2. Segmentation in lexical access

The literature on spoken word recognition frequently states that connected speech contains no acoustic analogue of the white spaces between words on the printed page (for recent examples, see Christiansen, Allen, & Seidenberg, 1998; McQueen, Cutler, Briscoe, & Norris, 1995). If words were written on the printed page as they sound the result would be something like (1) below:

(1) wordsinspeechwouldruntogetherlikethis

This visual caricature has been used to motivate investigation into lexical segmentation, since the difficulties that we experience in reading sentences like (1) appear not to be present when we listen to connected speech[1].

This chapter begins by reviewing the acoustic-phonetics literature to determine whether, as has been argued, there are no markers of word boundaries in connected speech. However, since work in phonetics has not focused on the mechanisms by which potential boundary cues can be processed during recognition, this chapter will be primarily concerned with reviewing the psycholinguistic literature on segmentation with reference to what is known about the acoustic properties of the speech stream.

In this review of the psycholinguistic literature on segmentation an important difference is between accounts in which knowledge of the statistical structure of lexical items is applied to segmentation, as opposed to accounts in which segmentation occurs through the identification of specific lexical items. One important issue in assessing different mechanisms by which lexical knowledge contributes to segmentation is the extent to which word boundaries may be ambiguous due to the presence of words embedded at the onset of longer words. This chapter therefore concludes with database searches evaluating

[1] Note that some languages such as Thai have orthographies which exclude spaces between words. Interestingly, (unlike in English) the presence or absence of spaces has no significant effect on the reading speed of Thai speakers (Kohsom and Gobet, 1997).

whether different assumptions regarding the nature of lexical representations alter the amount of ambiguity created by onset-embedded words.

2.1. Acoustic cues to segmentation

Since connected speech contains relatively few invariant cues to the identities of individual segments it would be surprising if cues to word boundaries were unambiguously marked in the speech stream. However, investigation of the acoustic properties of the speech stream will be necessary to evaluate any claim about the lack of marked word boundaries in connected speech. In the acoustic-phonetics literature on word boundaries two main classes of cue have been described; qualitative changes in speech segments that are at either the onset or offset of a word, and changes in the duration of segments or syllables depending on their location with respect to a word boundary.

2.1.1. Segmental cues to word boundaries

Acoustic analyses of cues to word boundaries have focused on minimal pairs for which only the location of a word boundary distinguishes between two interpretations. Examples of these sequences include *play taught* and *plate ought* or *grey day* and *grade A*. These minimally contrastive items have a long history in the phonetics literature (Jones, 1931), though the earliest survey of the acoustic properties of these stimuli was carried out by Lehiste (1960). Lehiste recorded three different speakers reading a selection of these minimal pairs with different segments located either side of the word boundary. She then carried out spectrographic analysis of the resulting speech waves, attempting to relate measured acoustic differences between pairs of stimuli with the success or failure of listeners in identifying the sequences.

Lehiste reports that listeners were unanimous in transcribing over two thirds of these minimally different pairs – suggesting that they were able to identify the differences between these pairs. Nonetheless, Lehiste found no evidence that there was any signal in the speech stream that uniquely marks the boundary between words – i.e. there is no segment equivalent to the spaces between words in written languages. The cues that existed to the location of word boundaries were, in many cases, unique to the particular combination of pre- and post-boundary consonants. For instance, for words beginning with a stressed vowel, glottal stops and laryngeal voicing indicated that that vowel was the

onset of a new word. Other cues for the detection of onsets include the aspiration of word initial voiceless stops (for instance the onset segment /t/ will be aspirated in *play taught* but not in *plate ought*) and other allophonic variations in the production of /l/ and /r/. Another consistently observed acoustic difference was lengthening of the pre-juncture vowel; for example the vowel /eɪ/ is of greater duration in *grey day* than in *grade A*. These differences were described by Lehiste as being properties of words that are delimited by a word boundary, not markers of the boundary itself.

However Lehiste's methodology for analysing and evaluating these acoustic differences is not sufficiently thorough. Since each stimulus pair that she tested includes several potential acoustic cues, it may be unclear which of these cues listeners used to identify word boundaries. Consequently, further work has been carried out using more tightly controlled stimuli to determine which (if any) of the acoustic differences noted by Lehiste can be used individually to identify the location of a word boundary.

One study by Christie (1974) used synthetic speech to investigate whether aspiration of a voiceless stop is a cue to the presence of a word boundary. Comparing synthesised pairs such as /eɪstɑː/ and /eɪst^{h}ɑː/ (which may be perceived as *a star* or *ace tar*) Christie showed that aspiration alone is sufficient to signal to subjects that a voiceless stop is word initial. However, since aspiration can only be a cue for the segmentation of words that start with a voiceless stop this result falls short of providing a general solution to the problem of lexical segmentation.

A more general study by Nakatani and Dukes (1977) used cross-spliced speech from several different minimal pairs (such as *play taught* and *plate ought*) to determine where cues to juncture were located in these stimuli. They examined four possible loci for segmental cues for word juncture – in the onset of the sequences (e.g. /pleɪ/ from *play* and *plate*), the offset of the sequence (/ɔːt/ from *ought* and *taught*), and either in the initial or final portion of the juncture segment (/t/ in the example above). Using stimuli resynthesised from different combinations of sections from the two 'parent phrases', they compared subjects' interpretations of these stimuli to determine which sections of each stimulus contributed most strongly to the placement of word boundaries.

Nakatani and Dukes found no evidence that the onset or offset of these sequences had any effect on participants' placement of word boundaries. Since it was these sections of the stimuli that accounted for the variation in duration reported by Lehiste, they concluded that vowel duration did not appear to act as a cue to word juncture. The main location in the speech stream that influenced boundary perception in their study was the onset of the second word. This suggests that the qualitative differences in onset segments (such as the allophonic variation observed by Lehiste) carry most information for the placement of word boundaries. Nakatani and Dukes also confirmed the role of aspiration for voiceless stops and suggested that glottal vowels and laryngealizations also function as cues for stimuli with vowel onsets.

One exception, where boundary cues were present in segments other than at the onset of a word, was the variation observed for /l/ (as in *we loan* vs. *we'll own*) and /r/ (*two ran* vs. *tour an*) which differed both word finally and word initially. The two sequences using these segments were only segmented correctly where both the initial and final sections of the juncture segment could be heard – these pairs otherwise being susceptible to 'doubling' (responses such as *we'll loan* or *tour ran*) or 'disappearance' of the juncture consonant (responses such as *we own* or *two an*).

Nakatani and Dukes concluded that, /l/ and /r/ aside, qualitative changes in onset-segments influence the perception of word boundaries and that these changes are more valuable than quantitative changes in variables such as vowel duration. Taken at face value, this marking of word-onsets suggests that a more realistic visual caricature of the properties of the speech stream might be as shown in (2) below:

(2) WordsInSpeechWouldRunTogetherLikeThis

However, such a conclusion may exaggerate the reliability of acoustic cues to word boundaries on several grounds. Firstly, as in the Lehiste study, the stimuli were recorded from a list (albeit scrambled). Such recordings will be closer to the citation form of the target words than would be expected in more naturalistic speech, which may enhance the strength of boundary cues. Research by Barry (1981) showed that the qualitative cues identified by Lehiste are much more variable when these minimal pairs occur in passages read as connected speech. Secondly, it is unclear whether there are boundary cues for all possible juncture segments. Nakatani and Dukes reported that even for un-spliced stimuli,

identification rates for some stimuli were as low as 33%. Since chance performance would produce 25% correct responses, and ruling out sequences with 'doubled' and 'disappeared' segments would produce 50% correct performance, this level of performance indicates that although segmental cues support boundary detection, they may be too weak and unreliable to be used as a sole cue for the segmentation of fluent speech.

A further problem with this work is that (as Nakatani and Dukes themselves concede) the stimulus sequences were presented in isolation during testing. Such a presentation format may preclude the use of temporal cues to word boundaries since listeners will be unable to make use of the rhythmic properties of connected speech. This mode of stimulus presentation may therefore conceal an important source of information that would be accessible in fluent speech – namely differences in the duration of vowels in open and closed syllables (i.e. the difference between *play* and *plate*). This review now turns to acoustic-phonetic work which directly investigates whether segment duration can act as a cue to the location of word boundaries.

2.1.2. Duration cues to word boundaries

In order for duration to act as a cue for detecting word boundaries, it is first necessary to demonstrate that durations of segments change to reflect the location of word boundaries (as suggested initially by Lehiste, 1960). However, there is rather less consensus regarding variation in segment duration in the acoustic-phonetics literature than there is in the literature on segmental cues. An early contribution was made by Lehiste (1972) who showed that there are reliable differences in the articulation of the syllable [sli:p] in words such as *sleep, sleepy* and *sleepiness*. As the number of syllables in the word increases, the duration of the syllable (and its vowel nucleus) decreases. Such a cue, if sufficiently reliable, might allow listeners to distinguish between syllables that make a word and those that are at the start of a longer word.

This temporal compression of vowels in polysyllabic words was followed up by Klatt (1976) in a review article on the nature and use of segment duration in English. Klatt's model of vowel duration was based on a number of factors, each of which could shorten the vowel by a proportion of its full length, up to a pre-determined minimum length. This very simple model provided a good match to vowel duration data collected by Umeda (1975). Factors that were listed by Klatt included the voicing of the segment following the

vowel (unvoiced consonants are preceded by shorter vowels than voiced consonants), shortening of non-phrase final vowels and shortening of unstressed vowels. Klatt also reports that syllables in polysyllabic words are 15% shorter than the equivalent syllable in a monosyllable (confirming Lehiste's findings).

However, describing an acoustic difference does not mean that this cue is actually used by listeners in identifying boundaries in connected speech. With this issue in mind, Nakatani and Schaffer (1978) used reiterant speech to investigate whether syllable duration can be used to place word boundaries. Reiterant speech, as originally described by Liberman and Streeter (1978), is generated by speakers replacing each syllable of a target word with a repeated syllable such as /ma/ (for example, the phrase, *new result* would be produced as /ma mama/). Such speech has been shown to preserve natural prosodic variation in the intonation, amplitude and duration of syllables but to remove sources of variation caused by the phonemic properties of the segments that are replaced by the repeated syllable.

Nakatani and Schaffer showed that such stimuli preserve the expected duration differences between mono-and poly-syllabic words as described by Klatt (1976), as well as the expected differences between the durations of stressed and unstressed syllables. They also showed that word-initial consonants were lengthened in these stimuli, providing support for one of the acoustic cues described by Lehiste (1960).

Using forced-choice boundary placement tests with these stimuli, Nakatani and Schaffer showed that listeners could reliably segment reiterant speech into the words intended by the speaker. Subjects performed best for sequences with unambiguous syllable stress patterns such as StrongStrongWeak (a stress pattern in which the word boundary must be after the first syllable since a weak syllable can not be an entire word for these adjective-noun combinations). However, even for sequences where stress location does not determine word boundaries, subjects still performed significantly above chance. For example a sequence of syllables with the metrical pattern StrongWeakStrong could come from a pair like *noisy dog* or a pair like *bold design*. Nonetheless, listeners were able to determine whether these sequences should have a boundaries placed after the first strong syllable (as in *"bold design"*) or after the second weak syllable (as in *"noisy dog"*).

By carrying out further listening tests using these ambiguously stressed stimuli re-synthesised with and without rhythm, pitch, amplitude and spectral differences, Nakatani

and Schaffer demonstrated that temporal properties of these stimuli carried the most important cue to segmentation: listeners only segmented sequences at better than chance performance where duration differences between syllables were preserved. They therefore concluded that relative duration, particularly the lengthening of syllables in monosyllabic words, is an important cue to the location of a word boundary. Although these findings are suggestive, results obtained using these rather artificial stimuli can not necessarily be extended to studies using more realistic speech stimuli. Since the phonetic properties of the constituent segments of a syllable can also alter syllable duration, the use of duration differences as a cue to the location of a word boundary will be more difficult in naturally occurring speech.

These results also do not demonstrate that syllable duration is a reliable cue to the location of all word boundaries. If this were the case then it would be necessary for all syllables that precede a word boundary to be lengthened (not just where the word before the boundary is monosyllabic). Various investigations have been carried out to investigate whether such lengthening (as observed by Nakatani, O'Connor, & Aston (1981) for reiterant speech) is also to be found in connected speech. Crystal and House (1990) carried out measurements of vowel duration in a wide range of spoken materials. They observed no tendency towards pre-boundary lengthening; indeed, where the final segment before a boundary was a vowel they observed that this segment was shortened (directly contradicting prior work by Lehiste 1960 using minimal pairs such as *grey day* and *grade A*).

However the studies carried out by Crystal and House have been criticised for using stimuli that were too weakly controlled to provide reliable data. Anderson and Port (1994) carried out more careful investigations of segmental duration as a cue for boundary detection using stimuli based around a template that controlled for the duration differences caused by segments with different manners of articulation (stop, fricative, approximant, etc.). Measures of segment duration obtained in these more constrained environments were then entered into a discriminant analysis to determine the amount of information carried by the temporal properties of speech segments. This showed that segment and syllable durations differed markedly with the metrical stress of syllables, but only weakly with the location of word boundaries.

These statistical analyses suggest that variation in duration may not carry information that directly contributes to the placement of word boundaries in all lexical environments. Syllable lengthening in monosyllabic words has been frequently reported in the literature, but is not an example of a general phenomenon of pre-boundary lengthening. Therefore even if listeners are efficient users of these duration cues they would only be of value in distinguishing monosyllables from polysyllables. Nonetheless, as will be discussed later on in the chapter, important theoretical issues in psycholinguistic accounts of lexical access and segmentation have focused on the question of how listeners distinguish monosyllables from longer words in which they are embedded (e.g. distinguishing *cap* from *captain*). Consequently, acoustic differences between syllables in mono-syllabic and polysyllabic words may yet play an important role in lexical segmentation.

2.2. Psycholinguistic accounts of lexical segmentation

As has been seen in the preceding review, work in acoustic-phonetics paints a confusing and often conflicting picture of the properties of the speech stream. Although there are acoustic cues to word boundaries in the speech stream, it has not been demonstrated that these acoustic differences are sufficient to permit the detection of word boundaries in connected speech. Furthermore, it is unclear whether the detection of these cues to word boundaries can be achieved in naturally produced stimuli and with a time course appropriate for the perception of connected speech. For these reasons, the psycholinguistic literature on lexical segmentation has generally focused on more salient and more widely applicable cues to speech segmentation.

Different accounts of lexical segmentation in the psycholinguistic literature have described processes operating at several distinct levels of representation of the speech signal. These have previously been categorised as operating at either a pre-lexical or lexical level (Gow & Gordon, 1995) distinguishing between processes that operate on a level of representation specific to lexical items and processes that occur at early stages of processing. In this review accounts of segmentation are distinguished by the type of information used to drive segmentation rather than the level of processing at which these mechanisms operate since this avoids potential confusion between sources of information (such as distributional statistics) which although non-lexical in origin may be processed at either a lexical or pre-lexical level.

2.2.1. Statistical accounts of segmentation

An important class of theories of segmentation proposes that the statistical or distributional properties of lexical items can be used as a cue to the location of word boundaries. Such accounts are particularly popular in the developmental literature since they suggest ways in which infants might learn to identify words in connected speech without any obvious cues to determine where boundaries between lexical items are located. The extent to which distributional information is utilised by adults (who have lexical knowledge to bring to bear on the segmentation problem) is unclear. It has been suggested that the processes of lexical access and identification will be much more efficient if the recognition system can reliably identify word onsets pre-lexically (see Briscoe, 1989). The argument is that if a pre-lexical segmentation strategy is used then instead of initiating frequent, unsuccessful lexical access attempts the recognition system can ensure that fewer inappropriate lexical access attempts will be made. The right segmentation strategy would therefore allow more efficient recognition without compromising accuracy[2].

Metrical segmentation strategy

One cue that has been proposed as a lexical segmentation strategy is metrical stress. As proposed by Anne Cutler and others (Cutler & Butterfield, 1992; Cutler & Norris, 1988; Grosjean & Gee, 1987), the metrical segmentation strategy (hereafter MSS), relies on the fact that the majority of content words in stress-timed languages like English have a metrically stressed syllable at their onset. Analysis of a large, phonemically transcribed corpus by Cutler and Carter (1987) showed that 1 in 3 English content words start with a stressed syllable. Furthermore, since these items are more frequent (mostly because

---

[2] This account is tied to an architecture or mechanism for recognition in which there is a specific computational cost involved in lexical search. Recent, parallel access accounts of lexical identification eschew the idea that lexical search (whether successful or not) imposes a specific computational load that should be minimised. Nonetheless, even if not measured in terms of the cost of unsuccessful lexical access attempts, if word boundaries can be detected pre-lexically it might be expected that this source of information would assist lexical access in a parallel processing system.

monosyllabic words, which are all stress-initial, are of high token frequency), a strategy of placing word boundaries before stressed syllables would correctly locate the onsets of 90% of content words in the London-Lund corpus of spoken conversation.

Experimental evidence also suggests that the presence of a strong syllable is used by listeners as a cue to the start of a new word. The word-spotting paradigm has been used to show that listeners are faster to detect monosyllables followed by a strong syllable than by an unstressed (weak) syllable (Cutler & Norris, 1988; Norris, McQueen, & Cutler, 1995; Vroomen, van Zon, & de Gelder, 1996; see McQueen, 1996, for a review of research using the word spotting task). These experiments are reviewed in more detail in Chapter 4.

The MSS has also been proposed as an account of how infants learn to divide the speech stream into words. It has been demonstrated that English-speaking 9-month-old infants display a preference for hearing words that conform to the predominant strong-weak stress pattern (Echols, Crowhurst, & Childers, 1997; Jusczyk, Cutler, & Redanz, 1993; Morgan, 1996; see Jusczyk, 1997 for a review of this and related work).

However, the MSS as described will only operate successfully for open-class or content words that begin with a stressed syllable (Cutler & Carter, 1987). For closed-class words, the reverse (weak-initial) stress pattern is generally found. In the corpus investigated by Cutler and Carter (1987), 69% of weak syllables are at the onsets of closed-class words, with fewer than 5% being the initial syllables of open-class words. In order to utilise the MSS effectively Cutler and Carter propose that two separate strategies operate for accessing the lexical representations of words in separate stores of open- and closed-class items. Strong initial syllables are used to access the open-class lexicon while words beginning with unstressed syllables are looked up in the store of closed class words.

Evidence supporting this dual-lexicon and dual-access strategy account comes from naturally occurring ‘slips-of-the-ear’ and laboratory induced boundary misperceptions (Cutler & Butterfield, 1992). Listeners are more likely to incorrectly add a word boundary before a strong syllable and are more likely to delete word boundaries before weak syllables. In adding or removing words from an utterance, these misperceptions tended to preserve the relationship between initial stress and lexical class: words created from weak syllables were more likely to be closed-class and words with strong initial syllables more likely to be open-class.

In this form, however, the MSS is computationally under-specified. The algorithm proposed by Cutler and Carter (1987) requires information about the metrical structure of extended sequences of syllables. This requires a system capable of storing information about the incoming input until a stressed syllable is received at which point lexical access can be initiated (Grosjean & Gee, 1987; Mattys, 1997). This requires a buffer capable of storing a representation of the input such that it can be analysed retro-actively on the arrival of a strong syllable. The exact construction of such a system is not clearly described by Cutler and Carter; furthermore it is unclear how to reconcile such a theory with what is known of the on-line nature of lexical access in connected speech.

As will be reviewed subsequently in this chapter, there is a great deal of evidence (beginning with the Cohort model of Marslen-Wilson and Welsh, 1978), that listeners construct an on-line interpretation as the speech signal unfolds, without the discontinuities and backtracking required by a stress-based model. Although versions of the MSS have been incorporated into on-line models such as Shortlist (Norris, McQueen & Cutler, 1995), there appears to be some distance between the implementation used in Shortlist (where metrical stress provides a boost to lexical items in the competition process) and the original 'dual-lexicon' conception of the MSS. In its current form, Shortlist does not draw any distinction between the representation and processing of open- and closed-class words.

A final area where the MSS is not completely specified is its requirement that listeners are able to detect syllable boundaries prior to lexical access. This pre-lexical syllabification is required for the MSS to place word boundaries but has not been described so far in the literature. While sonority hierarchies do provide a preliminary grouping of segments into syllabic units, phenomena such as re-syllabification (whereby a sequence such as *band ate* will be syllabified as *ban date*) mean that word boundaries will not necessarily fall at syllable boundaries. One approach shown to be effective for the identification of syllable and word boundaries uses distributional or phonotactic regularities in segment sequences.

Distributional regularity

The use of distributional regularity as a cue to lexical segmentation follows the assumption that chunking the speech stream into frequently occurring sequences will extract linguistically coherent units (Harris, 1955). Computational simulations have

demonstrated the effectiveness of this assumption in extracting words and morphemes from orthographically coded texts (Wolff, 1977) and from natural and artificial speech corpora (Cairns, Shillcock, Chater & Levy, 1997; Perruchet & Vinter, 1998; see Brent, 1999a for a review of several such algorithms).

For instance, Brent and colleagues (Brent & Cartwright, 1996; Brent, 1997; Brent 1999b) describe several variants of a symbolic algorithm that uses distributional regularity to find the set of lexical items contained in a corpus of utterances. The algorithm operates by minimising the description length of the corpus. That is, it compares different sets of lexical items that can be used to transcribe the utterances in the corpus, and chooses the set that uses the minimum number of lexical items, while minimising the total length of these lexical items and maximising the product of the frequency of occurrence of each lexical item. The lexicon discovered by this distributional regularity (DR) algorithm for a phonologically transcribed corpus of child-directed speech corresponds fairly closely to the words contained in the orthographic transcription of this corpus. Performance was improved by providing phonotactic constraints on the system's segmentations (Brent & Cartwright, 1996); these constraints will be described subsequently.

Similar systems have also been developed using on-line learning in a neural network. For instance, one influential simulation reported by Elman (1990) used a simple recurrent network to predict the next input segment in a small artificial corpus. Elman reports that output error drops as the network is presented with more of a word in the training set and rises sharply at the offset of each word. Information from this 'saw-tooth' error could therefore be used to determine which sequences of input segments constitutes a 'word' in the language that the network is exposed to. This system thus provides an alternative implementation of the Harris (1955) 'successor count' approach to lexical segmentation. The network's output error will be high where multiple phonemes can follow the current input – ie. at a word boundary. Thus, the system can be used to determine which sections of the speech input cohere as linguistically salient units.

Simulations reported by Cairns et al. (1997) extended this recurrent network account to a larger corpus transcribed into a distributed phonological representation, again using the 'predict-the-next-segment' task. They tested whether increased prediction error in the network could be used as a cue to determine the location of word boundaries in a corpus of conversational speech. Using a maximally efficient error threshold, Cairns et al. found

that this network successfully identifies 21% of word boundaries in a test set, with a hits to false alarms ratio of 1.5:1. However, they note that the network makes little or no distinction between syllable and word boundaries.

This suggests that the lexical effects obtained in previous small scale simulations do not scale up to more realistic training sets. The network is extracting phonotactic constraints that allow the detection of boundaries between well-formed syllables, not boundaries between words (cf. Gasser, 1992). The success of this approach in detecting word boundaries reflects the fact that many words in English (and high frequency words in particular) tend to be monosyllabic. However, Cairns et al. do not view this result as entirely negative – since the majority of early acquired words in English are mono-syllabic (Aslin, Woodward, LaMendola, & Bever, 1996), such an approach provides an interesting account of how the infant comes to 'bootstrap' segmentation prior to lexical acquisition.

There is also developmental evidence supporting the use of distributional regularity as a cue to the discovery of word units. Preferential listening experiments reported by Jusczyk and Aslin (1995) have shown that infants familiarised with isolated words, will then listen longer to sequences that contain those words. Similarly, infants familiarised on sequences will then listen longer to single words contained in those sequences. However, such results are not equivalent to demonstrating true segmentation where words learnt from connected speech must then be detected in connected speech. Experiments using adult volunteers have shown that extracting word units from the middle of an utterance is considerably more difficult than detecting isolated words or words at the onset or offset of an utterance (Dahan & Brent, 1999). These findings therefore support the predictions of distributional accounts of segmentation such as IncDROP (Brent, 1997) in which segmentation is acquired by detecting and re-using units discovered at the onset and offset of sequences.

Phonotactic accounts

An account that is closely related to these distributional regularity theories involves the use of phonotactic information as a cue to lexical segmentation. Phonotactic accounts are based on the same assumption as in distributional regularity accounts of segmentation: that chunking the speech stream into frequently occurring sequences extracts linguistically

coherent units (Harris, 1955). However the procedures used in phonotactic accounts of segmentation take the opposite approach – looking for word boundaries rather than looking for words. Phonotactic accounts assume that infrequently occurring sequences are likely to contain boundaries between distinct linguistic units. This idea has been implemented in many different forms, using different styles of algorithm to learn the location of potential word boundaries and to place these boundaries into test sequences.

Brent and Cartwright (1996) illustrated the value of phonotactics by incorporating two constraints on the lexical items detected by their DR algorithm. The segmentation performance of the system was improved, for instance, where it was provided with a list of phoneme sequences that never occurred word-internally. These illegal sequences can therefore be assumed to contain a word boundary whenever they appear in a test corpus. Another cue that was also incorporated into these systems was that all words detected must contain a vowel. An alternative implementation of this 'possible word constraint' has also been investigated in lexical competition models (Norris, McQueen, Cutler and Butterfield, 1997). The review of phonotactic accounts presented here, considers constraints on permissible phoneme sequences as a cue to segmentation.

Cairns, Shillcock, Chater, & Levy (1997) reviewed different phonotactic accounts of segmentation, comparing the amount of supervision provided to help the system learn which sequences of segments contain a word boundary. For instance, if the system was told where every word boundary was in the training sequences (but not during the test sequences), it would be described as weakly bottom-up. A more developmentally plausible form of this weakly bottom-up account has been proposed where only segments either side of utterance boundaries are explicitly marked during training (Aslin et al, 1996). Fully bottom-up accounts have also been proposed in which entirely unsegmented utterances are supplied to the system (as was the case for the prediction networks described in the previous section).

A further distinction made by Cairns et al. (1997) was whether phonotactic knowledge was applied categorically or probabilistically in parsing the speech stream into lexical items. Systems with a categorical threshold would assume that any sequence of segments that never occurred word internally must contain a word boundary when found in a test sequence. A probabilistic account would only place boundaries in a phoneme sequence that was below some threshold probability in the training set.

In terms of the distinctions described by Cairns et al. (1997), the phonotactic constraints incorporated into the Brent and Cartwright (1996) system showed no development of phonotactics since legal and illegal sequences were specified *a priori*. Thus phonotactic knowledge as incorporated in this algorithm falls outside of the Cairns et al. categorisations. Later algorithms developed by Brent, such as IncDROP (Brent, 1997), are classified by Cairns as weakly bottom-up, since they receive supervisory input to allow phonotactics to be learnt from utterance boundaries.

Harrington, Watson and Cooper (1989) describe a segmentation algorithm that Cairns et al. (1997) would also categorise as weakly bottom-up, since it acquired sequential constraints by analysis of a training corpus in which all word boundaries are marked. The algorithm encoded knowledge of sequential dependencies, by learning which trigrams must contain a word boundary (such as the sequence /mgl/ which can only occur across a word boundary – e.g. *same glove* containing the sequence /m#gl/) and which trigrams can occur either within a word or across a boundary (such as the sequence /ndl/ in the sequence which occurs in the word *handle*).

Harrington et al. then used a tree-searching algorithm to parse a phonemically-transcribed test corpus into sequences of trigrams, incorporating word boundaries where necessary. Since this system distinguishes categorically between permissible and non-permissible sequences, performance is limited. Cairns et al. (1997) demonstrated that greatly improved performance could be obtained by replacing this categorical distinction between legal and illegal phoneme sequences with a probabilistic cut-off for classifying sequences as containing a boundary or not.

A similar system was also investigated by Aslin et al. (1996) using a network trained to identify phrase and utterance boundaries in a corpus of child-directed speech. Note that unlike other neural network simulations, Aslin and colleagues used a simple feed-forward system that only receives phoneme trigrams as input (a 3 segment window that slides over the training set). Thus this system uses an equivalent input representation to the trigram-based lexical parser described by Harrington, Watson and Cooper (1989).

Since the system receives trigrams as input there is no opportunity for the network to discover lexical items longer than 3 segments in length. Nonetheless, the system is very successful, detecting over half the word boundaries in a test corpus. Given the large

proportion of monosyllables in the training set it is unclear whether this greater level of performance reflects the different task or the different corpus used in this simulation.

These phonotactic accounts of segmentation have been especially influential in the developmental literature since they suggest a means for the acquisition of lexical segmentation without requiring prior lexical knowledge. Evidence supporting infants' knowledge of phonotactics has come from preferential listening experiments suggesting that infants begin to distinguish between sequences that are phonotactically common from those that are infrequent or illegal in their native language in the first year of life (Friederici & Wessels, 1993; Jusczyk, Luce, & Charles-Luce, 1994). Furthermore, research has shown that 9-month-old infants are able to use the fact that certain phoneme sequences are more likely to occur between words than within a word to segment a novel sequence (Mattys, Jusczyk, Luce & Morgan, 1999). Finally, recent work by Saffran, Aslin and Newport (1996) has demonstrated that 8-month-old infants are capable of learning transitional probabilities (i.e. what phonemes are likely to follow on from other phonemes) from only a few minutes exposure to artificial speech stimuli, and can then use this information in detecting familiar sequences.

Higher-level prosodic cues

The metrical segmentation strategy proposed by Cutler and Norris uses one source of prosodic information (the rhythmic alternation of strong and weakly stressed syllables in English) as a cue for the lexical segmentation of connected speech. However, this is only one form of prosodic information that could be used for lexical segmentation. As was seen in the review of the phonetics literature, the duration of segments carries much more information than the level of stress associated with their constituent syllable. Furthermore, information carried by intonation patterns in connected speech – rising and falling fundamental frequency (F0) contours – may also carry useful information for the segmentation of words in connected speech.

Phenomena such as phrase-final lengthening in combination with declining intonation contours may therefore provide a cue to the location of prosodic boundaries. As described by Christophe, Guastie, Nespor, Dupoux and Ooyen (1997), segmentation into phonological phrases provides a useful first step towards extracting lexical and syntactic units from the speech stream. Preferential listening experiments have shown that infants

are able to use this information to distinguish between sequences that are produced as a single lexical item or between two words in adjacent phrases (Christophe, Dupoux, Bertoncini & Mehler, 1994).

Prosodic boundary information may also contribute to the identification of phonotactic cues to word boundaries (such as the use of utterance boundaries in the simulations of Aslin et al. 1996 and Christiansen et al. 1998). Note however, that since phonological phrases are likely to contain more than a single lexical item, the detection of these units does not provide a solution to the problem of segmentation. Christophe et al. therefore suggest that extracting single lexical units may require supplementary segmentation strategies, such as the distributional and phonotactic cues that have been described previously.

Summary

We have seen how a variety of different forms of statistical regularities in the speech stream can be used as a cue in learning to segment connected speech. Each of the cues proposed has been shown to be effective in computational simulations, and also has evidence to support its use in infants' segmentation of connected speech prior to the acquisition of lexical items.

Recurrent network simulations reported by Christiansen, Allen and Seidenberg (1998) investigated the strength of combining different combinations of these cues. Their simulations combined the prediction task used by Elman (1990) and Cairns et al. (1997) with the boundary identification task used by Aslin et al. (1996) along with a metrical stress prediction task. By comparing networks on different combinations of these three tasks, they were able to investigate the effect of each of several different distributional approaches to segmentation (prediction, boundary detection and metrical stress information). They show that best performance is obtained by combining multiple sources of constraint in a single network.

However even when all three tasks are combined (allowing the network to locate approximately 46% of word boundaries), Christiansen et al. fail to find evidence of the lexical effects observed in the small scale simulations reported by Elman (1990). The network is unable to use the fact that phoneme sequences become unique towards the end of words to reduce prediction error and identify word boundaries with confidence. Instead,

the network uses phonotactic constraints at the syllable level to reduce error on the prediction task and identify sequences that are likely to include a word boundary. It is therefore unclear whether the statistical regularities detected by recurrent networks are able to account for lexical acquisition as well as lexical segmentation within a single system. Although these neural networks are capable of detecting a substantial proportion of word boundaries, their success may only reflect the effectiveness of statistical cues in detecting syllable boundaries.

In contrast, the symbolic systems described by Brent and colleagues (Brent & Cartwright, 1996; Brent, 1997; Brent, 1999b), use distributional regularity as an explicit cue for the task of acquiring lexical items. They can then apply this lexical knowledge in the segmentation of subsequent sequences. One conclusion to draw from this comparison of symbolic and connectionist segmentation systems is that current connectionist learning algorithms are insufficiently powerful to use distributional regularity as a cue to discovering words in the speech stream (as opposed to discovering cues to word boundaries).

A recent review of different segmentation algorithms, however, suggests that the 'unbounded' and 'undecaying' memory that is required for these symbolic algorithms to operate contributes to their psychological implausibilty (Brent, 1999a). The symbolic algorithms are also reliant on sequences of invariant segments as input – an unrealistic assumption considering the temporal and spectral variability observed in real speech. There is therefore reason to expect that the performance of these algorithms will get worse when trained on real speech. Connectionist systems on the other hand, have been shown to retain good performance in the face of input variation (Christiansen & Allen, 1997). Consequently there is reason to believe that less powerful, connectionist accounts of lexical segmentation and acquisition, bolstered by additional mechanisms for lexical acquisition where necessary, will turn out to be more psychologically plausible as an account of the acquisition of lexical segmentation and identification. This approach will be pursued in more detail in Chapter 3.

In developing various statistical accounts of segmentation, researchers have described the success of each algorithm in terms of the percentage of boundaries that are successfully identified whilst minimising the ratio of hits to false alarms. However, none of the strategies presented so far comes close to locating 100% of word boundaries – the level of

performance that would be required to use segmentation independently of the identification of words in connected speech. Even combining different sources of distributional information (as in the simulations reported by Christiansen, Allen & Seidenberg, 1998) does not achieve a sufficiently high success rate to produce a solution to the segmentation problem.

Consequently these accounts are unlikely to provide a complete account of segmentation – either in terms of being able to identify all lexical boundaries without identifying lexical items, or by being able to account for the noise and variability that is inherent in connected speech. Accounts of spoken word recognition have therefore continued to incorporate mechanisms by which the identification of individual lexical items can contribute to the detection of word boundaries.

2.2.2. Lexical accounts of segmentation

Lexical accounts of segmentation can generally be divided into two main classes; those accounts that propose that segmentation is achieved by the sequential recognition of individual words in the speech stream (Cole & Jakimik, 1980; Marslen-Wilson & Welsh, 1978) and those that propose that segmentation arises through competition between lexical items that cross potential word boundaries (McClelland & Elman, 1986; Norris, 1994). This review will investigate each of these accounts in turn.

Sequential recognition

One early and influential account of lexical segmentation was proposed as part of the Cohort model of spoken word recognition (Marslen-Wilson & Welsh, 1978). An important property of the Cohort theory is that the word recognition system responds to incoming acoustic information in a maximally efficient manner; words are identified as soon as the speech stream uniquely specifies a single lexical item. This proposal has received widespread support from empirical data showing that words that have an early uniqueness point (i.e. that diverge from all other lexical items early on in the word) can be identified more rapidly than words with a late uniqueness point. Effects of uniqueness point have been demonstrated in many tasks such as shadowing (Marslen-Wilson, 1985), word monitoring (Marslen-Wilson & Tyler, 1975), gating (Grosjean, 1980; Tyler, 1984), lexical decision (Marslen-Wilson, 1984; Taft & Hambly, 1986), cross-modal priming (Marslen-Wilson, 1990; Zwitserlood, 1989) and in effects on eye-movements (Allopenna,

Magnuson, & Tanenhaus, 1998). They have also been demonstrated in languages other than English including French (Radeau & Morais, 1990) and Dutch (Tyler & Wessels, 1983).

The early identification of words predicted by the Cohort model plays a valuable role in lexical segmentation. Since words can be identified at their uniqueness point, which is often before their acoustic offset, it is possible to identify the offset of the current word and to interpret speech following that offset as coming from subsequent words. Thus a system in which words are recognised before their offset does not require marked lexical boundaries. It would be a straightforward matter for such a system to identify the start of a word as being the section of speech that comes after the offset of the current word. Thus the maximally efficient processing proposed in the Cohort model provides a means by which adult listeners can lexically segment connected speech.

It has recently been demonstrated that a simple recurrent network (Elman, 1990) trained to map from a phonemically coded representation of the speech stream to either a localist or distributed lexical representation of the current word provides a direct implementation of many of the desirable properties of the Cohort model (Norris, 1990; Gaskell & Marslen-Wilson, 1997). These networks learn to partially activate all the lexical candidates that match the current input, with the degree of activation approximating the probability of that lexical item given the current input. The full details of these simulations will be described in more detail in Chapter 3; this chapter focuses on the sequential recognition account of segmentation that these networks implement.

As was the case for the Cohort model described by Marslen-Wilson and Welsh (1978), the networks operate in a maximally efficient manner with candidates becoming de-activated when mismatching input is received and full activation only occurring when a single lexical candidate matches the speech stream. These networks will similarly use a sequential recognition strategy in dividing the speech stream into words – a new word is activated when input is received that follows on from the end of an already identified word.

However, gating experiments have shown that listeners do not always identify words before their acoustic offset (Bard, Shillcock, & Altmann, 1988; Grosjean, 1985). As will be discussed in more detail in Chapter 4, a lexical segmentation strategy reliant on

sequential recognition would be disrupted by the 20% of words that are not recognised until after their acoustic offset in the Bard et al. (1988) gating study.

Investigation of the structure of a large lexical database supports the results of these gating experiments by showing that many words do not diverge from other lexical candidates until after their final phoneme (Luce, 1986). These are words that are embedded at the onset of longer lexical items (such as *cap* embedded in *captain, captive* etc.). Luce argues that such items would prove problematic for the Cohort model since it would not be possible to rule out longer competitors before the offset of a word. These embedded words will be particularly problematic, since short words (which are most likely to be embedded in this way) occur more frequently in natural language. By Luce's calculations, 38% of words in English (when weighted by frequency of occurrence) become unique after their offset. Consequently a sequential recognition account of segmentation would not be feasible for sequences that contained onset-embedded words.

The recurrent network simulations reported by Gaskell and Marslen-Wilson (1997) and Norris (1990) make this failure of sequential recognition accounts particularly transparent. At the end of a sequence of phonemes like /kæp/ the output of the network is in an ambiguous state with both the embedded word *cap* and all longer candidates (*captain, captive* etc.) activated. In the case where the network is presented with an embedded word (such as *cap* in the sequence *cap fits*) the network will begin to activate a new set of candidates beginning with the segment /f/ (*feel, fall, fit,* etc.) at the onset of the following word. For this reason, the network never unambiguously activates short words and is therefore incapable of recognising words that are embedded at the onset of longer competitors. Thus the presence of large numbers of embedded words may prove fatal for accounts of lexical segmentation based on sequential recognition.

Further lexical database searches carried out by McQueen et al. (1995) report similarly pessimistic statistics regarding the presence of embedded words in English and draw stronger inferences from the failure of these recurrent network accounts. They firstly consider the possibility that syllabic information can constrain the numbers of embeddings that will be found. However, even when embedded words had to have an identical syllabification to the word in which they were embedded these searches still found large proportions of words with other lexical items embedded at their onset. McQueen et al. report that 57.5% of polysyllabic words have another word embedded as their initial

syllable. Further searches showed that excluding function words embedded in content words failed to substantially reduce the proportion of embeddings observed. Even including grammatical class as an additional constraint failed to remove lexical embedding as a problem for word recognition. McQueen et al. used the results of these corpus searches to draw conclusions about the necessity of competition between lexical hypotheses in models of spoken word recognition.

Lexical competition

The presence of large numbers of onset-embedded words has thus been used to argue against sequential recognition accounts of lexical segmentation. In order to recognise these embedded words, longer competitors need to be ruled out. For example, to recognise words in the phrase *"cap fits"* the recognition system needs to reject alternative interpretations of the initial syllable such as *captain*. Such a process is suggested to require information arriving after the offset of the embedded word – hence the delayed recognition observed in gating experiments by Grosjean (1985) and Bard et al. (1988).

One computational mechanism for this delayed recognition is provided by the TRACE model of speech perception (McClelland and Elman, 1986) and re-used in Shortlist (Norris, 1994). The network architecture and representations used in these models is reviewed in greater detail in Chapter 3. This section evaluates the account of lexical segmentation proposed in TRACE, which is based on inter-lexical competition.

In TRACE, the strength of evidence supporting each lexical item is represented by the activation of the relevant lexical unit. However, in addition to receiving input from lower levels of analysis, lexical units in TRACE are interconnected with inhibitory links. These inhibitory links produce competition between lexical units that is used to rule out mutually exclusive lexical hypotheses; for example, the hypotheses that *cat, cattle* and *catalog*, are all present in the sequence *"cattle hog"*.

In TRACE, the strength of the inhibitory connections between two lexical units depends on the number of segments shared by these lexical hypotheses – so there would be a strong, inhibitory connection between the unit representing *cat* and all other words that contain the segments /k/, /æ/ and /t/ in that order. The strength of these inhibitory connections is proportional to the number of segments shared between the competing

words – such that pairs like *cattle* and *catalog* are in much more direct competition than pairs sharing fewer segments.

This arrangement results in a large and highly interconnected network of lexical units, a portion of which is illustrated in Figure 2.1 below. The effect of this competition network is computationally fairly simple. The network instantiates a large, parallel constraint-satisfaction system (cf. Smolensky, 1986) where the problem being solved is to assign segments in the input to lexical items such that a consistent lexical parse of the input is achieved. That is, the lexical network should settle into a state in which only lexical units for appropriate words have been activated, and all segments in the input have been assigned to only one word.

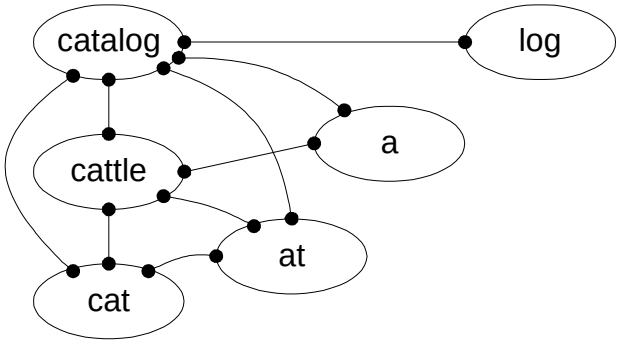

Figure 2.1:Pattern of inhibitory connections between a subset of candidates activated by the presentation of /kætælog/ in Shortlist. Adapted from Norris (1994), Figure 2.

In practice, since TRACE is required to activate an ongoing interpretation as segments are presented in the input, the process of constraint satisfaction will be an imperfect approximation to that obtained by a fully parallel process. However, in simulations reported by Frauenfelder and Peeters (1990), TRACE is able to cope with many forms of segmentation ambiguity. For instance, lexical garden-paths (where the input matches a longer competitor though, in fact, coming from two words such as the sequence *car pick* with the competitor *carpet*) can be correctly identified by TRACE.

It is important to note that these transitory ambiguities are only part of the problem faced by the recognition system in segmenting the speech stream. There exist sequences such as *car pit* and *carpet* where two competing sequences may contain identical segments. Although Frauenfelder and Peeters were able to demonstrate that TRACE could distinguish these two sequences when word boundary markers were placed in the input (i.e. when the sequence *car pit* was presented with a gap between the two words), this

form of explicitly marked word boundary is a poor representation of the properties of the speech stream.

Summary

In this review of lexical accounts of segmentation the difficulties faced by sequential recognition accounts in identifying words embedded at the onset of longer words have been described. Given these difficulties, the presence of large numbers of onset-embedded words in searches of phonologically transcribed lexical databases (Luce, 1986; McQueen et al., 1995) has been used to support accounts of identification that incorporate competition between lexical hypotheses which are able to identify onset-embedded words.

However, before concluding with Luce (1986) and McQueen et al. (1995) that the presence of words embedded at the onset of longer words rules out sequential recognition accounts of lexical segmentation, it is worthwhile to examine the kinds of embedded words that are found by these dictionary searches. To draw an inference from the structure of the language to the cognitive architecture underlying spoken word recognition requires that the assumptions used for the dictionary searches match what is known about the recognition system. As we will see, these assumptions can be called into doubt and we may therefore hesitate before accepting the conclusions of McQueen et al. (1995), that lexical competition is necessary in accounts of spoken word recognition.

2.3. Onset-embedded words and segmentation

The theoretical significance of onset-embedded words is that they are cases in which a sequential (Cohort-style) recognition strategy will break down. Given that such a system is hypothesised to operate in a ‘left to right’ fashion without backtracking, the cases that are of importance are those where one lexical item is phonologically embedded at the start of a longer item. The dictionary searches carried out by Luce (1986) and McQueen et al. (1995) identified such cases to different levels of phonological precision. In Luce's searches, embedded words were considered to be any word which contained the phonemes making up another word, while McQueen only considered syllabic embeddings to be relevant. Since the production of a segment will vary in different positions within a syllable, the stricter criteria employed by McQueen seem to be justified in this case. However, even these stricter criteria fail to consider the possibility that acoustic cues to

word boundaries may play a role in lexical segmentation and identification. This issue will be explored further in Chapters 4 and 5.

One assumption shared by the dictionary searches of Luce (1986) and McQueen et al. (1995) is that all the words listed in the dictionary are separate units in the mental lexicon. More specifically, they take a full-listing approach to the representation of morphologically complex words: assuming that the lexical representation of the word *dark* embedded in a transparently derived word such as *darkness* is an entirely separate representation, just as *whiskey* is separately represented from the embedded word *whisk*.

Experiments using cross-modal repetition priming (Marslen-Wilson, Ford, Older, & Zhou, 1996; Marslen-Wilson, Tyler, Waksler, & Older, 1994) cast doubt on this assumption, suggesting that where two lexical items are transparently related (such as *dark* and *darkness*), the lexical representation of the morphologically complex word is formed out of a representation of the stem combined with a representation of an affix. Consequently a lexical system that accessed the embedded word *dark* while processing a related word such as *darkness* would not be required to back-track or revise its hypothesis. In a morphologically-decomposed lexicon the presence of these onset-embeddings would help, not hinder, the recognition system. It is therefore worthwhile to revise these estimates of the proportion of words containing a lexical embedding in the light of a more realistic, morphologically-decomposed view of the mental lexicon[3].

Another common form of morphological relationship that can result in words being phonologically embedded in other words is compounding. However, the processes by which two morphemes combine to form a compound tend not to be as semantically transparent as those by which stems and affixes combine (consider, for instance, the

---

[3] Interestingly, since both Luce and McQueen used databases derived from dictionaries they excluded regular, morphologically inflected forms in their counts. Thus embeddings like *jump* in *jumped* and *jumping* or *cat* in *cats* would not be counted. Although not discussed by either authors, inflected forms are excluded from most dictionaries since they are phonologically and semantically transparent and can thus be derived from knowledge of the stem. In the dictionary searches reported here we investigate the effect of extending a form of this assumption to derivational and compounding morphology.

different relationsips involved with the morpheme *house* in *houseplant*, *houseboat* and *housewife* or the different mearnings of the *mill* in *windmill*, *sawmill*, *peppermill*). Consequently the lexical representation of a compound word may be less clearly formed out of its constituent morphemes than was the case for representations of derived words – despite the similar results obtained in priming experiments using compounding morphology (Xhou and Marslen-Wilson, in press). Consequently the database counts reported here will consider derivational and compounding morphology separately.

2.3.1. Counting embedded words

In order to measure the effect of excluding morphologically related words on the proportion of embedded words found, a lexical database that includes morphological decompositions is required. One such database is CELEX (Baayen, Pipenbrook, & Guilikers, 1995) which incorporates decompositions for all polymorphemic words. Although care must be taken since the database decomposes some semantically opaque items such as {apart}+{ment} for *apartment* and {black}+{mail} for *blackmail*, this information should at least permit an initial investigation of what proportion of embeddings would be ruled out by an morphemically-organised mental lexicon.

Materials

Since these counts use a different database to McQueen et al (1995), a first step will be to try to replicate their counts as closely as possible. With this aim in mind, the CELEX lemma database was used (which, in common with the LDOCE database used by McQueen, excludes the majority of inflected forms). In line with the procedure used by McQueen, multi-word and phrasal lemmas (e.g. *funny peculiar*, *billiard-table*) were removed. Also excluded were words that were one letter or one phoneme long (e.g. letters of the alphabet and exclamations like *oh*, *eh*, etc.), proper nouns and words with 7 or more syllables. These exclusions removed 12812 of the original 52447 lemmas in the CELEX database leaving 39635 lemmas (in which there were 33713 unique phonological

strings[4]). The database used for these searches is therefore approximately 30% larger than that used by McQueen.

Method

These phonological strings were searched for words embedded in longer words. As described by McQueen, embeddings were only counted if they perfectly matched the syllabification of the longer word (for example *can* is embedded in *canvas* but not in *cannibal*); also the stress value associated with each syllable was not used to exclude embedded words (for example, *can* was embedded in *canteen*). One departure from the method used by McQueen was that only the proportion of words embedded at the onset of longer words is reported since this is the more critical case for the sequential recognition account. Only statistics on onset-embedded words will be reported from both from the current searches and from the results of McQueen et al. (1995).

Results

The proportion of unique phonological strings between two and six syllables in length which contain an onset-embedded word is shown in Table 2.1. Comparing the current results (marked P(embed.) CELEX, + Phon) with those obtained by McQueen et al. (1995) it can be seen that there is some disagreement between the two sets of counts. McQueen found that 57.5% of polysyllables have a monosyllabic word embedded at their onset while these searches report that only 39.4% of polysyllables contain a monosyllabic word as their first syllable. Although the discrepancy is less marked for other lengths of embedded word, this difference does seem surprising. It is possible that this is merely the result of using a larger database. If, for instance, the additional words that are in CELEX and not in LDOCE do not have as many onset-embedded words in them, including these

---

[4] This discrepancy between the number of lemmas and the number of unique phonological strings reflects the presence of homophones in CELEX as well as separate entries for the same word occurring in different syntactic classes - e.g. noun and verb forms of *brush*. In counting the proportion of words that contain embedded words we will follow the procedure employed by McQueen et al. (1995) of counting the proportion of phonological strings that contain an embedded word.

additional words would reduce the overall proportion of polysyllabic words that contain an embedding.

To rule out this explanation searches were carried out using only words that are in both CELEX and LDOCE. This reduced the set of lemmas searched to 30296 lemmas containing 24756 phonologically unique strings. The results of these searches are also shown in Table 2.1 (column marked P(embed.) CELEX & LDOCE +Phon). Comparing these searches with the results obtained previously shows that of the 21 259 polysyllabic words in both CELEX and LDOCE, 38.4% contain an onset-embedded monosyllable (compared to 39.4% for the full CELEX database and 57.5% in McQueen et al., 1995). Thus there remains a large discrepancy in the counts of embedded words. Clearly, the difference between these counts and McQueen et al. are not just the result of using a larger database. It remains possible that there are additional monosyllabic words in LDOCE that are not included in CELEX that could distort these results. Another possibility is that in LDOCE, more polysyllables are transcribed as having full vowels in their initial syllable. This would have the effect of increasing the proportion of polysyllables having an onset-embedded monosyllabic word. Having established these discrepancies between LDOCE and CELEX, it is clear that further searches must all be done within the same lexical database. For this reason future searches will use the full CELEX database, where both phonological and morphological information can be considered in parallel.

2.3.2. Counting morphological embeddings

As argued previously, merely calculating the proportion of dictionary words that contain an embedded word may overestimate the problems created for a model of lexical access. Many of these embeddings may be morphological in nature – for example, a word like *darkness* would be counted as containing an embedded word *dark*. By a morphemic theory of lexical organisation, such an embedding may actually prove beneficial since accessing the semantics of the stem *dark* early on in the complex word will facilitate processing. The counts reported previously were therefore repeated, excluding cases in which the embedded word was morphologically related to the carrier.

Syllables in carrier word	Syllables in embedded word	P(embed.) LDOCE McQueen (1995)	Carrier words CELEX	P(embed.) CELEX + Phon.	Carrier words CELEX & LDOCE	P(embed.) CELEX & LDOCE + Phon.
2	1	0.651	12095	0.495	10239	0.468
3	1	0.525	9832	0.346	6764	0.313
	2	0.246		0.265		0.141
4	1	0.465	5429	0.288	3167	0.281
	2	0.140		0.105		0.093
	3	0.091		0.197		0.039
5	1	0.480	2159	0.328	933	0.328
	2	0.144		0.095		0.090
	3	0.058		0.031		0.025
	4	0.060		0.172		0.021
6	1	0.535	560	0.343	156	0.404
	2	0.186		0.141		0.115
	3	0.087		0.038		0.064
	4	0.058		0.036		0.045
	5	0.017		0.104		0.026

Table 2.1: Proportions of words with unique phonology containing other words embedded at their onset – comparison of McQueen et al. (1995) LDOCE and CELEX counts.

Method

To exclude embedded words that were morphologically related, the morphemic segmentations listed in the CELEX database were used. For example, the word *darkness* is decomposed into a stem {dark} and an affix {-ness} in the database. Consequently the word *dark* would not be counted as being embedded in *darkness*.

Given the discussion that exists in the experimental literature concerning the decomposition of compound words such as *darkroom*, two sets of counts were carried out. In the first, only forms in which one or more affixes (e.g. -ness, -able, -ment etc.) were added to an embedded word would be classed as being morphologically embedded. In a second set of searches a more relaxed definition of morphological relatedness was considered. In this second set, any embedded word that was included in the morphological decomposition of the carrier word would not be classed as embedded, regardless of whether subsequent units are classed as an affix or not.

Results

The results of these searches are shown in Table 2.2. Counts discarding derivational embeddings are shown in the CELEX +Phon -Deriv column, while counts discarding both derivational and compounding morphology are shown in the column labelled CELEX +Phon -Deriv -Compound. Results of the original searches of the CELEX database are included for comparison purposes.

As can be seen in Table 2.2, excluding morphological embeddings markedly reduces the proportion of polysyllabic words that contain an embedded word. Previous searches of the CELEX database revealed that 15187 out of 30075 polysyllabic words (50.5%) have at least one onset-embedded word. Follow-up searches excluding polysyllables that were derivationally related to their embedded words reduced the number of polysyllables with one or more embedded words to 11728 (39.0%). Excluding onset-embeddings in compound words reduces this number still further such that only 8077 polysyllables had a non-morphological embedded word (26.9%). Of these 8077 polysyllables, the average number of embeddings per word was just 1.04. This result indicates that by far the majority of polysyllabic words in English only have morphologically-related forms embedded within them[5].

---

[5] The greater numbers of embeddings rejected for being compounds rather than for being derived forms does not indicate that there are more compounds than derived forms in CELEX. Rather, it reflects the fact that compounding is more likely than derivational morphology to preserve the phonology of the stem without changes in syllabification as occur when adding the affixes *–able, -ily*, etc.

Syllables in carrier word	Syllables in embedded word	CELEX + Phon.	CELEX + Phon - Deriv.	CELEX + Phon - Deriv - Compound
2	1	0.495	0.416	0.240
3	1	0.346	0.316	0.242
	2	0.265	0.081	0.054
4	1	0.288	0.288	0.233
	2	0.105	0.077	0.065
	3	0.197	0.010	0.008
5	1	0.328	0.327	0.231
	2	0.095	0.090	0.082
	3	0.031	0.007	0.006
	4	0.172	0.013	0.012
6	1	0.343	0.343	0.257
	2	0.141	0.127	0.116
	3	0.038	0.018	0.009
	4	0.036	0.004	0.004
	5	0.104	0.020	0.020

Table 2.2: Proportions of words with unique phonology containing other words embedded at their onset. CELEX counts including phonology and excluding either derivationally embedded words or both derivational and compound words

Comparing words of different lengths, it appears that by far the greatest change between searches that consider morphological relationships and those that do not is to be found in comparisons where the embedded word is one syllable shorter in length than the carrier word. Where morphological structure was incorporated fewer embeddings were found in which the embedded word is one syllable shorter than the carrier. This is consistent with

the conclusion that the majority of embeddings are morphological in nature and are created by the addition of another morpheme to the offset a previously existing word.

One aspect of these searches that remains problematic is that not all of the morphological decompositions listed by CELEX will necessarily be semantically transparent. For example *apartment* would not be counted as containing the embedded word *apart* since it is listed in the database as {apart}+{-ment}. However, experiments using cross-modal priming (Marslen-Wilson, Tyler et al., 1994) suggest that derived forms which have an opaque relationship with their stem (such as *apartment* and *apart*) are not decomposed in the mental lexicon. Similar findings are also reported by Zhou and Marslen-Wilson (in press) for compound forms, again suggesting a decomposed lexical representation is only used for morphologically complex forms that are related to the meaning of their constituents in a semantically transparent manner.

2.3.3. Semantic relatedness

The standard means of assessing the semantic transparency of morphologically complex words is to obtain semantic-relatedness (SR) ratings from native speakers (Marslen-Wilson et al. 1994). For instance, two forms which are as transparently related (e.g. *darkness* and *dark*) would receive a high SR rating (in this case a mean rating of 8.5 out of 9 – where 9 is highly related in meaning and 1 is unrelated). Opaquely related forms such as *apartment* and *apart* would receive a substantially lower rating in this test (rating 2.1 out of 9). One concern about the results obtained in these corpus counts might therefore be that many of the morphological embeddings that were discarded will turn out to be semantically opaque and hence not decomposed in the mental lexicon.

Inspection of a database of semantic-relatedness judgements however, suggests that opaque forms (such as *apart-apartment*) are by far the minority of morphologically complex words. Of the derivational morphological embeddings rejected in our searches, 214 have been rated for semantic relatedness. Of these items, 82.7% are designated as transparent (SR > 7) and 4.7% are opaque (SR < 3.5). The same comparison, applied to the embeddings that were kept in as being non-morphological, shows the reverse pattern. Of the 236 pairs which had received a rating only a small proportion were rated as transparent (19.9%) and a larger proportion were rated as opaque (30.9%).

For compounds that contained an embedded word the statistics again showed that items considered to be morphological were more likely to be semantically transparent than semantically opaque. Out of 207 compounds with phonologically embedded words that had been rated, 52 were rated as transparently related to their embedded word (25.1%) while 27 were rated as being opaque (18.8%) with a mean SR rating for these 207 items of 5.3. This suggests that despite the reduced semantic relatedness of compound words to their constituents, embedded words were, for the most part, transparently related to the longer carrier word. As indicated in Table 2.2, there are very few morphologically unrelated embedded words found by this search. However, of the 46 items found that had a semantic relatedness rating, the majority of these items were opaque. Only 1 item (*laughter - laugh*) which was not decomposed by CELEX was rated as transparent (the lack of a decomposition for *laughter* appears to be an oversight in CELEX). For the 46 pairs that had received a rating, the overall mean semantic relatedness was 2.3 out of 9. Of these items, 35 pairs (76%) were rated as opaque.

Since the SR ratings compiled in this database were generated for use in morphology priming experiments it cannot be assumed that this data constitutes a random sample of the English morphological system. However, this point aside it is clear that the conclusions drawn from cross-modal priming experiments for the most part hold true. Derivationally related items are, by and large, transparently-related and hence are likely to be decomposed in the mental lexicon. Conversely items that are not morphologically decomposed in the CELEX database are almost entirely semantically opaque. The situation for compound forms is more mixed, with a range of transparent and opaque forms being listed as decomposed in the CELEX database.

2.3.4. Discussion

It has been shown that assuming a morphologically structured mental lexicon substantially reduces the proportion of lexical items that have a shorter word embedded within them. This reduction is especially apparent in the proportion of words of three or more syllables that have a polysyllable embedded within them. This makes good intuitive sense – the majority of long words are morphologically complex and it is therefore likely that they would have their stems embedded within them. However, as the figures above show, there are still large numbers of mono-morphemic bisyllabic words that have monosyllables

embedded at their onset. These items (such as *captain* containing the embedded word *cap*) may still present a problem to a sequential system that uses early recognition as a means of lexically segmenting the speech stream. Therefore, despite attempts to assess the number of embedded lexical items in a more realistic fashion, the central argument of Luce (1986) and McQueen et al. (1995) – that onset-embedded items require delayed recognition – can not at this point be rejected.

As was discussed in the review of the phonetics literature, however, there are acoustic cues, for instance the increased duration of syllables in monosyllabic words, that may contribute to the detection of a word boundary for these embedded words. Consequently, to the extent that dictionary searches assume sequences of undifferentiated phonemes as input to the recognition system, they will overestimate the degree of ambiguity created by onset-embedded words. However, since duration differences between syllables in short and long words have not unequivocally been shown to be used by listeners, it is not possible to assess whether these counts of embedded words still overestimate the ambiguities created by embedded words. This issue will be pursued further in experiments reported in chapters 4 and 5.

2.4. General Discussion

In this chapter a wide-ranging review of the literature on lexical segmentation has been presented. From the range of topics and different fields that have been covered it appears that the problem of how to detect the boundaries between words in connected speech has been an important issue for fields as diverse as acoustic-phonetics, computer speech recognition and developmental psychology as well as for psycholinguistics. One conclusion that can be drawn from this mass of data and theory is that the problem of segmenting speech is fundamentally unlike that of reading printed words on a page. The placement of word boundaries may in some cases be inferred from a combination of noisy and unreliable cues; however these cues are often absent. In view of this difficulty it is likely that adult listeners bring the full force of their lexical knowledge to bear on the problem of segmenting the speech stream into words.

Two alternative accounts of how listeners use lexical knowledge in the segmentation of connected speech have been described. As we have seen, the sequential recognition strategy proposed in the original form of the Cohort model (Marslen-Wilson and Welsh,

1978) has been strongly criticised for its apparent inability to deal with onset-embedded words. Recent models of spoken word recognition such TRACE (McClelland and Elman, 1986) have therefore incorporated competition across word boundaries to achieve lexical segmentation. This theoretical shift has been caused, in part, by experimental evidence from the gating task that will be reviewed in more detail in Chapter 4, and also by arguments based on the results of searches of lexical databases (Luce, 1986; McQueen et al., 1995).

However, it could be argued that inferring mental architecture from the contents of machine-readable dictionaries is at best a weak inference. Database searches reported in this chapter have shown that a morphemic view of the units of lexical representation substantially reduces the number of lexical embeddings observed in English. In the review of the phonetics literature it has also been suggested that acoustic differences between syllables in short and long words might permit listeners to distinguish embedded words from their longer competitors.

Each of these pieces of evidence questions the assumptions of the lexical database searches. In combination they undermine the strength of the argument from these searches that lexical competition is ‘necessary’ for the identification of onset-embedded words. However, perhaps the most direct challenge must come from computational modelling itself. As was discussed in the introductory chapter it is unsafe to conclude that a particular pattern of behavioural data requires a particular class of model. Onset-embedded words that require delayed recognition will only necessitate lexical competition if all other models are incapable of identifying these words. It is this question that is addressed in the following chapter.