of prosody can effectively resolve syntactic ambiguities, we can ask how they have their effect. One suggestion that was made earlier can be rejected. It might be that prosodic disambiguation is asymmetrical, so that a marked prosody can convey a marked structure but no prosody could disambiguate in favour of a normally-preferred structure. Speer *et* a/.'s (1996) work used a baseline prosody, without a break either before or after the ambiguous NP *(the door),* as well as the two prosodic patterns shown earlier in (1). This baseline was judged to be equally appropriate for either interpretation *(the door* as object of the first verb, or subject of the second). Using two different techniques (end-of-sentence comprehension time, and the time taken to name a visual probe that was a legitimate or an illegitimate continuation of the sentence; cf. Marslen-Wilson *et al.* 1992), Speer *et al.* reported both facilitation and interference as a result of different placements of a prosodic break, compared to the baseline condition.

Another question is whether prosody is used on-line to determine initial analysis, or simply after-the-fact to guide revision of an otherwise-preferred analysis that turned out to be grammatically or pragmatically inappropriate. Pynte and Prieur (1996) provide the most recent statement of the revision-support proposal as one of two possible accounts of their data on time taken to identify the occurrence of a target word in a prosodically-appropriate or inappropriate sentence. However, the proposal does

not offer an attractive account of how prosody can disambiguate utterances that are fully ambiguous apart from prosody. Research on 'on-line' effects in auditory sentence processing may also provide evidence against the proposal. Marslen-Wilson *et al.* (1992) played their subjects an auditory string that, apart from prosody, was ambiguous between NP- and S-complement interpretations (3).

(3) The teacher noticed one girl from her class … WAS

The phrase *one girl from her class* is temporarily ambiguous between being the direct object of *notice* and the subject of a yet-to-appear complement sentence. Marslen-Wilson *et al.* measured the time to name a probe word *(was* in example (3)) when the string had been recorded with *one girl...* as part of a sentence complement and when it had been recorded with *one girl...* as direct object. Note that the word *was* fits with the sentence complement analysis (where *was* can play the role of verb to the subject *one girl...*); it does not fit with the direct-object analysis. Times were faster when the probe word fit with how the sentence was recorded, strongly suggesting that the listener used prosody to help in analysing the structure of the sentence.

This evidence does not fully rule out Pynte and Prieur's (1996) revision-support account of prosody. Watt and Murray (1996) provide some methodological criticisms of the Marslen-Wilson *et al.* experiments and present data suggesting that they may be replicable only under severely constrained conditions. Further, it is not inconceivable that any effects observed using this task reflect revision processes invoked in trying to fit the probe word into the sentence. Clearly, better on-line research techniques are needed before the issue can be considered settled (cf. Ferreira *et al.* 1996, for further discussion).

Even if experimental evidence is not yet adequate to demonstrate conclusively that parsing decisions (not just parsing reanalysis processes) are guided by prosody, it is interesting to consider the possible ways in which prosody could guide parsing. One way, implicit in much early research, is for prosody to provide local cues. A prosodic break, for instance, could be a local signal to terminate a syntactic phrase (Marcus and Hindle 1990). An alternative hypothesis is that the listener constructs a full prosodic representation, presumably along the lines described by Pierrehumbert (1980), and this representation serves as one input to the parser (cf. Slowiaczek 1981, for an early precursor to this proposal; see Schafer 1996, for a careful examination of the hypothesis and comparisons with other hypotheses; see Beckman 1996, for an analysis of how the prosodic representation might be constructed from the speech signal).

Schafer (1996) presents some evidence in favour of the full prosodic representation hypothesis combined with the concept of'visibility' (cf. Frazier and Clifton 1998). which claims that syntactic nodes within the current phonological phrase are more visible than nodes outside it, and hence preferred as attachment sites. She demonstrated fewer VP interpretations (47 vs. 64 per cent) of sentences like (4)—interpretations in which the prepositional phrase *with a mean look* is taken to modify the verb rather than the noun—when a phonological phrase (PPh) boundary intervened

between *angered* and *the rider* (4a) than when it did not (4b; IPh denotes intonational phrase boundary).

(4) a. (The bus driver angered L-)$_{PPh}$ (the rider with a mean look L-)$_{PPh}$ (L%)$_{IPh}$

b. (The bus driver angered the rider with a mean look L-)$_{PPh}$ (L%)$_{IPh}$

This finding would not be predicted by a local cue mechanism, since the phonological phrase boundary did not occur at a point of ambiguity or a point where a phrase could be ended, even though it did contribute to the full prosodic description of the utterance (note, all content words except *driver* had a H* accent in Schafer's materials).

A full prosodic representation may play a role in interpreting sentences semantically as well as integrating them into discourses. In other research, Schafer (1996) presents evidence that intonational phrases (rather than phonological or intermediate phrases, which she claims play a role in parsing) are the domains within which semantic interpretation is completed. Listeners presented with an ambiguous word like *glasses* seem to have committed more fully to its preferred meaning when an intonational phrase boundary intervenes between the ambiguous word and its disambiguation than when a phonological phrase boundary does. The presence of the intonational phrase boundary increased the amount of disruption in end-of-sentence comprehension time when the utterance forced *glasses* to be analysed in its unpreferred (spectacles) sense.

While only a modest amount of research indicates that prosody plays a role in semantic interpretation, there is ample evidence that it figures importantly in how pragmatic factors affect the construction of a discourse interpretation. Prosody highlights the information in an utterance that is salient to the discourse as it has developed (Bolinger 1978). For instance, it is appropriate to place a pitch accent, signalling focus, on the phrase that answers a vvA-question (thus, *GEORGE bought the flowers* but not *George bought the FLO WERS* appropriately answers the question *Who bought the flowers?).* Accented words, as well as words on which focus is appropriately placed, are identified faster (as measured by a phoneme-detection task) than non-accented words (Cutler and Fodor 1979; Cutler and Foss 1977), as well as being better remembered in their surface form (Birch and Garnsey 1995). They are taken as 'new' as opposed to 'given' (Chafe 1974; Halliday 1967). If a phrase that should be treated as given receives accent, comprehension can be disrupted; failing to place a pitch accent on a new phrase seems to disrupt comprehension even more (Bock and Mazella 1983; Nooteboom and Terken 1982; Terken and Nooteboom 1987). Going beyond the given/new contrast, placing a pitch accent on a phrase that selects between two contrasting possibilities in a discourse context can facilitate comprehension. Sedivy *et al.* (1995, see also Eberhard *et al.* 1995) showed that listeners who were told to select the *LARGE red square* selected it rapidly when the options were a large red square, a small red square, a large blue circle, and a small yellow triangle. The accent on *LARGE* was apparently interpreted as contrastive, allowing the listener immediately to select the one figure that contrasted with another in size.

The use of prosody in discourse interpretation is guided by the listener's knowledge of the prosodic structure of his or her language, not just by a crude principle such as

'important words are accented'. For instance, Birch and Clifton (1995) replicated Bock and Mazzella's (1983) finding of faster comprehension and higher prosodic acceptability judgements when focus fell on new information than the given information contained in the answer of a question-answer pair. They went further, though, by demonstrating that not every piece of new information in the focused phrase had to receive a pitch accent. Following a question like *What did Tina do when the neighbours were away?*, listeners were as quick and accurate at understanding the answer *She walked the DOG* , where only *dog* receives pitch accent, as the answer *She WALKED the DOG*, where both pieces of new information receive accent. This follows from Selkirk's (1984, 1995) theory of focus projection in English. According to this theory, an English listener's knowledge of language permits FOCUS to spread from a pitch-accented argument of a phrase *(the dog)* to the unaccented head of the phrase *(walked)*, and then to the whole phrase. Since the whole phrase receives FOCUS (even without all being accented), the whole phrase can be treated as new information. And since this is a property of English language structure, it shows that the effects of prosody are mediated by the listener's knowledge of language structure, perhaps by the creation of a full prosodic representation.

We will close this section by mentioning briefly two other discourse roles of prosody. First, prosody is clearly relevant to the interpretation of anaphors. *Mary hit Sue and then she BIT her* is surely different from *Mary hit Sue and then SHE bit HER* (cf. Solan 1980). Further, as discussed by Cutler *et al.* (1997), there may be a close tie between unaccented words and anaphoric devices generally: both are used to refer to entities already introduced into the discourse. Finally, as also discussed by Cutler *et al.*, prosody can be used to impose structure on entire discourses. It can be used to signal, among other things, the introduction of a new topic or the end of an old one, or even the end of a speaker's turn.

5.7 The architecture of the listening system

The process sketched in Fig. 5.1 converts a spoken input to a representation of meaning. We have drawn it as encompassing various levels of processing, with a unidirectional flow of information from the input of sound to the output of utterance meaning. But the flow of information in the process of comprehension has been a fiercely disputed topic in Psycholinguistics. Thus there is an abundance of experimental evidence pertaining to the question of autonomy versus interactivity of the various operations described in the preceding sections. In particular, the relationship of prelexical processing to lexical information, and of syntactic processing to information from the semantic and discourse context, have been the object of research attention.

Boland and Cutler (1996) have pointed out that current models of spoken-language comprehension can no longer be crudely characterized as in general interactive, or in general autonomous. Computational implementation, and refinement of model specification, has meant that it is necessary to consider the relationships between

individual sub-components of each model; models may allow various degrees of interaction or autonomy and these may differ across processing levels. In this final section we consider the directionality of the flow of information in particular parts of Fig. 5.1.

5.7.1 Decoding, segmenting, and lexical processing

Space considerations prohibit even a summary of the enormous literature on the question of whether lexical information constrains prelexical processing. A recent review of this literature (Norris *et al.* 1998) concludes, however, that there is no necessity for models of this aspect of the listening process to include top-down connections—that is a reversal of the information flow, wherein lexical processing passes information back to affect the decoding processes etc. The literature in question contains numerous cases in which experimental findings have been held to warrant top-down information flow, but in which subsequent experimental or theoretical work has shown this claim to be unjustified.

One such case history concerns compensation for coarticulation (a shift in the category boundary for a particular phoneme distinction as a function of the preceding phonetic context). Elman and McClelland (1988) apparently induced such compensation from lexical information; the preceding phonetic context supplied in their experiment was in fact a constant token ambiguous between [s] and *[J],* but it occurred at the end of *Christma** versus *fooli**. Listeners' responses to the phoneme following this constant token were shifted in the same direction as was found with the truly different phonemes at the end of *Christmas* and *foolish.* Elman and McClelland simulated their result in TRACE (a connectionist model of spoken-word recognition, see section 5.4.1) and attributed it to TRACE'S feedback connections between the lexical and the phoneme level. Norris (1993), however, simulated the same experimental findings in a network with no feedback connections. Subsequent studies then showed that the contextual dependence of compensation for coarticulation apparently reflects listeners' knowledge of transitional probabilities (Pitt and McQueen, 1998). Thus, both empirical and theoretical arguments disproved Elman and McClelland's original claim.

Norris *et al.* (1998) have argued, furthermore, that top-down feedback from the lexical level to prelexical processing stages cannot even improve recognition performance. After all, the best word-recognition performance is achieved by selection of the best lexical match(es) to whatever prelexical representation has been computed. Adding feedback from the lexical level to the prelexical level does not improve the lexical level's performance, but merely confirms it. Indeed, simulations with TRACE have shown that the overall accuracy of the model is neither better nor worse if the top-down connections which the model normally contains are removed (Frauenfelder and Peeters, in press).

This is not to deny that top-down information flow can result in alteration of prelexical decisions. For instance, if the output of prelexical processing is the string of phonetic representations /s*i/ in which the * represents some unclearly perceived stop

consonant, top-down activation from the lexicon (which contains the word *ski*, but neither *spee* or *stee)* might change the prelexical decision from uncertainty, to a certain decision that there had been a [k]. But if in fact there had not been a [k], because the speaker had actually made a slip of the tongue and said *spee* or *stee*, then the top-down information flow would, strictly speaking, have led to poorer performance by the prelexical processor, since it would have caused a wrong decision to be made about the phonetic structure of the input.

Thus top-down connections can clear up ambiguity in prelexical processing, but they do so at a potential cost; and more importantly, they do not result in an improvement of word recognition accuracy. There seems no need to build such connections into the blueprint of the listener.

5.7.2 Word recognition and utterance context

While the listener's knowledge of his or her lexicon may not directly feed into perceptual decisions of what segments are being heard, top-down influences may play a bigger role at higher levels of processing. The substantive context in which an ambiguous word, such as *bank* or *bug*, is heard clearly influences its interpretation. You do not think that a police agent is talking about insects if you hear him talking about putting a bug in a suspect's room. The interpretation of this observation, however, has shifted over the years. For a while, it was popular to suggest that a listener's context-based expectations played essentially the same role in word recognition as did perception of the physical signal (see Riesbeck and Schank 1978, for a particularly extreme statement of this position, extended to all of language comprehension). Experimental work reviewed earlier in this chapter (e.g. Swinney 1979) led to the opposite conclusion, that words were recognized (at least in the sense of the mental representations of all their senses being activated) regardless of context. Context was left the role of selecting from among the activated alternatives.

More recent work, also reviewed earlier, suggests that a strong enough context can effectively eliminate measurable activation of inappropriate word senses. Still, current theoretical opinion is sharply divided about the direction of information flow between the word recognition system and utterance-level processing. Some word recognition models (e.g. TRACE, McClelland and Elman 1986) assume that utterance context can activate mental representations of words directly, implying a top-down flow of information from higher-level processing to lexical processing. (Note, however, that the 1986 implementation of TRACE does not actually incorporate levels of processing above word recognition.) Other models (e.g. the Cohort model, Marslen-Wilson 1990. or Shortlist, Norris 1994) propose that the activation of words is immune from higher-level influence (although again, these models have not been implemented with utterance-level processing). In these models, as described in section 5.4.1, activation is automatic and may be initiated by partial information about a word; activated candidates are checked against current contextual representations, and early and powerful effects of context reflect the rapidity with which this check can lead to inhibition of inappropriate candidates.

No empirical data are as yet available to decide this issue. In line with the conclusion of section 5.7.1, it might therefore seem unnecessary at this point to build top-down information flow into the blueprint of the listener's word recognition system.

5.7.3 Syntactic and semantic processing

A similar theoretical contrast exists concerning how meaning and plausibility might influence the extraction of a message from a sentence. Here, though, as noted by Boland and Cutler (1996), some theoretical positions in which context and plausibility select from among several alternative structural analyses are termed 'interactive', while theories of word recognition that make a similar claim were termed 'autonomous'. This is partly because of Frazier's (1979, 1987) garden-path theory of parsing, which claims that a single structural analysis of a sentence is constructed on the basis of speed and economy, and later evaluated against context. In this context, a theory in which multiple candidates are allowed to compete with one another is interactive.

In Frazier's original theory, only a very limited amount of grammatical information was assumed to be used in constructing the initial analysis of a sentence. Recent work has expanded the range of grammatical information that seems to play an immediate role in initial sentence processing, most notably to include prosodic information. Prosody may be used to create a full prosodic representation of a sentence, developed in parallel with a mutually-constraining syntactic representation (cf. Frazier 1990, for an architecture that would permit this), or it might be viewed as another informational constraint in a constraint-satisfaction model such as that of MacDonald *et al.* (1994).

The question of the relation between syntactic and higher-level processing, however, still occasions much debate. Perhaps new advances will shortly be made here with new techniques. Electrophysiological studies of listening to spoken sentences, for instance, show clearly separable effects of violations of grammaticality and violations of semantic structure (Friederici 1998; Hagoort and Brown, in press). This suggests at least that comprehension models should incorporate appropriate distinctions between syntactic and semantic processing.

Currently, however, models are concerned to account for the many research results of the past decade showing that semantic context and plausibility, but also frequency of usage, are taken into account very quickly during parsing. Some models focus on how these factors may guide decisions among alternative syntactic structures. Tanenhaus *et al.* (in press) explicitly present their model as a model of such decisions, acknowledging that other theories must be devised to explain where the structures being decided among come from. MacDonald *et al.* (1994) suggest that the structures are simply projected from the lexical heads of phrases, a suggestion that has been criticized as inadequate by Frazier (1995). Other models (e.g. Frazier and Clifton 1996) focus more on the process by which structural analyses are initially constructed and less on how eventual selections are made. A compromise model was proposed by Boland (1997), involving constraint-based selection in combination with parallel autonomous generation of alternative structures. However, no completely satisfactory theory of how syntactic and extra-syntactic information are co-ordinated in

comprehending language is in our opinion as yet available. As Boland and Cutler (1996) concluded, the field has moved beyond a simplistic modular/ interactive contrast, but the more refined models which are now needed have not as yet been formulated and tested. The coming few years should prove exciting and productive for researchers involved in investigating spoken-language comprehension.

Acknowledgements

We thank Brechtje Post for the ToBI transcription in Fig. 5.2, and James McQueen for helpful comments on the text.

References

Abney, S. (1989). A computational model of human parsing. *Journal of Psycholinguistic Research,* 18, 129-44.

Allbritton, D., McKoon, G., and Ratcliff, R. (1996). The reliability of prosodic cues for resolving syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 22, 714-35.

Baayen, R. H., Dijkstra, T., and Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language,* 37, 94-117.

Bader, M. (1994). *The assignment of sentence accent during reading.* Paper presented at the CUNY Sentence Processing Conference, March, 1994, New York City.

Bard, E. G., Shillcock, R. C, and Altmann, G. T. M. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception and Psychophysics,* 44, 395^08.

Beckman, M. (1996). The parsing of prosody. *Language and Cognitive Processes,* 11, 17-67.

Beckman, M. E. and Ayers, G. M. (1993). *Guidelines for ToBI labelling, version 2.0.* Ohio State University.

Beckman, M. E. and Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology,* 3, 255-309.

Bever, T. G. (1970). The cognitive basis for linguistic structures. In *Cognition and the development of language* (ed. J. R. Hayes), pp. 279-352. Wiley, New York.

Birch, S. and Clifton, C, Jr (1995). Focus, accent, and argument structure. *Language and Speech,* 33, 365-91.

Birch, S. and Garnsey, S. M. (1995). The effect of focus on memory for words in sentences. *Journal of Memory and Language,* 34, 232-67.

Bock, J. K. and Mazzella, J. R. (1983). Intonational marking of given and new information: Some consequences for comprehension. *Memory and Cognition,* 11, 64—76.

Boland, J. E. (1997). The relationship between syntactic and semantic processes in sentence comprehension. *Language and Cognitive Processes,* 12, 423-84.

Boland, J. E. and Cutler, A. (1996). Interaction with autonomy: Multiple output models and the inadequacy of the Great Divide. *Cognition,* 58, 309-20.

Bolinger, D. (1978). Intonation across languages. In *Universals of human language, Vol 2: Phonology* (ed. J. J. Greenberg,), pp. 471-524. Stanford University Press.

Bond, Z. S. (1981). Listening to elliptic speech: Pay attention to stressed vowels. *Journal of Phonetics.* 9, 89-96.

Bond, Z. S. and Small, L. H. (1983). Voicing, vowel and stress mispronunciations in continuous speech. *Perception and Psychophysics,* 34, 470-74.

Bradley, D. C, Sanchez-Casas, R. M., and Garcia-Albea, J. E. (1993). The status of the syllable in the perception of Spanish and English. *Language and Cognitive Processes,* 8, 197-233.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound.* MIT Press, Cambridge, MA.

Bresnan, J. (1982). *The mental representation of grammatical relations.* MIT Press, Cambridge, MA.

Brown, G. D. A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavior Research Methods, Instrumentation and Computers,* 16, 502-32.

Caramazza, A., Laudanna, A., and Romani, C. (1988). Lexical access and inflectional morphology. *Cognition,* 28, 297-332.

Carroll, P. J. and Slowiaczek, M. L. (1987). Modes and modules: Multiple pathways to the language processor. In *Modularity in sentence comprehension: Knowledge representation and natural language understanding* (ed. J. L. Garfield), pp. 221-48. MIT Press. Cambridge. MA.

Chafe, W. L. (1974). Language and consciousness. *Language,* 50, 111-33.

Chen, H.-C. and Cutler, A. (1997). Auditory priming in spoken and printed word recognition. In *The cognitive processing of Chinese and related Asian languages* (ed. H.-C. Chen), pp. 77-81. Chinese University Press, Hong Kong.

Cluff, M. S. and Luce, P. A. (1990). Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. *Journal of Experimental Psychology: Human Perception and Performance,* 16, 551-63.

Connine, C. M., Titone, D., Deelman, T., and Blasko, D. (1997). Similarity mapping in spoken word recognition. *Journal of Memory and Language,* 37, 463-80.

Conrad, C. (1974). Context effects in sentence comprehension: A study of the subjective lexicon. *Memory and Cognition,* 2, 130-8.

Cooper, W. and Paccia-Cooper, J. (1980). *Syntax and speech.* Harvard University Press, Cambridge, MA.

Cooper, W. and Sorenson, J. (1981). *Fundamental frequency in speech production.* Springer. New York.

Cowan, N. (1984). On short and long auditory stores. *Psychological Bulletin,* 96. 341-70.

Cutler, A. (1986). *Forbear* is a homophone: Lexical prosody does not constrain lexical access. *Language and Speech,* 29, 201-20.

Cutler, A. and Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language,* 31, 218-36.

Cutler, A. and Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language,* 2. 133^t2.

Cutler, A. and Chen, H.-C. (1995). Phonological similarity effects in Cantonese word recognition. *Proceedings of the Thirteenth International Congress of Phonetic Sciences, Stockholm,* 1, 106-9.

Cutler, A. and Chen, H.-C. (1997). Lexical tone in Cantonese spoken-word processing. *Perception and Psychophysics,* 59, 165-79.

Cutler, A. and Fodor, J. A. (1979). Semantic focus and sentence comprehension. *Cognition.* 7. 49-59.

Cutler, A. and Foss, D. J. (1977). On the role of sentence stress in sentence processing. *Language and Speech,* 20, 1-10.

Cutler, A. and Norris. D. G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Hitman Perception and Performance,* 14, 113-21.

Cutler, A. and Otake, T. (1994). Mora or phoneme? Further evidence for language-specific listening. *Journal of Memory and Language,* 33, 824-44.

Cutler, A. and Otake, T. (1999). Pitch accent in spoken-word recognition in Japanese. *Journal of the Acoustical Society of America.*

Cutler, A., Mehler, J., Norris, D. G., and Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language,* 25, 385-400.

Cutler, A., Dahan, D., and Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech,* 40, 141-201.

Donselaar, W. van, Koster, M., and Cutler, A. *Voornaam* is not a homophone: Lexical prosody and lexical access in Dutch. (Manuscript.)

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C, and Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research,* 24, 409-36.

Elman, J. L. and McClelland, J. L. (1986). Exploiting lawful variability in the speech wave. In *Invariance and variability in speech processes* (eds J. S. Perkell and D. H. Klatt), pp. 360-86. Erlbaum, NJ.

Elman, J. L. and McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language,* 27, 143-65.

Fear, B. D., Cutler, A., and Butterfield, S. (1995). The strong/weak syllable distinction in English. *Journal of the Acoustical Society of America,* 97, 1893-904.

Ferreira, F., Anes, M. D., and Horine, M. D. (1996). Exploring the use of prosody during language comprehension using the auditory moving window technique. *Journal of Psycholinguistic Research,* 25, 273-90.

Foss, D. J. and Gernsbacher, M. A. (1983). Cracking the dual code: Toward a unitary model of phonetic identification. *Journal of Verbal Learning and Verbal Behavior,* 22, 609-32.

Fox, R. A. and Unkefer, J. (1985). The effect of lexical status on the perception of tone. *Journal of Chinese Linguistics,* 13, 69-90.

Frauenfelder, U. H. (1991). Lexical alignment and activation in spoken word recognition. In *Music, language, speech and brain* (eds J. Sundberg, L. Nord , and R. Carlson), pp. 294-303. Macmillan, London.

Frauenfelder, U. H. and Peeters, G. Simulating the time-course of spoken word recognition: An analysis of lexical competition in TRACE. In *Symbolic connectionism* (eds J. Grainger and A. M. Jacobs). Erlbaum, NJ. (In press.)

Frazier, L. (1979). *On comprehending sentences: Syntactic parsing strategies.* **Indiana** University Linguistics Club, Bloomington.

Frazier, L. (1987). Sentence processing: A tutorial review. In *Attention and performance,* Vol. 12 (ed. M. Coltheart), pp. 559-86. Erlbaum, NJ.

Frazier, L. (1989). Against lexical generation of syntax. In *Lexical representation and process* (ed. W.D. Marslen-Wilson), pp. 505-28. MIT Press, Cambridge, MA.

Frazier, L. (1990). Exploring the architecture of the language system. In *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (ed. G. T. M. Altmann), pp. 409-33. MIT Press, Cambridge, MA.

Frazier, L. (1995). Constraint satisfaction as a theory of sentence processing. *Journal of Psycholinguistic Research*, 24, 437-68.

Frazier, L. and Clifton, C, Jr (1996). *Construal.* MIT Press, Cambridge, MA.

Frazier, L. and Clifton, C, Jr. (1998). Sentence reanalysis, and visibility. In *Sentence Reanalysis* (eds J. D. Fodor and F. Ferreira), pp. 143-76. Kluwer, Dordrecht.

Frazier, L. and Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14, 178-210.

Friederici, A. D. (1998). The neurobiology of language processing. In *Language comprehension: A biological perspective* (ed. A. D. Friederici), pp. 263-301. Springer, Heidelberg.

Fujimura, O. and Lovins, J. B. (1978). Syllables as concatenative phonetic units. In *Syllables and segments* (eds A. Bell and J. B. Hooper), pp. 107-20. North-Holland. Amsterdam.

Gaskell, M. G. and Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, 12, 613-56.

Glanzer, M. (1976). Intonation grouping and related words in free recall. *Journal of Verbal Learning and Verbal Behavior*, **15,** 85-92.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, **105,** 251-79.

Goldinger, S. D., Luce, P. A., and Pisoni, D. B. (1989). Priming lexical neighbours of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, 28, 501-18.

Goldinger, S. D., Luce, P. A., Pisoni, D. B., and Marcario. J. K. (1992). Form-based priming in spoken word recognition: The roles of competition and bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 1211-38.

Gow, D. W. and Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance*, 21, 344-59.

Greenspan, S. L. (1986). Semantic flexibility and referential specificity of concrete nouns. *Journal of Memory and Language*, 25. 539-57.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception and Psychophysics*, 38, 299-310.

Grosjean, F. and Gee, J. (1987). Prosodic structure and spoken word recognition. *Cognition*, 25, 135-55.

Hagoort, P. and Brown, C. M. Semantic and syntactic effects of listening to speech compared to reading. *Neuropsychologic!.* (In press.)

Halliday, M. A. K. (1967). *Intonation and grammar in British English.* Mouton. The Hague.

Howes, D. (1966). A word count of spoken English. *Journal of Verbal Learning and Verbal Behavior*, 5, 572-604.

Jennings, F.. Randall, B,, and Tyler. L. K. (1997). Graded effects of verb subcategory preferences on parsing: Support for constraint-satisfaction models. *Language and Cognitive Processes*, 12, 485-504.

Kjelgaard, M. M. and Speer, S. R. Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language.* (In press).

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics,* 7, 279-312.

Klatt, D. H. (1989). Review of selected models of speech perception. In *Lexical representation and process* (ed. W. D. Marslen-Wilson), pp. 169-226. MIT Press, Cambridge, MA.

Kolinsky, R. (1992). Conjunction errors as a tool for the study of perceptual processing. In *Analytic approaches to human cognition* (eds J. Alegria, D. Holender, J. Morais, and M. Radeau), pp. 133^9. North Holland, Amsterdam.

Kong, Q. M. (1987). Influence of tones upon vowel duration in Cantonese. *Language and Speech,* 30, 387-99.

Konieczny, L., Hemforth, B., Scheepers, C, and Strube, G. (1997). The role of lexical heads in parsing: Evidence from German. *Language and Cognitive Processes,* 12, 307-48.

Kratochvil, P. (1971). An experiment in the perception of Peking dialect tones. In *A symposium on Chinese grammar* (ed. I.-L. Hansson). Curzon Press, Lund.

Lackner, J. R. and Garrett, M. F. (1972). Resolving ambiguity: Effects of biasing context in the unattended ear. *Cognition,* 1, 359-72.

Ladd, D. R. (1996). *Intonational phonology.* Cambridge University Press.

Lahiri, A. and Marslen-Wilson, W. D. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition,* 38, 245-94.

Lehiste, I. (1973). Phonetic disambiguation of syntactic ambiguity. *Glossa,* 7, 107-22.

Lehiste, I., Olive, J., and Streeter, L. (1976). Role of duration in disambiguating syntactically ambiguous sentences. *Journal of the Acoustical Society of America,* 60, 1199-202.

Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition,* 21, 1-36.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review,* 74, 431-61.

Lucas, M. M. (1987). Frequency effects on the processing of ambiguous words in sentence contexts. *Language and Speech,* 30, 25-46.

Luce, P. A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception and Psychophysics,* 39, 155-8.

Luce, P. A., Pisoni, D. B., and Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In *Cognitive models of speech processing* (ed. G. T. M. Altmann), pp. 122-147. MIT Press. Cambridge, MA.

MacDonald. M. C, Pearlmutter, N. J., and Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review,* 101, 676-703.

Maddieson, I. (1984). *Patterns of sounds.* Cambridge University Press.

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

Marcus, S. M. and Frauenfelder, U. H. (1985). Word recognition: Uniqueness or deviation? A theoretical note. *Language and Cognitive Processes,* 1, 163-9.

Marcus, M. and Hindle, D. (1990). Description theory and intonation boundaries. In *Cognitive models of speech processing* (ed. G. T. M. Altmann), pp. 483-512. MIT Press, Cambridge, MA.

Marslen-Wilson, W. D. (1990). Activation, competition and frequency in lexical access. In *Cognitive models of speech processing* (ed. G. T. M. Altmann), pp. 148-72. MIT Press, Cambridge, MA.

Marslen-Wilson, W. D. and Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes and features. *Psychological Review,* 101, 653-75.

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

Marslen-Wilson, W. D., Brown, C. M., and Tyler, L. K. (1988). Lexical representations in spoken language comprehension. *Language and Cognitive Processes.* 3. 1-16.

Marslen-Wilson, W. D., Tyler, L. K., Warren, P.. Grenier, P., and Lee, C. S. (1992). Prosodic effects in minimal attachment. *Quarterly Journal of Experimental Psychology,* **45A,** 730-87.

Marslen-Wilson, W. D., Tyler, L. K., Waksler. R„ and Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review,* 101, 3-33.

Martin, J. G. and Bunnell, H. T. (1981). Perception of anticipatory coarticulation effects. *Journal of the Acoustical Society of America.* 69, 559-67.

Martin, J. G. and Bunnell, H. T. (1982). Perception of anticipatory coarticulation effects in vowel-stop consonant-vowel sequences. *Journal of Experimental Psychology: Hitman Perception and Performance,* 8, 473-88.

McClelland, J. L. and Elman. J. L. (1986). TheTRACE model of speech perception. *Cognitive Psychology,* 18, 1-86.

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language,* 39, 21-46.

McQueen, J. M. and Cutler, A. (1992). Words within words: Lexical statistics and lexical access. *Proceedings of the Second International Conference on Spoken Language Processing. Banff, Canada,* 1, 221-4.

McQueen, J. M. and Cutler, A. (1998). Morphology in word recognition. In *The handbook of morphology* (eds A. Spencer and A. M. Zwicky). pp. 406-27. Blackwell, Oxford.

McQueen, J. M., Norris. D. G.. and Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning. Memory, and Cognition,* 20, 621-38.

McQueen, J. M., Norris. D. G., and Cutler. A. Lexical influence in phonetic decision-making: Evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance.* (In press.)

Mehler, J.. Dommergues, J.-Y., Frauenfelder. U. H„ and Segui. J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior,* 20. 298-305.

Miller, J. L. (1981). Some effects of speaking rate on phonetic perception. *Phonctica.* 38. 159-80.

Miller, J. L. and Liberman, A. L. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception and Psychophysics,* 25. 457-65.

Nespor, M. and Vogel, I. (1986). *Prosodic phonology.* Foris, Dordrecht.

Nooteboom. S. G. and Terken, J. M. B. (1982). What makes speakers omit pitch accents: An experiment. *Phonetica,* 39. 317-36.

Norris, D. G. (1993). Bottom-up connectionist models of'interaction'. In *Cognitive models of speech processing* (eds G. T. M. Altmann and R. Shillcock), pp. 211-34. Erlbaum, NJ.

Norris, D. G. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition,* 52, 189-234.

Norris, D. G., McQueen, J. M., and Cutler, A. (1995). Competition and segmentation in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 21, 1209-28.

Norris, D. G.. McQueen, J. M., Cutler, A., and Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology,* 34, 191-243.

Norris, D. G., McQueen, J. M., and Cutler, A. (1998). Merging phonetic and lexical information in phonetic decision-making. (Manuscript.)

Oden, G. C. and Spira, J. L. (1983). Influence of context on the activation and selection of ambiguous word senses. *Quarterly Journal of Experimental Psychology,* 35, 51-64.

Onifer, W. and Swinney, D. A. (1981). Accessing lexical ambiguities during sentence comprehension: Effects of frequency-of-meaning and contextual bias. *Journal of Verbal Learning and Verbal Behavior,* 17, 225-36.

Orsolini, M. and Marslen-Wilson, W. D. (1997). Universals in morphological representation: Evidence from Italian. *Language and Cognitive Processes,* 12, 1—47.

Otake, T., Hatano, G., Cutler, A., and Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language,* 32, 358-78.

Otake, T., Hatano, G., and Yoneyama, K. (1996a). Speech segmentation by Japanese listeners. In *Phonological structure and language processing: Cross-linguistic studies* (eds T. Otake and A. Cutler), pp. 183-201. Mouton de Gruyter, Berlin.

Otake, T., Yoneyama, K.„ Cutler, A., and Van der Lugt, A. (1996ft). The representation of Japanese moraic nasals. *Journal of the Acoustical Society of America,* 100, 3831^42.

Peretz, I., Lussier, I., and Beland, R. (1996). The roles of phonological and orthographic code in word stem completion. In *Phonological structure and language processing: Cross-linguistic studies* (eds T. Otake and A. Cutler), pp. 217-26. Mouton de Gruyter, Berlin.

Pierrehumbert, J. B. (1980). The phonology and phonetics of English intonation. Unpublished Ph.D. thesis, MIT.

Pisoni, D. B. and Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition,* 25, 21-52.

Pitt, M. A. and McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language,* 39, 347-70.

Pollard, C. and Sag, I. A. (1994). *Head-driven phrase structure grammar.* CSLI, Stanford.

Price, P. J., Ostendorf, M„ Shattuck-Huffnagel, S., and Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America,* 90, 2956-70.

Pustejovsky, J. (1995). *The generative lexicon.* MIT Press, Cambridge, MA.

Pynte, J. and Prieur, B. (1996). Prosodic breaks and attachment decisions in sentence processing. *Language and Cognitive Processes,* 11, 165-92.

Rayner, K. and Pollatsek. A. (1989). *The psychology of reading.* Prentice-Hall, Englewood Cliffs, NJ.

Rayner, K. and Pollatsek, A. (1996). Reading unspaced text is not easy: Comments on the implications of Epelboim *et* a/.'s (1994) study for models of eye movement control in reading. *Vision Research,* 36, 461-70.

Riesbeck, C. K. and Schank, R. (1978). Comprehension by computer: Expectation-based analysis of sentences in context. In *Studies in the perception of language* (eds W. J. M. Levelt and G. B. Floris d" Arcais), pp. 247-94. Wiley, New York.

Ryan, J. (1969). Grouping and short-term memory: Different means and patterns of grouping. *Quarterly Journal of Experimental Psychology.* 21, 137-47.

Samuel, A. G. (1989). Insights from a failure of selective adaptation: Syllable-initial and syllable-final consonants are different. *Perception and Psychophysics,* 45, 485-93.

Schafer, A. (1996). Prosodic parsing: The role of prosody in sentence comprehension. Unpublished Ph.D. thesis. University of Massachusetts, Amherst.

Schafer, A., Carter, J., Clifton, C, Jr., and Frazier. L. (1996). Focus in relative clause construal. *Language and Cognitive Processes,* 11, 135-63.

Schriefers, H., Zwitserlood, P., and Roelofs. A. (1991). The identification of morphologically complex spoken words: Continuous processing or decomposition? *Journal of Memory ami Language,* 30, 26-47.

Sebastian-Galles, N., Dupoux, E.. Segui, J., and Mehler, J. (1992). Contrasting syllabic effects in Catalan and Spanish. *Journal of Memory and Language.* 31, 18-32.

Sedivy, J., Tanenhaus, M., Spivey-Knowlton. M.. Eberhard. K., and Carlson. G. (1995). Using intonationally-marked presuppositional information in on-line language processing: Evidence from eye movements to a visual model. *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society,* 375-80. Erlbaum. Hillsdale. NJ.

Segui, J. (1984). The syllable: A basic perceptual unit in speech processing? In *Attention and performance X: Control of language processes* (eds H. Bouma and D. G. Bouwhuis). pp. 165-82 . Erlbaum, NJ.

Segui, J., Frauenfelder, U. H., and Mehler, J. (1981). Phoneme monitoring, syllable monitoring and lexical access. *British Journal of Psychology,* 72, 471-77.

Seidenberg, M. S., Tanenhaus. M. K.. Leiman. J. M.. and Bienkowski. M. (1982). Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology.* 14. 489-537.

Selkirk, E. O. (1984). *Phonology and syntax: The relation between sound and structure.* MIT Press, Cambridge, MA.

Selkirk, E. O. (1995). Sentence prosody: Intonation, stress, and phasing. In *Handbook of phonological theory* (ed. J. Goldsmith), pp. 550-69. Blackwell. Oxford.

Shen, X. S. and Lin, M. (1991). A perceptual study of Mandarin tones 2 and 3. *Language and Speech,* 34, 145-56.

Shattuck-Hufnagel, S. and Turk. A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research.* 25. 193-248.

Shillcock, R. C. (1990). Lexical hypotheses in continuous speech. In *Cognitive models of speech processing* (ed. G. T. M. Altmann), pp. 24-49. MIT Press, Cambridge. MA.

Silverman. K. E. A., Blaauw, E., Spitz, J., and Pitrelli. J. F. (1992). Towards using prosody in speech recognition understanding systems: Differences between read and spontaneous

speech. Paper presented at the Fifth DARPA Workshop on Speech and Natural Language, Harriman, NY.

Slowiaczek, L. M. (1990). Effects of lexical stress in auditory word recognition. *Language and Speech,* 33, 47-68.

Slowiaczek, M. L. (1981). Prosodic units as language processing units. Unpublished Ph.D. thesis. University of Massachusetts, Amherst.

Slowiaczek, M. L. and Clifton, C, Jr (1980). Subvocalization and reading for meaning. *Journal of Verbal Learning and Verbal Behavior,* 19, 573-82.

Small. L. H., Simon, S. D., and Goldberg, J. (1988). Lexical stress and lexical access: Homographs versus nonhomographs. *Perception and Psychophysics,* 44, 272-80.

Solan, L. (1980). Contrastive stress and children's interpretation of pronouns. *Journal of Speech and Hearing Research,* 23, 688-98.

Speer, S. R. and Kjelgaard, M. M. (1998). Prosodic facilitation and interference in the resolution of temporary syntactic ambiguity. (Manuscript.)

Speer, S. R., Crowder, R. G., and Thomas, L. M. (1993). Prosodic structure and sentence recognition. *Journal of Memory and Language,* 32, 336-58.

Speer, S. R., Kjelgaard, M. M., and Dobroth, K. M. (1996). The influence of prosodic structure on the resolution of temporary syntactic closure ambiguities. *Journal of Psycholinguistic Research,* 25, 249-72.

Spivey-Knowlton, M. and Sedivy, J. C. (1995). Resolving attachment ambiguities with multiple constraints. *Cognition,* 55, 227-67.

Streeter, L. A. and Nigro, G. N. (1979). The role of medial consonant transitions in word perception. *Journal of the Acoustical Society of America,* 65, 1533-41.

Suomi, K., McQueen, J. M., and Cutler, A. (1997). Vowel harmony and speech segmentation in Finnish. *Journal of Memory and Language,* **36,** 422^4.

Swinney, D. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior,* 18, 645-59.

Tabossi. P. (1988a). Accessing lexical ambiguity in different types of sentential contexts. *Journal of Memory and Language.* 27, 324-40.

Tabossi. P. (19886). Effects of context on the immediate interpretation of unambiguous nouns. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 14, 153-62.

Tabossi. P., Colombo, L., and Job, R. (1987). Accessing lexical ambiguity: Effects of context and dominance. *Psychological Research,* 49, 161-7.

Taft, L. (1984). Prosodic constraints and lexical parsing strategies. Unpublished Ph.D. thesis, University of Massachusetts.

Taft, M. (1986). Lexical access codes in visual and auditory word recognition. *Language and Cognitive Processes.* 1. 297-308.

Taft, M. and Chen, H.-C. (1992). Judging homophony in Chinese: The influence of tones. In *Language processing in Chinese* (eds H.-C. Chen and O. J. L. Tzeng), pp. 151-172. Elsevier, Amsterdam.

Tanenhaus. M. K. and Donenwerth-Nolan. S. (1984). Syntactic context and lexical access. *Quarterly Journal of Experimental Psychology,* **36A,** 649-61.

Tanenhaus, M. K., Leiman, J. M., and Seidenberg, M. S. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior.,* 18, 427-40.

Tanenhaus, M. K., Garnsey, S„ and Boland, J. (1990). Combinatory lexical information and language comprehension. In *Cognitive models of speech processing* (ed. G. T. M. Altmann), pp. 383-408. MIT Press, Cambridge, MA.

Tanenhaus, M. K., Boland, J. E., Mauner, G., and Carlson, G. N. (1993). More on combinatory lexical information: Thematic structure in parsing and interpretation. In *Cognitive models of speech processing* (eds G. T. M. Altmann and R. Shillcock). pp. 297-319. Erlbaum, NJ.

Tanenhaus, M. K., Spivey-Knowlton, M. J., and Hanna, J. E. Modeling thematic and discourse context effects within a multiple constraints framework: Implications for the architecture of the language comprehension system. In *Architectures and mechanisms for language processing* (eds M. Crocker, M. Pickering, and C. Clifton). Cambridge University Press. (In press.)

Tanehaus, M. K. and Donenwerth-Nolan, S. (1984). Syntactic context and lexical access. *Quarterly Journal of Experimental Psychology.* **36A,** 649-61.

Terken, J. and Nooteboom, S. G. (1987). Opposite effects of accentuation and deaccentuation on verification latencies for Given and New information. *Language and Cognitive Processes, 2,* 145-64.

Trueswell, J. C. and Tanenhaus, M. K. (1991). Tense, temporal context and syntactic ambiguity resolution. *Language and Cognitive Processes.* 6. 303-38.

Tsang, K. K. and Hoosain, R. (1979). Segmental phonemes and tonal phonemes in comprehension of Cantonese. *Psychologia,* 22, 222^1.

Tyler, L. K. (1989). The role of lexical representation in language comprehension. In *Lexical representation and process* (ed. W. D. Marslen-Wilson), pp. 439-62. MIT Press, Cambridge, MA.

Vroomen, J. and de Gelder, B. (1995). Metrical segmentation and lexical inhibition in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance,* 21, 98-108.

Vroomen, J. and de Gelder, B. (1997). Activation of embedded words in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance.* 23. 710-20.

Wales, R. and Toner, H. (1979). Intonation and ambiguity. In *Sentence processing: Psycholingaistic studies presented to Merrill Garrett* (eds W. E. Cooper and E. C. T. Walker), pp. 135-158. Erlbaum. NJ.

Wallace, W. P., Stewart, M. T., Sherman. H. L.. and Mellor. M. D. (1995). False positives in recognition memory produced by cohort activation. *Cognition.* 55. 85-113.

Watt, S. M. and Murray, W. S. (1996). Prosodic form and parsing commitments. *Journal of Psycholinguists Research,* 25, 291-318.

Whalen, D. H. (1984). Subcategorical mismatches slow phonetic judgments. *Perception and Psychophysics,* 35, 49-64.

Whalen, D. H. (1991). Subcategorical phonetic mismatches and lexical access. *Perception and Psychophysics,* 50. 351-60.

Whitney, P., McKay, T., Kellas, G., and Emerson, W. A. (1985). Semantic activation of noun concepts in context. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 11, 126-35.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition,* 32, 25-64.

Zwitserlood, P., Schriefers, H., Lahiri, A., and Donselaar, W. van (1993). The role of syllables in the perception of spoken Dutch. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 19, 260-71.