# Lexical Segmentation in Spoken Word Recognition

**Matthew Harold Davis**

Birkbeck College,

University of London

Thesis submitted for the degree of Doctor of Philosophy

March 2000

# Abstract

This thesis examines an important issue in spoken word recognition; how the perceptual system segments connected speech into lexical units or words. Research on this topic has investigated the role of different sources of information in dividing up the speech stream: acoustic cues in the speech signal, statistical regularities in the structure of the language or through the identification of individual lexical items.

This research focuses on cases in which the location of word boundaries may be ambiguous by one or more of these segmentation mechanisms using words embedded at the onset of longer words (such as *cap* in *captain*). The ambiguities proposed for onset-embedded words have motivated accounts of segmentation based on competition between alternative parses of speech into words. In these accounts, the recognition of embedded words is delayed until after their offset when subsequent input rules out longer competitors. In this thesis it is demonstrated that training a simple recurrent network to activate a representation of all the words in a sequence allows a connectionist network to learn the appropriate delay to allow the identification of onset-embedded words without requiring directly implemented competition between words.

Both lexical competition and recurrent network models assume ambiguity between onset-embedded words and equivalent syllables in longer competitors. Acoustic analysis carried out in this thesis confirms the presence of reliable acoustic differences between syllables in short and long words. A series of experiments using gating and cross-modal priming suggest that the perceptual system uses these acoustic differences to discriminate embedded words from the onset of longer competitors and that match or mismatch with longer competitors may be less important for the identification of onset-embedded words. These results are interpreted within a revised version of the recurrent network model, incorporating input representing the acoustic differences between syllables in short and long words.

# Acknowledgements

I am extremely grateful to my supervisors, William Marslen-Wilson and Gareth Gaskell, for their guidance and support throughout the course of this work. I have also benefited from discussions with John Bullinaria, Morten Christiansen, Gary Cottrell, Tom Loucas and Billi Randall on many of the topics contained in this thesis.

At a physical rather than intellectual level, this thesis could not have been written without the assistance of Dr. Whelan, Mr. Cobb and many others at the Middlesex hospital. The care and support of friends and family during my illness helped make this time a frustration rather than a torment.

A further debt of thanks is also owed to my colleagues past and present in the Centre for Speech and Language both at Birkbeck and in Cambridge who provided such a friendly atmosphere in which to work. In particular, thanks must go to Billi Randall and Tom Loucas who helped make my time at CSL so enjoyable. Thanks are also due to Lolly Tyler who faced my divided loyalties in the past year with great stoicism and to Helen Moss and Kathy Rastle for help with proof-reading.

I would like to thank Joe Devlin, Mike Ford and Richard Russell for hospitality and barbecues in Cambridge and Liz Gresham for assisting with matters of style. I also wish to thank my parents for their encouragement throughout my studies. I have kept them from retirement for more years than I should have.

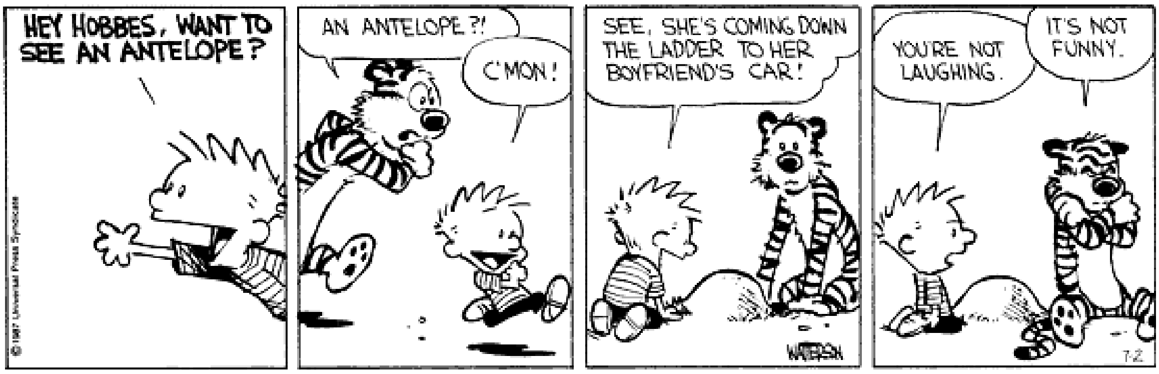
Most important of all though, I would like to thank Maggie Kemmner for advice and encouragement throughout and for the love and support that helped me through the most difficult times.

This research was supported by a studentship from the Engineering and Physical Sciences Research Council.

# Contents

# List of tables

# List of figures


HEY HOBBES, WANT TO SEE AN ANTELOPE?
© 1987 Universal Press Syndicate
AN ANTELOPE?!
C'MON!
SEE, SHE'S COMING DOWN THE LADDER TO HER BOYFRIEND'S CAR!
WATTERSON
YOU'RE NOT LAUGHING.
IT'S NOT FUNNY.
7-2

# 1. Introduction

One of the most fundamental of human skills is the perception of connected speech. The communication abilities provided by spoken language are the most obvious division between humans and other animals. This thesis investigates one small aspect of the cognitive skills that contribute to our ability to understand speech – the segmentation and recognition of words in connected speech.

## 1.1. Spoken word recognition

Word recognition plays a central role in the processes by which acoustic waveforms are converted into a representation of the meaning of utterances. Accounts of spoken language comprehension typically postulate initial processing stages which extract relevant perceptual information from the acoustic signal. The term *word recognition* applies to the processes by which these perceptually-derived input representations make contact with stored representations of the words being identified. Processing stages following word recognition are then concerned with integrating the individual syntactic and semantic properties of the recognised words into a representation of an utterance's meaning.

In this very abstract description, the processes associated with spoken word recognition appear little different from those involved in the comprehension of printed text. Indeed, early models of spoken word recognition were derived from existing knowledge of visual word recognition (Forster, 1976; Morton, 1969). This debt remains apparent in some more recent accounts of spoken language processing (see for instance, Bradley & Forster, 1987; Luce, Pisoni, & Goldinger, 1990). However, with increasing competence in the experimental manipulation of speech stimuli (mostly through the use of digital computers for the recording and playback of speech) many of the unique properties of spoken language have become available to psycholinguistic investigation.

Perhaps the most obvious difference between spoken and written language is that while written language can be perceived in parallel for as long as is required for processing (at least at the level of individual words) spoken language is sequentially ordered and transient. Since only a small amount of speech can be retained in the auditory system in an unanalysed or echoic form (see for instance Crowder & Morton, 1969; Huggins, 1975) it seems that the processing infrastructure for speech perception must operate rapidly if connected speech is to be processed efficiently. The question of how closely the processing of connected speech tracks the auditory input is an important issue in research on speech perception (Mattys, 1997). However, the immediacy with which we perceive and are able to respond to incoming speech (see for instance Marslen-Wilson & Tyler, 1980) suggests that fast and effective processing of spoken language is a normal property of adult comprehension.

Another important problem for spoken language comprehension that is absent in reading printed words is created by inherent variability in the speech signal. The development of written language and use of the printing press required explicit agreement about the exact form of written letters (though consider the difference between **a**, ***a*** and **A**). The evolution of language and the development of the speech articulators in humans allowed no equivalent standardisation (though Stevens & Blumstein (1981) have described potentially invariant cues to the identification of some classes of phonemic segment). For these reasons, variation in the acoustic form of the speech signal is pervasive at many levels of analysis.

One source of variation in the acoustic form of the speech signal is caused by differences between speakers. These may be caused by differences in accent or dialect, or through differences between individual speakers in the rate and pitch of their speech. Experimental evidence suggests that familiarity with the particular characteristics of individuals' speech can facilitate identification of words even with several days delay between successive experiences of a given speaker uttering a particular word (Goldinger, 1996a).

Another form of variation is caused by pressure to fit the discrete phonological gestures associated with segments into a connected stream of speech. The resulting coarticulation and deletion of segments, where speech sounds are altered through the influence of preceding and following phonemes, may result in dramatic differences in the production of individual words. The word *stand* for instance will rarely be pronounced in its canonical form /stænd/ but may surface as /stæn/ in *stand down*, as /stæŋ/ in *stand close* or as /stæm/ in *stand back* (for experimental and computational investigations into the processing of phonological variation, see Gaskell, Hare, & Marslen-Wilson, 1995; Gaskell & Marslen-Wilson, 1996).

A third class of problem specific to spoken as distinct from written language is that of segmentation. Alphabetic writing systems are composed of discrete units (letters) formed into larger chunks (words) which are then organised into longer sequences (sentences and paragraphs). There is therefore a physical separation of orthographic units at multiple levels which is likely to provide at least an initial structure for the mental representation of written language. However, this physical organisation is not found in the structure of spoken language, which is to a great extent continuous and formed out of connected and co-articulated units not bounded by spaces or other breaks in the speech signal. Consequently the processes by which speech is divided into low level units, as well as the nature of these units, remains open to widespread debate. Different authors have proposed that perceptual processing of speech proceeds from spectral representations (Klatt, 1979), phonetic features (Warren & Marslen-Wilson, 1987; 1988), phonemes (Pisoni & Luce, 1987) or syllables (Mehler, Dommergues, Frauenfelder, & Segui, 1981).

A similar debate exists regarding the representation of higher-level lexical units in spoken word recognition. Most authors have assumed that the units of lexical storage correspond to an orthographic word as written on the page. However, research on the representation and processing of derivational morphology has been used to argue for a decomposed mental lexicon in which representations of stems and affixes combine to represent morphologically complex words (Marslen-Wilson, 1999; Marslen-Wilson, Tyler, Waksler, & Older, 1994). Conversely, experimental evidence demonstrating a ‘word superiority effect’ for familiar word combinations (e.g. *greasy spoon*) has been used as evidence for stored representations of larger units consisting of more than one orthographic word (Harris, 1994; Harris, 1996). This conflict suggests that the lexicon is unlikely to contain one single size of lexical unit. Consequently questions about the computational mechanisms that divide the speech stream into lexical units are likely to be essential in describing the operation of the language processor.

## 1.2. Lexical segmentation

The segmentation of connected speech into lexical units is a major topic for psycholinguistic research and is the primary focus of the research reported in this thesis.

The use of the term ‘lexical segmentation’ throughout this thesis is not intended to imply that only processes at a lexical level are involved in segmentation; rather this term refers to the segmentation of meaningful units, as distinct from the segmentation of lower-level phonetic or phonemic units in the speech stream. The term ‘segmentation’, used in this thesis without a modifier, refers to the segmentation of connected speech into lexical units. These units may or may not correspond to orthographic or dictionary words depending on the details of the particular theory being discussed at the time.

Various theories have been proposed describing the nature of the information used to divide the speech stream into lexical units. There have been three main classes of proposal. Firstly specific acoustic markers in the speech stream are used in segmentation (Lehiste, 1972; Nakatani & Dukes, 1977; Nakatani & Schaffer, 1978). Secondly, knowledge of the statistical or distributional structure of lexical items in the language can provide a cue to segmentation; this statistical approach may be applied in different domains, including phonology (Brent & Cartwright, 1996; Cairns, Shillcock, Chater, & Levy, 1997); metrical stress (Cutler & Norris, 1988; Grosjean & Gee, 1987) or prosody (Christophe, Guasti, Nespor, Dupoux and Ooyen, 1997). Thirdly, segmentation is achieved through the identification of lexical items in connected speech (Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986; Norris, 1994).

While there is considerable experimental evidence showing that each of these strategies is effective in lexical segmentation, very little research has attempted to provide a unified account of the effect that different combinations of these strategies have on the segmentation of words in connected speech (for exceptions see Christiansen, Allen & Seidenberg, 1998; Norris, McQueen, Cutler & Butterfield, 1997; Norris, McQueen & Cutler, 1995). One aim of this thesis is to attempt to evaluate these different accounts of segmentation by investigating which mechanisms operate in the case of words whose boundaries are predicted to be ambiguous by one or more of these theories. The words concerned are embedded at the onset of longer words (such as *cap* in *captain*). This thesis uses sentences containing these embedded words to investigate how processes using different sources of information may interact during the segmentation and identification of words in connected speech.

## 1.3. Models of spoken word recognition

Although the work reported in this thesis is intended to address issues of lexical segmentation and word recognition relevant to all current theories, this research relies on a general theoretical framework. This section describes the main assumptions behind this framework and reviews the experimental results that have been used to support these assumptions.

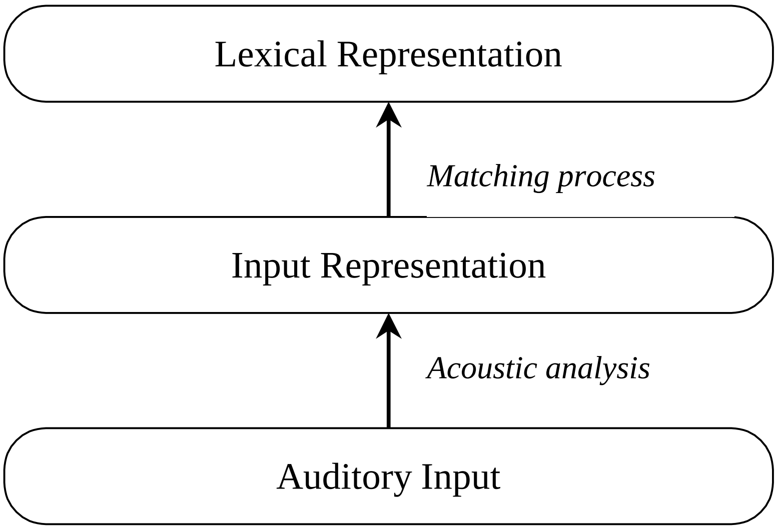


**Figure 1.1: Processing stages involved in auditory lexical access**

The general approach to speech recognition that is taken in the literature involves progressive stages of abstraction, going from a representation of the acoustic signal, to low-level linguistic or perceptual units, to lexically based representations and ultimately to syntactic and semantic properties of the spoken input. In descriptions of this framework, a commonly made distinction is between processes that are involved in deriving an initial representation of the speech input, and later stages involved in accessing lexical representations that match the input representation (see for instance Tyler & Frauenfelder, 1987). In the initial stages, *acoustic analysis* is carried out to construct a form-based representation of the speech signal. Later stages of analysis then involve a *matching process* whereby this input representation is matched to a representation of specific lexical candidates. The goal of this lexical access process is to identify or recognise specific lexical items contained in the speech signal. These processing stages are illustrated in Figure 1.1.

Despite general agreement regarding this processing framework for speech perception, there is still considerable debate and disagreement regarding the nature of the representations that are involved in each of these stages, as well as discussion of how different levels of representation interact during processing. One important issue regards the nature of the pre-lexical representation of the speech input that contacts the mental lexicon.

### 1.3.1. The input representation

Psycholinguistics has inherited from the linguistic tradition an assumption that a string of undifferentiated phonemes (represented as bundles of phonetic features) is an appropriate representation of the speech stream. For a historical perspective on this assumption see Chomsky and Halle (1968). For a more recent discussion of the limitations of this approach from a phonetic perspective see Manaster-Ramer (1996) and Port (1996). Some accounts of spoken word recognition, such as the Neighbourhood Activation Model of Luce and colleagues (Luce et al., 1990; Pisoni & Luce, 1987) assume a phonemically-labelled level of representation as an input to the lexical identification system. Similarly, computational models such as TRACE (McClelland & Elman, 1986) and Shortlist (Norris, 1994) incorporate a phonemically coded representation at a pre-lexical level - although these may be preceded by other lower-level representations of the speech input.

Alternatively, some authors have proposed that the primary input representation of the speech signal is in terms of larger, syllabic units. Evidence supporting this position has come from experiments in French showing that the detection of word fragments is facilitated in words that contain these fragments as syllables compared to words in which these fragments cross a syllable boundary (Mehler et al., 1981). Although this finding has been replicated in other languages such as Spanish, it appears that languages such as Japanese and English are an exception to this pattern. Input representations of Japanese speakers appear to be based around a smaller, moraic unit (Otake, Hatano, Cutler, & Mehler, 1993) while research in English has failed to provide unequivocal evidence of pre-lexical representations organised at any single level (Dupoux & Hammond (submitted), see Pallier, Christophe, & Mehler (1998) for a recent review of this research).

Other accounts of lexical access and speech perception propose a sub-phonemic representation as input to the lexical access process. For instance, Warren and Marslen-

Wilson (1987, 1988) postulate a featural representation of the speech input, with fine-grained sub-phonemic information playing an important role in the recognition process. Experiments showing that mismatching information at a sub-phonemic level can disrupt the recognition process have been cited as evidence to support these accounts (Andruski, Blumstein, & Burton, 1994; Marslen-Wilson, Moss, & van Halen, 1996; Marslen-Wilson & Warren, 1994).

This conflict between accounts proposing different sizes of units as being of primary importance in spoken word processing may reflect a distinction between experimental tasks that tap into perceptual awareness of the form of speech and tasks that infer the structure of pre-lexical representations from the properties of the recognition process. Some theoretical accounts suggest a functional separation between representations involved in the pre-lexical processing of the speech signal and representations that are involved in producing a 'percept' of the speech input (Marslen-Wilson & Warren, 1994). Recent computational models (Gaskell & Marslen-Wilson, 1997; Norris, McQueen & Cutler, in press) consequently provide separate levels of representation for pre-lexical acoustic processing and for the perceptual representations used in phoneme detection and other similar tasks. This separation between acoustic processing and perceptual representation suggests that the units involved in forming a speech percept may not be the same units by which the acoustic signal is processed pre-lexically.

## 1.3.2. The matching process

Many different models have been proposed as accounts of the process by which input representations are mapped onto representations of lexical form. Rather than listing the models themselves, this section will illustrate the theoretical distinctions made by these models. Describing these distinctions is the most productive means of categorising the various models described in the literature.

### *Parallel vs. serial search*

One important distinction made in the literature is between models based on a serial comparison of the input representation with successive lexical items (as in the search model of Forster, 1976; Bradley & Forster, 1987; Forster, 1989) and those based on the parallel comparison of multiple candidates (as originally proposed for visual word recognition in the logogen model of Morton (1969) and extended to spoken word recognition by Marslen-Wilson & Welsh (1978)).

In the serial search model, lexical access occurs via a search through a frequency-ordered list of word candidates. Each candidate is compared in turn to the currently perceived input, with lexical access occurring when a match is found between a lexical candidate and the current input. This ordered search provides a very natural account of effects of word frequency, whereby highly frequent words are accessed more quickly than low frequency words (Rubenstein, Garfield, & Millikan, 1970). However, since this effect is less reliable with spoken than written words (Bradley & Forster, 1987; Marslen-Wilson, 1984; though see Marslen-Wilson, 1990) it is unclear whether frequency effects can be used as evidence for serial search models of spoken word recognition.

In contrast, parallel access models propose that multiple lexical candidates are compared simultaneously, with the activation of any given item indicating the current degree of fit of a lexical hypothesis (cf. Selfridge, 1959). By these accounts, identification involves accumulating sufficient evidence to activate a single lexical item past some threshold level. As a consequence, lexical access is no longer an all-or-nothing process, but may initially involve the simultaneous, partial access of multiple candidates before a single lexical item is recognised. Parallel activation models can also account for frequency effects, either through postulating differences in the resting activation of the representational units for words with different frequencies (McClelland & Rumelhart, 1981; Morton, 1969) or through stronger connections to units involved in the representation of higher frequency words (Plunkett & Marchman, 1991; Seidenberg & McClelland, 1989).

One advantage of parallel access over serial search models is that they offer a simple account of the activation of multiple lexical candidates during spoken word recognition. For instance, Zwitserlood (1989) showed that during the auditory presentation of a fragment of the Dutch word *kapitein* (captain), significant facilitation could be observed for words that are semantically and associatively related to a competing word *kapitaal* (capital) that also matched the fragment. This transient activation of multiple lexical items (and associated meanings) is readily accommodated in models proposing that the speech input activates all the candidates that match the initial spoken input, with selection processes operating to narrow down the set of activated candidates to those that continue to match the speech input (Cohort model – Marslen-Wilson & Welsh, 1978; TRACE – McClelland & Elman, 1986; Shortlist – Norris, 1994).

Although these results may be simulated in a serial system by initiating searches at regular intervals during the speech stream and accessing the meaning of matching candidates after each search, this simulation of parallel effects is only achieved through reducing the role of serial search to a minimum, and making the recognition process functionally equivalent to a parallel access account. The research carried out in this thesis will be based around a parallel-access, activation-based account of the word recognition process.

A further implication of the results presented by Zwitserlood (1989), as well as other results used to motivate the Cohort model (Marslen-Wilson & Welsh, 1978) is that the spoken word processing system tracks the speech input, continuously updating the activation of different representations in the light of new input. Consequently these data are beyond the scope of spoken word recognition models which do not make explicit predictions for processes that occur during the time-course of the speech signal, such as the Neighbourhood Activation Model of Luce and colleagues (Luce et al., 1990). This, and other models that do not adequately capture the temporal processing of the speech input will not be considered in detail in this thesis.

### ***Autonomy, interaction and competition***

Models using an activation metaphor to represent the goodness of fit between currently available evidence and lexical items provide a coherent account of the matching process. However, it remains unclear what sources of evidence are evaluated in this way during lexical access. One commonly made distinction in the literature is between accounts whereby only information from lower levels affects the activation of lexical candidates – *autonomous* models, such as those proposed by Forster (1976) and Norris (1994) – and interactive models in which information from higher levels (such as from syntactic or semantic constraint) can also influence the recognition system (such as in TRACE - McClelland & Elman, 1986).

In a parallel activation account, the terms autonomous and interactive can be interpreted as constraints on the direction of connectivity and types of connections that can be made to and from lexical units. More specifically, an autonomous model is one in which lexical units only receive input from units at lower levels. An interactive model is one in which these lexical units may also receive feedback from higher levels of representation such as syntax or semantics.

However, with the advent of parallel distributed processing models in which connection strengths are acquired through the application of a gradient descent learning algorithm (Rumelhart, Hinton & Williams, 1986), this categorical distinction between autonomous and interactive models may not provide the best characterisation of different styles of computation. For instance, in many computational simulations the notion of a distinctly lexical level of representation becomes blurred. For example in the triangle framework of Plaut and colleagues (Plaut, McClelland, Seidenberg, & Patterson, 1996) distributed representations of orthography, phonology and semantics form a lexical system through the connectivity that exists between these levels. The nature of the processing interactions that develop between different representations depends on the regularities that exist between different domains and only indirectly on the assumptions made by the modeller (see for instance Harm & Seidenberg, 1999).

These terms are similarly problematic when used in interpreting the results of behavioural experiments. For instance, the Ganong effect whereby ambiguous phonemes are categorised according to the lexical status of the string in which they are contained has been interpreted as evidence for interactive effects on phonetic categorisation (Ganong, 1980). Although this influence of lexical information on phonetic categorisation may be simulated through top-down interactions between lexical and pre-lexical information (in TRACE, for example), these results need not imply top-down connectivity but only that participants’ responses are made from a level of representation that includes lexical influences. Elman and McClelland (1988) showed that phonemes disambiguated by lexical context influence categorisation of subsequent phonemes in the same way as would be expected for unambiguous phonemes. Thus the Ganong effect can also affect mechanisms involved in compensation for co-articulation. Although their results were initially interpreted as evidence for top-down, lexical influences on phonemic processing, subsequent experiments (Pitt & McQueen, 1998) and simulations (Cairns, Shillcock, Chater, & Levy, 1995; Norris, 1993) show that transitional probabilities between phonemes provide a non-lexical account for this effect. Consequently, in order to describe a particular pattern of experimental results as indicating top-down influences or interactive processes it must be demonstrated that no alternative, non-interactive account of the results must be possible.

Another term commonly used to describe models of speech perception is *competition*. This term typically describes models in which the activation of any given lexical candidate is affected by the presence or absence of other activated candidates. For example, if two or more lexical candidates match the current input, a model incorporating competition will activate each candidate to a lesser degree than if only a single candidate matches the input. In localist connectionist models, the presence of competition is commonly modelled as inhibitory connections between jointly activated units (e.g. *captain* and *captive* which would both be activated by the speech input /kæptɪ/)

However, just as caution is advised in the use of the words ‘autonomous’ and ‘interactive’, similar caution should be exercised in using the word ‘competition’. Behavioural data described as indicating competition need not be simulated using direct inhibitory connections and consequently may not only support computational models that include direct competition between lexical units. For instance, recurrent neural networks trained by back-propagation produce output activations that are dependent on the number of simultaneously-activated candidates by representing the probability of all outputs given the current input. Such behaviour would commonly be assumed to reflect competition between output units, yet these recurrent networks do not incorporate direct inhibitory connections between units within a processing level (e.g. Gaskell & Marslen-Wilson, 1997; in press).

It would therefore appear that terms such as autonomy, interaction and competition that are used to relate behavioural data to theoretical accounts are best used sparingly in the absence of implemented computational systems based on these theories. Processing distinctions made on behavioural grounds may not be as effective in distinguishing between different models as previously considered. Throughout this thesis care will therefore be taken to describe precisely the behavioural evidence from experiments using terms that are independent of their simulation in computational models.

### 1.3.3. Lexical representations

Traditional assumptions about the role of lexical representations viewed these as a bridge between the form of words and their meaning. Consequently, accounts of lexical access in spoken word recognition considered the goal of the process to be to activate a representation of the meaning of lexical items or words in the speech stream. This focus on access to the meanings of words is apparent in work investigating effects of preceding sentential context on the meanings activated in response to words such as *bank* which have multiple meanings (Simpson, 1984; Swinney, Onifer, Prather & Hirshkowitz, 1979).

Recent accounts of sentence processing have been proposed that draw parallels between the resolution of lexical ambiguities (such as those created by homophones) with syntactic ambiguities such as "*the spy saw the cop with the binoculars*" (MacDonald, Perlmutter, & Seidenberg, 1994). The debate between these statistical (constraint-satisfaction) and syntactic (garden-path) accounts is still far from resolved (Frazier & Clifton, 1996). However, both classes of accounts propose that lexically represented information plays an important role in guiding the parsing process. Consequently, lexical access not only activates representations of semantic and conceptual knowledge, but also accesses information about the syntactic constructions in which a word is used.

## 1.4. Computational modelling

As was illustrated in the preceding discussion, the implementation of computational models is vital for determining whether or not a descriptive model is capable of accounting for a particular set of experimental data. Consequently this thesis combines empirical investigations of spoken word recognition with computational simulations of the time course of identification of words in connected speech. In order to implement a computational model it is necessary to specify every component and assumption that is associated with a theory. By making theories explicit in a working system, insights can be gained into the nature of the tasks undertaken by the language processor. It is also possible to test whether current experimental data can be accounted for by models derived from these theories.

Computational accounts of lexical segmentation and spoken word recognition come in many forms. One important distinction is between models that are implemented using symbolic algorithms (INCDROP - Brent, 1997; PARSER - Perruchet & Vinter, 1998 and MK10 - Wolff, 1977) and connectionist models implemented using networks of simple, neuron-like processing units (Distributed Cohort Model - Gaskell & Marslen-Wilson, 1997; TRACE - McClelland & Elman, 1986; and Shortlist - Norris, 1994). Both symbolic and connectionist models provide numerous advantages over conventional, descriptive theorising, but there are also many differences between them, perhaps the most important of which is in the style of processing they assume.

Conventional computer languages impose strict constraints on the manner and order in which operations can occur on input representations. Computation occurs through a series of discrete, serial steps involving transformations and calculations operating on abstract, symbolic representations. In contrast, connectionist models rely on a parallel process mapping from one representation to another through the operation of simple, distributed processing elements (Rumelhart & McClelland, 1986a; McClelland & Rumelhart, 1986). Although each style of computation could be considered theoretically neutral (after all, both types of system are Turing equivalent and are implemented on serial computers), in practice the two types of model lend themselves most naturally to different types of process and different types of explanation. Connectionist models view cognition as involving similarity-based processes of generalisation, while symbolic models operate through mechanisms of categorisation and  rule application.

This distinction between symbolic and connectionist computation is most clearly apparent in the debate on supposed differences in the representation and processing of regular and irregular forms of the English past tense. Symbolic computational systems require separate processes for regular and irregular verbs (Pinker & Prince, 1988; Prasada & Pinker, 1993) whereas connectionist accounts propose that both types of verbs are processed within a single system (Rumelhart & McClelland, 1986b; Plunkett & Marchman, 1991; Plunkett & Marchman, 1993). Although this debate is too lengthy to summarise here, it is becoming increasingly apparent that it will not be resolved either by behavioural data (Ullman et al., 1997) or by computational simulations alone (Joanisse & Seidenberg, 1999). Converging evidence from multiple sources is therefore required to constrain theorising. This multi-disciplinary combination of computational modelling and experimental investigation is utilised in the research reported in this thesis.

The computational simulations carried out here have used connectionist networks exclusively. Although there is nothing intrinsic to these simulations that precludes the use of symbolic algorithms, a list of the desirable computational properties of models of spoken word recognition suggest that connectionist architectures are more parsimonious. For instance, these models are required to account for effects of partial information, to display probabilistic operation, and to be robust in the face of noisy input. Furthermore an important property of the word recognition system is that it be capable of acquiring new vocabulary throughout childhood and to retain this knowledge in the face of progressive damage with ageing. Many of these desirable properties follow naturally from the use of a connectionist model.

One further debate that is mentioned here only in passing, is the distinction between connectionist models built exclusively using localist representations and those that use distributed representations as well. Despite the prevalence of arguments in favour of distributed systems in the psychological literature, it is clear that many useful computational properties can be gained by incorporating localist representations (Page, in press), though possibly at the cost of a certain amount of neural plausibility. Although the models reported in this thesis use a localist representation of phonetic features at the input level and a localist lexical representation at the output, they develop distributed internal representations through training. As will be discussed in Chapter 3, this use of localist input and output representations in a system trained to produce distributed internal representations is chosen for convenience of implementation only, and not through any ideological commitment to either localist or distributed models.

## 1.5. Overview of thesis

This review of the theoretical background provides a framework with which to describe the work carried out as part of this thesis. Since the research approaches segmentation from both an experimental and computational perspective, the review of lexical segmentation in Chapter 2 provides a theoretical introduction without including excessively detailed accounts of either computational or experimental investigations. These will be reserved for more focused reviews at the start of the relevant chapters: Chapter 3 for computational modelling of lexical segmentation and identification and Chapter 4 for experimental investigations of segmentation and lexical access.

The review of lexical segmentation in Chapter 2 outlines an apparent conflict between processes of segmentation that have been hypothesised to operate at different levels of representation of the speech stream. It is argued that onset-embedded words provide a test-case for resolving this conflict since they are predicted to be temporarily ambiguous by several accounts of lexical segmentation. The main body of the thesis will focus on investigating the recognition of onset-embedded words with a view to resolving the conflict between acoustic and lexical accounts of segmentation. Chapter 2 concludes by reporting dictionary searches establishing the extent of the problems created by embedded words within the morphologically decomposed lexicon proposed by Marslen-Wilson, Tyler, Waksler and Older (1994).

Chapter 3 starts with an introductory review of computational models of lexical identification. This review focuses on a particular problem for models of spoken word recognition regarding the identification of onset-embedded words. The apparent need to delay identification of these words until after their acoustic offset has been used to motivate models of spoken word recognition that incorporate inhibitory competition between lexical units. However, simulations reported in this chapter describe how a simple recurrent network can learn to identify embedded words without explicitly implemented competition. The relationship between this model and accounts of vocabulary acquisition is discussed, in particular looking at the relationship between processes involved in learning the statistical structure of speech and processes involved in extracting meaning from sequences of sounds.

Since the recurrent network models reported in Chapter 3 predict a different time course of identification for embedded words than previous, competition-based models, Chapter 4 begins by describing the results of previous experiments that might decide between models with inhibitory competition between lexical units and the recurrent networks described in Chapter 3. Given the lack of appropriately constructed experiments in the literature, Chapter 4 then describes the development of stimuli for experiments with the potential to test these two conflicting accounts of the identification of onset-embedded words. Since computational models of lexical segmentation make strong and possibly unjustified assumptions about the nature of the speech input, the experimental stimuli developed for this experiment are subjected to detailed acoustic analyses to ensure that investigations are based on an accurate description of the acoustic properties of the speech stream. The importance of this analysis is confirmed by the results of a gating study suggesting that acoustic differences between embedded words and longer competitors can be used by the perceptual system. However, caveats regarding the role of response biases in the gating task limit the interpretations of these results with respect to the computational models described previously.

Chapter 5 reports the results of a series of cross-modal priming experiments carried out on the stimuli described in Chapter 4. Using the magnitude of repetition priming as a measure of lexical activation it is shown that acoustic cues allow the perceptual system to differentiate onset-embedded words from longer competitors even before the offset of the embedded word. Despite these acoustic cues, priming experiments tracking the activation of competing interpretations show that longer competitors of onset-embedded words remain active in contexts designed to produce ambiguity for these words.

Since there is now experimental evidence supporting the role of acoustic cues to word length in the identification of onset-embedded words, Chapter 6 describes modifications to the previously developed recurrent network model to incorporate input cues analogous to those hypothesised to be responsible for the discrimination of onset-embedded words. Two sets of simulations are reported, exploring the effect of input cues that require adaptive processing of the preceding spoken context in order to be utilised effectively. The results of these simulations, within the limits created by the small scale of the network, show a similar time course of identification for onset-embedded words and longer competitors as the priming experiments reported in Chapter 5.

In Chapter 7 a further prediction of the model regarding the role of following context in the identification of onset-embedded words is tested. The models reported in Chapter 6 predict that information after the offset of an embedded word plays an important role in ruling out longer competitors, thereby supporting the recognition of an embedded word. One further gating experiment and two cross-modal priming experiments tested this prediction using a set of stimuli derived from those that were used in the initial series of experiments. Results of these experiments suggest that the activation of onset-embedded words appears to be unaffected by the presence or absence of phonological mismatch with the longer words in which they are embedded. Comparisons are made between this behavioural profile and the predictions of the recurrent network account of word recognition presented in the previous chapters. Finally, in Chapter 8, conclusions are drawn from the experimental and computational work presented in this thesis. Future work is also proposed to extend and test the model developed in this thesis.

# 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 [sliːp] 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 monosyllabic (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.

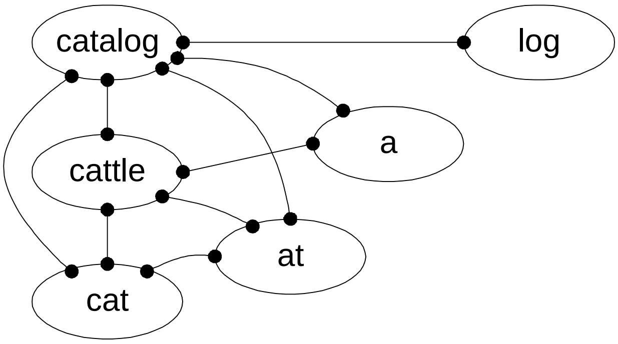


**Figure 2.1:Pattern of inhibitory connections between a subset of candidates activated by the presentation of /kætælog/ in Shortlist. Adapted from Norris (1994), Figure 2.**

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 <u>beneficial</u> 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.

| | | P(embed.) | Carrier words | P(embed.) | Carrier words | P(embed.) |
|---|---|---|---|---|---|---|
| Syllables in carrier word | Syllables in embedded word | LDOCE McQueen (1995) | CELEX | CELEX + Phon. | CELEX & LDOCE | 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.

# 3. Connectionist models of spoken word recognition

Since the seminal work presented in the two Parallel Distributed Processing volumes that are most responsible for reintroducing connectionist modelling to Cognitive Science (Rumelhart & McClelland, 1986a; McClelland & Rumelhart, 1986) computational implementations of psychological theories have increasingly used artificial neural networks. Although connectionist models are almost universally simulated on serial, symbolic computers, it has been claimed that the computational properties of connectionist models provide a theoretical framework for cognition that goes beyond the re-implementation of traditional symbol-and-rule based processing accounts (see Fodor & Pylyshyn, 1988; Smolensky, 1988; Clark, 1993 for further discussion). In modelling the perception of spoken language, the power of neural networks to account for the processing of noisy and probabilistic information makes them unrivalled as psychological models of perceptual processing (see Bishop (1995) for a more thorough discussion of probabilistic interpretations of neural networks and Robinson (1994) for an application of neural networks in speech processing).

A further advantage of the connectionist approach is that the use of neural network learning algorithms provides a means by which to incorporate insights from development into cognitive theorising (see for instance Plunkett & Sinha, 1992; Elman et al., 1996; Quartz & Sejnowski, 1997). Although the use of gradient descent learning algorithms significantly increases the complexity of the resulting network (since the solutions obtained using these algorithms are often computationally opaque) a valuable constraint is placed on the modeller by the requirement that systems make explicit assumptions about the developmental process. Since the modeller is required to specify what information is available to the network prior to and during acquisition, developmental connectionism allows investigation of which properties of the fully trained model depend on the structure of the system and which depend on the environment to which it is exposed during training. Behavioural data from language learners can then provide an empirical test of implemented models. This interplay between modelling and empirical data is best illustrated in the literature on the inflectional morphology of the English past tense (Rumelhart & McClelland, 1986b; Pinker & Prince, 1988; Plunkett & Marchman, 1991;

Plunkett & Marchman, 1993) however such an approach is likely to be equally informative in the literature on spoken word recognition and language acquisition.

This chapter begins by reviewing the existing literature on connectionist models of spoken word recognition, contrasting two distinct styles of model, models incorporating direct competition between localist lexical representations, and distributed systems in which effects of competition emerge as a consequence of the probabilistic behaviour of a recurrent neural network. As discussed in the previous chapter, distributed systems have been suggested to be of limited value in accounting for the identification of words embedded at the onset of longer words. In the current chapter, a recurrent network model is investigated which not only resolves this difficulty, but also incorporates potentially more realistic assumptions regarding the nature of the problem solved during lexical acquisition. The relationship between this  account and other developmental theories of lexical segmentation is explored in more detail in a set of simulations investigating the relationship between lexical and distributional learning of segmentation.

## 3.1. The TRACE model

Arguably the most influential connectionist account of spoken word recognition is the TRACE model (McClelland & Elman, 1986). This influence is illustrated by the fact that it is still used by researchers to provide an accurate fit to novel psycholinguistic data more than 10 years after its development (Allopenna, Magnuson, & Tanenhaus, 1998). However, as will be described subsequently, the architecture of TRACE is complex and reliant on hard-wired connections. Consequently TRACE may not be amenable to the developmental approach to connectionist modelling that is proposed here. The model is also unable to account for some recent experimental data suggesting a role for bottom-up inhibitory processes in lexical access.

The architecture of TRACE emerged from applying the structure of the interactive activation and competition (IAC) model of visual word recognition (McClelland & Rumelhart, 1981) to the auditory domain. In keeping with the previous IAC model, TRACE consists of 3 levels of units representing phonetic features, phonemes and words, analogous to letter features, letters and words in IAC. In TRACE, as in other localist models, each hypothesis as to the identity of a feature, phoneme or word in a section of speech is represented by the activation of a single unit. Connections between mutually consistent units in different levels are bi-directional and excitatory, while connections between units within a level are mutually inhibitory. At the lexical level this allows TRACE to resolve the conflict between lexical items that share segments, ensuring the activation of words that make up a consistent lexical segmentation of the speech stream. That is, given a sequence of speech as input, TRACE will activate lexical units so as to account for all the segments in the input without inappropriately assigning the same segment to multiple lexical items.

One important difference between TRACE and IAC (reflecting the obvious difference between speech and text) is that while different units in IAC represent spatially separated information (distinct letters or words for example) units in TRACE represent temporally distinct information occurring sequentially in the speech stream. The time dimension is represented in TRACE by parallel duplication of units representing features, phonemes and words at each time step. Thus there will be a separate representation for the same linguistic unit occurring at different times in the input. In processing temporally extended information, units accumulate information represented at previous time steps in processing. By combining activations over adjacent time steps, TRACE is not restricted to a strictly sequential, left-to-right recognition process. This proves crucial in allowing the model to use following context in the recognition of onset-embedded words.

However, this spatial representation of temporal information has been criticised. One consequence of the TRACE architecture is that the entire network must be duplicated across different time steps. It is suggested that the number of units and connections required for a realistically sized lexicon is unfeasible and hence that TRACE is too inefficient to be a plausible account of spoken word recognition (Norris, 1994). Similarly it is argued that by representing temporal information spatially (i.e. using different units to represent the same event occurring at different points in time) the TRACE model prevents a form of generalisation that is argued to be vital in the processing of spoken language – namely that the same linguistic units (words, phonemes or features) can be reused at different points during a sequence (Port, 1990). TRACE has to enforce this generalisation by weight sharing between units at different points in time.

Nonetheless, it is unclear what criteria we are to use in determining whether any given model can be implemented in neural hardware and with an adult-sized lexicon. Although duplicated units and weight sharing may be limitations of a particular implementation, since it may be possible to implement the same computational assumptions within a more realistic architecture it is unlikely that these difficulties alone are sufficient to rule out the TRACE account. The TRACE model should therefore stand or fall on its account of psycholinguistic data on word recognition.

### *Psycholinguistic data on word recognition*

The time course of spoken word recognition as modelled by TRACE captures many of the important aspects of the experimental data used to support the Cohort model of Marslen-Wilson and Welsh (1978)[1]. At the onset of a word both Cohort and TRACE predict that many lexical items are activated in parallel with candidates dropping out of this activated set when mismatch occurs in the speech stream. Thus TRACE provides an account of behavioural data showing sensitivity to the point at which the speech stream uniquely specifies a single lexical item - the uniqueness point (Marslen-Wilson, 1984). However the mechanisms by which TRACE produces this activation profile differ from those envisioned in the Cohort theory. McClelland and Elman do not incorporate bottom-up inhibitory connections between phoneme units and lexical units representing words that do not contain those phonemes. Instead TRACE uses inhibitory connections at the lexical level to rule out potential candidate words. For this reason, the model is only able to reduce the activation of a lexical item following mismatch where there are other more active units that can provide inhibition at the lexical level.

Consequently TRACE predicts that in cases where input that mismatches with an activated candidate is presented, there should be little or no decrease in lexical activation if the input does not match an alternative lexical item. This prediction was not borne out by experiments reported by Marslen-Wilson & Gaskell (1992; described in more detail in Gaskell, 1994) since mismatch that creates a non-word (e.g. *sausin* mismatching with *sausage*) reduces priming to an associatively related target as effectively as mismatch that produces an alternative word (*cabin* versus *cabbage*). Gaskell (1994) shows that the standard TRACE model is unable to account for this data and suggests that accounts incorporating bottom-up inhibition (such that the mismatching final segment of *sausin*

[1] Indeed early versions of the TRACE model were called Cohort

reduces the activation of *sausage*) may provide a better account of effects of mismatch on lexical activation than models that rely solely on intra-lexical competition.

### *Competition and lexical segmentation*

Despite this experimental evidence, direct lexical-level competition between activated word units is suggested as a necessary property of TRACE for other reasons. In particular, inhibitory connections between lexical units that span potential word boundaries allow TRACE to use lexical competition to segment the speech stream into words. This is of particular importance in allowing TRACE to recognise onset-embedded words such as *cap* embedded in *captain* (McQueen et al., 1995).

As described in the previous chapter, the recognition of these onset-embedded words requires that longer lexical items that are 'carriers' of the embedded words can be ruled out. Since word boundaries are seldom explicitly marked in connected speech, ruling out longer competitors may not be straightforward. Simulations with TRACE (Frauenfelder & Peeters, 1990) show how mismatching input combined with lexical competition provides a mechanism for the identification of onset-embedded words. To take an example sequence "*cap fits*" , information coming after the offset of the embedded word*cap*, is inconsistent with longer competitors (such as *captain*). This mismatch will reduce the activation of units representing longer words (through the activation of other lexical items). The decreased competition from carrier words will then allow the identification of an embedded word.

For the example sequences used by Frauenfelder and Peeters, it was shown that the lexical competition network in TRACE allows mismatching input to facilitate the recognition of embedded words, even where mismatch may be delayed with respect to a word boundary (such as for the sequence *cap tucked*, where the start of the following word matches the longer competitor *captain*). Thus lexical competition not only provides a means of ruling out mismatching input within a word – as in the case of cohort-competitors like *cabin* and *cabbage* that share the same onset – but also allows the use of mismatch after a word boundary in the identification of embedded words.

In identifying onset-embedded words and longer competitors, TRACE predicts a distinct time course of activation. Since the magnitude of competition between lexical units is dependent on the number of other items that share constituent phonemes, longer words will have more lexical competitors. Thus long words such as *captain* which have inhibitory connections to and from words that are aligned with the second syllable (*tin*, *tinsel*, etc.) will receive greater overall inhibition than short words such as *cap*. For this reason, embedded words will be more active than longer competitors during early stages of processing (i.e. at the offset of /kæp/, the lexical unit for *cap*, will be more active than the unit for *captain*). This bias towards short words supports the identification of embedded words since it provides them with the additional activation required for them to win out in competition with longer lexical items.

In recurrent network simulations described in this chapter, it will be shown that models that incorporate bottom-up inhibition do not require this short word bias to identify onset-embedded words. Mismatching input can be used to rule out longer lexical items without requiring that embedded words are initially more active. However, in models lacking bottom-up inhibition the only means by which lexical items can be ruled out is through competition at the lexical level. Thus TRACE requires a short word bias so that lexical competition can be used to identify onset-embedded words. Some models such as Shortlist (Norris, 1994) provide for both mechanisms – bottom-up inhibition and lexical competition. Nonetheless, in simulations that incorporate inhibitory connections between lexical items a short word bias will still be observed in the model. As will be discussed in subsequent chapters, this discrepancy between models that incorporate lexical competition and those employing only bottom-up inhibition may provide a tool with which to falsify competition based accounts of spoken word recognition.

### *Discussion*

The TRACE model has proved successful in accounting for a large body of data on the time course of spoken word recognition. However, in the years since the original development of the model an increasing amount of experimental evidence from word recognition (Marslen-Wilson and Gaskell, 1992) and phoneme detection (Frauenfelder, Segui, & Dijkstra, 1990) have challenged the interactive activation and competition assumptions of the TRACE model. Furthermore recurrent neural networks have shown how systems can not only account for the processing of temporally structured input, but also suggest an account of how systems can learn the sequential structure of a domain from exposure to an appropriate training set. Given the relevance of vocabulary acquisition in accounts of spoken word recognition and the importance of developmental evidence in evaluating computational models, it is perhaps unsurprising that alternatives to the hard-wired connections used in the current implementation of TRACE have been sought.

However, despite the fact that supervised learning rules can operate as effectively in a localist model as in distributed networks (see Page, in press for further discussion) other aspects of the architecture of TRACE preclude modification to incorporate data from acquisition. One problem is that inhibitory connections between lexical items are hard-wired, with connection strengths that depend on the overlap between the phonologies of competing words. There is therefore no simple way of adding lexical units for new words without making substantial alterations to the lexical competition network. This problem is compounded by the duplication of units and connections at each time slice. For this reason it is unclear how TRACE could learn novel words and generalise newly learnt vocabulary to other time steps in the input. These acquisition issues can be resolved through the use of recurrent neural networks for spoken word recognition.

## 3.2. Recurrent network accounts

As discussed in the introductory section of this chapter in connection with the TRACE (see also Elman, 1990; Port, 1990), the spatial representation of temporal information fails to preserve the similarity between two identical patterns presented at different points in time. For example consider the following two input vectors (adapted from Elman, 1990):

[ 0 0 0 **1 1 1** 0 0 0 ]    (1)

[ 0 0 0 0 0 0 **1 1 1** ]    (2)

**→ time →**

As can be seen, sequences (1) and (2) contain two identical patterns displaced in time. However, to a network which represents this temporal displacement spatially (where each segment in the sequence is represented at a separate unit) there would be no similarity between the two vectors. Alternative approaches such as the ‘moving window’ input used in NET-Talk (Sejnowski & Rosenberg, 1987) are able to resolve this problem by having an input window slide over the vectors one step at a time. Thus the sequences in (1) and (2) would be processed when they are centred in the input space, preserving their similarity.

However the moving window approach has limitations, one of which is illustrated in the more complex set of vectors below (taken from Abu-Bakar and Chater, 1995):

[ A B C - - - ] (3)

[ A A B B C C ] (4)

[ A A B - - - ] (5)

**→ time →**

Here we have a set of three sequences in which the duration of each segment in the input changes in proportion to the overall rate at which the vector is presented. This can be considered analogous to one of the problems found in speech recognition since the absolute duration of segments and syllables will vary with the overall rate at which an utterance is produced (Crystal & House, 1990). A system able to process these time-warped sequences should be able to recognise that sequence (4) is identical to sequence (3) but produced at a slower rate, ignoring the greater overlap between it and sequence (5) which it more superficially resembles. Such a problem cannot be resolved in a moving-window approach since the network does not represent duration information dynamically and is therefore unable to correct for rate of presentation in order to recognise the similarity between sequences (3) and (4).

One type of connectionist architecture that is better able to process these time-warped sequences is a recurrent or simple recurrent network (Abu-Bakar & Chater, 1995; Elman, 1990; Norris, 1990; Port, 1990). These networks represent sequential information in a temporal fashion – as a sequence of vectors represented at the same set of input units at different points in time. Recurrent connections (often just at the hidden units, but potentially at all sets of units) allow the network to use a representation of states at previous time steps in interpreting the current input. When training these networks, information is preserved from the previous time step only (for a simple recurrent network or SRN; Elman, 1990) or over many time steps (fully recurrent networks; Rumelhart, Hinton and Williams, 1986). SRNs can be trained using straight back-propagation whereas fully recurrent networks require training using back-propagation through time, with weight sharing to ensure that weight changes remain identical across each unfolded time step. Given the greater computational cost involved in simulating fully recurrent networks, the majority of the work reviewed here and all the simulations reported in this thesis will use the computationally cheaper simple recurrent network architecture illustrated in Figure 3.1 below.

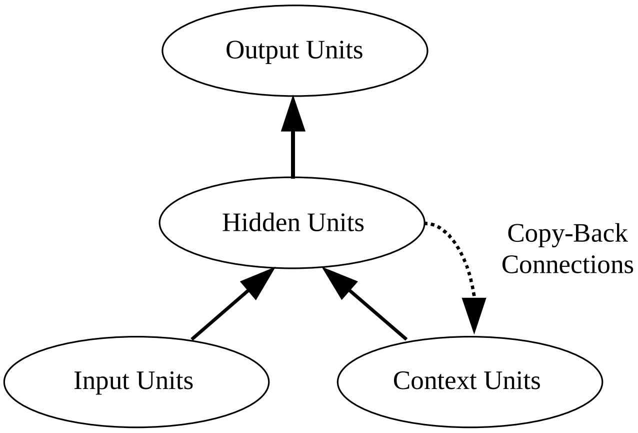


**Figure 3.1: A simple recurrent network (Elman, 1990). Solid arrows represent trainable connections, broken arrows show hidden-unit activations from the previous time step, copied back to the context units on a one-to-one basis.**

### 3.2.1. Prediction tasks and lexical identification

As described in the review of lexical segmentation in Chapter 2, one influential simulation investigated the computational properties of simple recurrent networks trained to predict the next input in a stream of speech segments (Elman, 1990). The novel result from these simulations is that in carrying out this prediction task the network displays sensitivity to the structure of lexical items in the training set. Elman reports that prediction error drops as the network is presented with more of a word and rises sharply at the offset of each word. As described in Chapter 2 this sharp rise in error could be used as a cue to the location of a word boundary. This section reviews whether these simulations also have the potential to provide an account of lexical identification.

The decrease in prediction error towards the ends of words could be considered analogous to cohort effects, in which information accumulates over time until a word can be uniquely identified (Marslen-Wilson and Welsh, 1978; Marslen-Wilson, 1984). At the uniqueness point of a word, Elman’s network could potentially predict the next segment with zero error. In this circumstance the network has passed through a sequence of states that are unique to a single lexical item. However the network's internal representation at this point is not equivalent to a lexical representation for that word. Since the task the network is doing only requires prediction of subsequent input, not the storage of prior input, the network need not uniquely represent each lexical item even in cases where prediction error is zero[2].

Furthermore in line with early incarnations of the cohort model this account has problems with onset embedded words (for the orthographic input used in the original Elman (1990) simulation these were "*the*" and "*they*" ). Any attempt to use the prediction task to uniquely represent these items will be in vain. Where embedded words have an identical input representation to the onset of longer competitors the system will not be able to distinguish embedded words from competitors until a following context has been presented (c.f. delayed recognition in TRACE), at which point subsequent words are the focus of the network's output.

This problem with onset-embedded words may go some way towards explaining why lexical effects observed in the small-scale Elman simulations do not scale to realistically sized training sets. As described by Cairns et al. (1997) and Christiansen et al. (1998), networks carrying out the prediction task do not reduce output error following the uniqueness point of a word when trained on corpora with large numbers of lexical items. Thus the lexical effects obtained in previous small scale simulations do not scale up to more realistic training sets. Although the phonotactic regularities extracted by these networks are valuable in detecting word boundaries (see the discussion of these distributional accounts of lexical segmentation in Chapter 2) they are not equivalent to lexical identification. This review will therefore focus on recurrent network simulations of tasks that require explicit lexical identification.

[2] Consider the pair of words *transmission* and *remission*. After the uniqueness point of these words, a network may well be able to predict the next segment without error. However, since the task required of the network is predicting future input it is not necessary for the system to produce a different representation depending on the onset of the word. Such a representation would be required if the network is to distinguish one word from the other.

### 3.2.2. Modelling word recognition in connected speech

A connectionist view of the task of recognising words in connected speech consists of a mapping from a representation of the speech input to a lexical and/or semantic representation of the words contained in the speech stream. In producing models of this mapping, different assumptions have been made regarding the properties of the target representation, as well as using different recurrent network architectures to perform the mapping.

Norris (1990) reported simulations using simple recurrent networks to investigate spoken word recognition. He trained a network to map a sequence of speech segments[3] onto a localist representation of the identity of the current word (i.e. activating one node out of a set of units each representing a single word in the network's vocabulary). The results of this simulation capture the left-to-right properties of the cohort model very naturally. At each segment in the input, the network activates an output representation indicating the identity of the current word in the speech stream. In cases of ambiguity, several lexical units will be activated in proportion to the likelihood that they represent the current word in the speech stream. For example, before the uniqueness point of the pair of words *delimit* and *deliver* (*v* in *deliver, m* in *delimit)* each word is activated equally. As soon as input is received that allows discrimination of the two words, the inappropriate word is deactivated and the appropriate lexical unit becomes fully activated.

This simulation of cohort effects arises as a consequence of the probabilistic nature of processing in the network and its training regime. As demonstrated by Servan-Schreiber, Cleeremans, & McClelland (1991) for a simple recurrent network in which a single unit is active in each target representation, output activations represent the conditional probability (given the current input) of each output unit being active in the training set. Hence, where four items of equal frequency match the input, units representing each item will be activated to 0.25, three matching items would be activated to 0.33, and so on. This

[3] Norris in fact used a representation of the letters in each word as the input to the network. Equivalent performance would have been observed had a phonemically coded input representation been used instead.

probabilistic account of cohort effects is a great strength of SRN models of spoken word recognition.

However as discussed by Norris (1990, 1994), this model of spoken word recognition is not without its problems. Most importantly, the network will be unable to recognise onset-embedded words such as *cap*. Since input to the model will be identical for an embedded word and the start of a longer competitor, embedded words will not become unique until after their offset – when mismatch between the following context and longer words can rule out all lexical items other than the embedded word. However at the point where all other words are ruled out, the network will no longer be activating the identity of the embedded word at the output and will instead be attempting to identify subsequent words in the input. Therefore at no point in the speech stream will the network uniquely identify (and hence fully activate) onset-embedded words.

This failing of the Norris (1990) SRN model directly parallels the limitations of sequential recognition accounts of lexical segmentation discussed in the previous chapter. The approach used by Norris (1994) to resolve this problem is to add the lexical competition mechanism that allowed TRACE to recognise onset-embedded words. Norris uses the activated lexical units from a recurrent network as a 'Shortlist' of potential candidates for a lexical competition network[4] in which mutual inhibition allows the selection of the candidate (or candidates) that best match the speech input. Thus Shortlist divides the lexical access process into two stages (bottom-up activation of lexical candidates, followed by competition between these candidates). More recent versions of Shortlist increase the separation between these two processing stages by 'resetting' the activation of the words in the current Shortlist after each segment has been presented at the input. Recent implementations have also incorporated a variety of different distributional cues (metrical stress and phonotactics) through the addition of penalty terms in the competition stage for lexical hypotheses that violate these different constraints (Norris, McQueen, & Cutler, 1995; Norris, McQueen, Cutler, & Butterfield, 1997).

[4] In fact, rather than implementing a large recurrent network model, Norris simulates the output of an idealised recurrent network by repeated searches of a lexical database for words after the presentation of each input segment.

This more complex model has proved successful in accounting for a wide-range of psycholinguistic data – though at the expense of making the behaviour of Shortlist for any given input rather hard to predict. Since the network includes two different mechanisms for ruling out mismatching candidates (through mismatch in the bottom-up activation stage or through lexical competition in the Shortlist) the computational cause of any behaviour produced by the network may be unclear. It is therefore important to evaluate the contribution and possible performance of each component of Shortlist separately. This thesis will focus on the computational properties of the recurrent network component of the Shortlist model.

Some recent work by Gaskell & Marslen-Wilson (1997) proposed a model of speech perception based on a recurrent network trained to activate a distributed representation of both lexical form and semantics. The simulations reported by Gaskell and Marslen-Wilson used a simple recurrent network trained to map a stream of phonetic feature information to a distributed lexical/semantic representation and a phonological representation of the current word. This 'Distributed Cohort' model provides a reasonably accurate simulation of the results of experiments comparing phoneme detection and lexical decision for cross-spliced tokens of words and non-words (Marslen-Wilson & Warren, 1994; see Norris, McQueen and Cutler, in press for further discussion). Lexical influences on both of these tasks are interpreted as evidence for a model of speech perception that combines semantic and phonological information in the target representation.

In investigations of lexical access, the properties of the mapping from a phonetic input to a lexical/semantic representation are of greatest relevance. The simulations reported by Gaskell and Marslen-Wilson show that a system mapping from connected speech to a distributed semantic representation has essentially the same computational properties as SRNs with localist output representations used by Norris (1990). Specifically the network activates the representation of each word in proportion to the conditional probability of that item given the current input to the network. However whereas in Norris's simulations this is a result of averaged activations at localist lexical units, in the distributed account proposed by Gaskell and Marslen-Wilson this is a consequence of activating a 'lexical blend' – an averaged pattern of activation obtained from the distributed representation of all the words that match the current input. As discussed by Gaskell and Marslen-Wilson (in press) this introduces inherent limitations on the representational capacity of blends in differently structured representational spaces, limitations that have been supported by recent priming studies (Gaskell & Marslen-Wilson, submitted).

Despite the different mechanisms by which this partial activation is produced, the model presented in Gaskell and Marslen-Wilson (1997) retains the limitation discussed by Norris (1990) with respect to onset-embedded words. Since such items do not become unique before their offset, the network cannot distinguish them from longer competitors. Gaskell and Marslen-Wilson discuss a mechanism by which this problem can be resolved within their network (without the addition of a direct, lexical competition as included in Shortlist). Their proposal is that during training the network's target representation is not changed until a single segment at the start of the following word has been presented. In this way the network can use a segment following the offset of an embedded word to rule out longer competitors. However, this account places an arbitrary and fixed limitation on the extent to which recognition can be delayed. So for sequences where the following context forms a lexical garden-path with a longer competitor (such as for the sequence *car pick* where the onset of the following word continues to match the longer word *carpet*) the network will be incapable of identifying the embedded word.

Simulations reported by Content and Sternon (1994) also use a delayed output to model effects of following context in the recognition of embedded words. Like Norris (1990) they used a localist lexical representation, however they added an additional group of outputs to encode the identity of the previous word in the input[5]. In this way the network continues to represent hypotheses regarding the identity of the preceding word and can update these activations in the light of following context. However, this still places a limit on the degree of delay over which following context can be used – i.e. that the ambiguity created by embedded words must be resolved by the offset of the following word. Since extreme cases may violate this assumption (consider for instance, "*I like that cat a lot*" versus "*I like that catalog*" ) Content and Sternon's approach may not offer a general solution to the problems created by embedded words. Nonetheless, it appears that delayed

[5] Content and Sternon actually used a single set of lexical units with a probe input to determine whether these output units should represent the current word or the previous word in the input

recognition in recurrent networks offers some scope for further investigation. One goal of the models developed in this thesis will be to investigate network architectures and training regimes capable of producing a more general solution to the problem of the delayed recognition of onset-embedded words.

## 3.3. Simulation 1 – A distributed account of lexical acquisition

One strength of recurrent neural networks in simulating word recognition is that, since the system starts with randomly configured connections and is then trained to identify words, these models have the potential to account for data from vocabulary acquisition as well as word recognition. However, in accounting for developmental data all the networks reported so far (Content & Sternon, 1994; Gaskell & Marslen-Wilson, 1997; Norris, 1990) are of limited plausibility since they learn from a training set in which the target output is a representation of the current word in the speech stream. Generating such a training set requires the speech stream to have already been segmented into lexical units.

In the previous chapter, a variety of computational accounts were reviewed suggesting that lexical segmentation can be learnt from distributional analysis of large phonologically transcribed corpora. It might therefore be proposed that these mechanisms could simply be combined with word recognition networks to provide an account of spoken word recognition that incorporates insights from the developmental literature. However the solution to the segmentation problem required to train a recognition network goes beyond what can be obtained by a distributional analysis of the speech stream. Recurrent networks that are used to map speech to a lexical/semantic representation of the current word require not only that the location of word boundaries be specified (in order to know when to change the target representation from one word to the next) but also that the identity of the words separated by that boundary be known (in order to set the correct target lexical representation). This is equivalent to assuming that the language learner knows the set of one-to-one correspondences between the speech stream and lexical/semantic representations before vocabulary acquisition can begin. It is as if the language learner knew which concept each word in an utterance referred to *prior* to learning the meaning of words.

Very little of infant directed speech consists of single word utterances, even when care-givers are explicitly instructed to teach their children a new word (Aslin, et al.,1996).

Furthermore, infants are not only able to learn words in contexts where they are provided with an explicit pairing between a word and a concept (see for instance Carey & Bartlett, 1978; or the recent review by Bloom & Markson, 1998). Consequently the word-by-word assumption of the recurrent network models (and of some other connectionist models of vocabulary acquisition, e.g. Plunkett, Sinha, Møller, & Strandsby, 1992) is not supported by the available data. Since vocabulary acquisition can occur when learners hear multiple words with multiple possible referents, there is an additional problem that must be solved by a model of lexical acquisition. As well as discovering an appropriate segmentation of the speech stream, vocabulary acquisition will involve discovering the conceptual representation that corresponds to each lexical item or word in an utterance (see Siskind, 1996, for a more formal computational description of this problem).

The modelling investigated in this thesis incorporates this problem of discovering correspondences between units in the speech stream and units of conceptual representation into a model of word recognition. The approach taken here is to assume that this problem requires the language learner to extract generalisations from the occurrence of lexical items in many different utterances. Thus, the language learner hearing phrases such as "*look at the cat*" "*that cat is sitting on the fence again*" "*does the cat want feeding?*" and so on, will learn to associate the sequence of sounds /kæt/ with whatever conceptual representation is commonly contained in the scenarios that these utterances refer to (presumably a representation of a small, furry domestic mammal). This must occur despite the fact that the appearance of the referent will not uniquely coincide with the sound sequence /kæt/. Instead a number of possible concepts will be plausible as the referents of a longer sequence of speech. In this way, even though utterances in child directed speech will seldom contain single words (Aslin, et al., 1996) and will not be spoken in contexts where only one potential referent is present in the world, the language learner can learn to extract one-to-one correspondences between form and meaning.

This approach provides an account of the source of the supervisory input used in training a recurrent network account of word recognition – namely from the non-linguistic environment in which infants experience spoken language. This view of lexical acquisition as involving a mapping from a spoken utterance to a conceptual representation of the world referred to by that utterance has some similarities with the work of Gleitman

(1994). However the recurrent networks used here provide an explicit computational system in which the acquisition of this mapping from form to meaning can be simulated.

The goal of the modelling reporting in this chapter is thus to investigate whether recurrent networks can be trained to recognise words in connected speech *without* requiring a pre-segmented training set. This work is an extension of previous research using simple recurrent networks to model spoken word recognition (Norris, 1990; Content and Sternon, 1994; Gaskell and Marslen-Wilson, 1997), extending these accounts to deal with phenomena that have proved difficult for these models. Most prominent amongst these is the recognition of onset-embedded words and whether a network can acquire a one-to-one mapping from sound to meaning without a training set in which these correspondences are pre-specified.

### *Modelling the identification of embedded words*

The model that has been motivated here is one in which the task of the recognition system is to activate a representation of an entire sequence of words. Thus, whereas previous models only maintained the activation of an embedded word over a single segment (Gaskell & Marslen-Wilson, 1997) or a single word of following context (Content & Sternon, 1994), the network investigated here must maintain an active lexical representation of all the words that have been heard until the end of the current sequence. This should ensure that the system has adequate time to resolve any temporary ambiguities created by the presence of onset-embedded words.

This approach can be thought of as suggesting that word recognition is an emergent property of the process of identifying entire sequences. Lexical items will be an important level of regularity that exists between sequences of speech and the meanings communicated by those sequences. One justification for this approach comes from a consideration of the problems involved in lexical acquisition where one-to-one correspondences between speech and meaning must be learnt from experience.

### *Modelling lexical acquisition*

Like previous models of spoken word recognition (and unlike the distributional accounts of lexical segmentation reviewed in Chapter 2) a supervised learning process is used in training this recurrent network model. This reflects the assumption that lexical acquisition involves a process of associating form and meaning. Since supervised learning systems require an external teacher to determine the target activation for the network at all points during training, it is necessary to specify where the teacher input comes from. As in other models of word learning (e.g. Plunkett et al., 1992) we assume that vocabulary acquisition involves learning a mapping from form to meaning, in which meaning representations are in part derived from the non-linguistic context in which the language learner hears spoken sequences.

However, in contrast to the Plunkett et al. (1992) account of vocabulary acquisition and the other models of spoken word recognition discussed previously it is not assumed that correspondences between the speech stream and lexical or semantic representations are available to the learner on a one-to-one basis. These one-to-one correspondences must be acquired by the network through generalisation from the experience of hearing sequences of lexical items in different contexts (see Goldowsky & Newport, 1993, for a similar approach to modelling lexical acquisition). Thus, in the example given previously, experiencing the word *cat* in different utterances in which various conceptual representations can be inferred as being a likely referent of that utterance, the system must learn to associate the sequence of sounds /kæt/ with the appropriate semantic representation (cf. de Sa, 1994).

The manner in which this assumption is implemented in this model is perhaps rather less realistic than this description implies. Firstly the model has a representation of all the words in the current sequence as a target during training. This is an unrealistic assumption since it is likely that only a subset of the words in any utterance have an obvious interpretation during acquisition. This reduced capacity has however been suggested to facilitate vocabulary acquisition in a model developed by Goldowsky and Newport (1993), though the same may not necessarily be true for the model proposed here. However since a competent adult listener must be able to activate a representation of all the words in a sequence, the assumption that all words are active in the target representation was incorporated. In this way the model should attain the adult levels of performance that are necessary in order to compare the network’s output with behavioural data.

One approach to language comprehension that is similar to the account proposed in this model is that described in the St. John and McClelland (1990) model of sentence processing. The goal of their model was to activate a ‘sentence gestalt’; a representation capturing the thematic relationships between constituents in a sentence. St. John and McClelland did not specify this representation beforehand, but allowed back-propagation to generate this sentence gestalt, by probing a level of representation with thematic roles for which the network had to output the item that filled that role in the sequence. Although this query network works well, it has the effect of making the target activation for the network undetermined at the start of training. In the modelling proposed here the utterance level representation was generated beforehand. Although this requires stronger assumptions regarding the nature of representations provided before and during training, this has the advantage that the goal of the network’s training regime and its performance following training will be rather more transparent.

To allow easy interpretation of the network’s output, the utterance representation is composed of localist lexical units, each representing a word in the network’s vocabulary. Although this provides an output representation structured in terms of discrete lexical units, this aspect of the model is intended as no more than a computational convenience. Although infants clearly bring a well formed representation of the structure of the outside world to the language learning situation, the lexical output is not intended to suggest that conceptual representations are fixed prior to lexical acquisition. Contextually variable, distributed output representations would offer a more complete account of the vocabulary acquisition process, since the network could then simulate the extraction of invariant lexical/semantic representations from noisy and contextually variable meanings. However, since these distributed outputs would substantially increase the computational demands of the network, current simulations all incorporated localist units. Thus the network is limited to modelling phenomena arising from the process by which spoken input is mapped onto conceptual representations during acquisition and for the time course of processing in adults. Given the similarity between the performance of recurrent network models trained to map speech to localist (Norris, 1990) or distributed semantic representations (Gaskell & Marslen-Wilson, 1997) it was not expected that the use of localist representations would substantially alter these aspects of the behaviour of the network – aside from making it dramatically quicker to train. A further advantage of using localist units is that they side-step the binding problem that is incurred in producing distributed representations of syntactic sequences (see Sougné (1998) for further discussion and a review).

Despite the presence of lexical units in the target output the network's task in identifying lexical items is far from trivial since, in contrast to previous simulations, the target pattern during training is an unordered representation of all the words in a sequence – not just the current word at any point in the input. The training set therefore does not contain any information about which segments in the speech stream map onto individual lexical items. Furthermore, since the target remains static throughout each sequence of words the network is not provided with any information about the location of word boundaries. Finally since words in all positions are represented over the same units, no information is provided about the order in which words occur in the training sequences. The network is therefore trained on a many-to-many mapping between the speech stream and lexical representations from which it must extract one-to-one correspondences between words in the speech stream and lexical items.

During training the target output for the network is to activate units representing all the words in the current sequence. However, during testing, the network will clearly not be able to activate units representing the final words in a sequence until those words have been presented in the input. The network can therefore not be expected to learn the training set to perfection. However, as is the case in networks trained to predict segments and boundaries in utterances (e.g. Christiansen et al., 1998), drawing a distinction between the task on which the network is trained, and network performance during testing may provide a more clearly elaborated psychological account. In the case of the networks investigated here, the immediate task for the network is to associate strings of phonemically coded segments with a representation of the lexical items contained in that sequence. During testing, the performance of the network will be compared to the time-course of activation of individual lexical items inferred from psycholinguistic data on the recognition of words in connected speech. Thus the model is not directly trained to produce the behavioural profile observed during testing.

### 3.3.1. Method

***Training set***

For all the simulations reported in this thesis, the training set for the network was constructed from an artificial language containing 7 consonants and 3 vowel segments coded over a set of 6 binary phonetic features adapted from Jakobson, Fant and Halle (1952). These segments and their feature representations are listed in Table 3.1. In creating the input for these networks, phoneme vectors were concatenated one after the other to make input sequences for the network. No attempt was made to incorporate coarticulation or variability into these input sequences, though incorporating such information should not substantially alter the results reported from these simulations (see Gaskell, Hare, & Marslen-Wilson, 1995; Gupta & Mozer, 1993).

| Transcription | | Phonetic Features | | | | | |
|---|---|---|---|---|---|---|---|
| IPA | MRPA | Vocalic | Consonant | Voiced | Nasal | Diffuse | Grave |
| p | p | 0 | 1 | 0 | 0 | 1 | 1 |
| t | t | 0 | 1 | 0 | 0 | 1 | 0 |
| k | k | 0 | 1 | 0 | 0 | 0 | 1 |
| b | b | 0 | 1 | 1 | 0 | 1 | 1 |
| d | d | 0 | 1 | 1 | 0 | 1 | 0 |
| n | n | 0 | 1 | 1 | 1 | 1 | 0 |
| l | l | 1 | 1 | 1 | 0 | 1 | 0 |
| ɪ | I | 1 | 0 | 1 | 0 | 1 | 0 |
| æ | & | 1 | 0 | 1 | 0 | 0 | 0 |
| ɒ | o | 1 | 0 | 1 | 0 | 0 | 1 |

**Table 3.1: Phonetic feature representation used as input for the computational simulations**

These 10 phonemes were placed into a CVC syllable template which was used to create a vocabulary of 20 lexical items, of which 14 words were monosyllabic and 6 were bisyllabic. To allow investigation of the time course of recognition, lexical items varied in the point at which they became unique from all other words in the networks vocabulary. The artificial language therefore included ‘cohort’ pairs such as *lick* and *lid*, that share the same onset and become unique on their final segment, as well as two pairs of onset-embedded words (e.g. *cap* and *captain*) where the monosyllable is not uniquely identifiable until following context rules out longer competitors. There were also two pairs of offset-embedded words (such as *lock* and *padlock*[6]) to allow comparison of the network's sensitivity to preceding and following context in the recognition of embedded words. These 20 vocabulary items are shown in Table 3.2.

| Length | Type | Word | Phonology | Word | Phonology |
|---|---|---|---|---|---|
| Bisyllable | + onset-embedding | captain | /kæptın/ | bandit | /bændıt/ |
| | + offset-embedding | topknot | /topnot/ | padlock | /pædlok/ |
| | non-embedding | landed | /lændıd/ | picnic | /pıknık/ |
| Monosyllable | onset-embedded | cap | /kæp/ | ban | /bæn/ |
| | offset-embedded | knot | /not/ | lock | /lok/ |
| | cohort competitors | dot | /dot/ | dock | /dok/ |
| | | lick | /lık/ | lid | /lıd/ |
| | non-embedded | tap | /tæp/ | bat | /bæt/ |
| | | knit | /nıt/ | cat | /kæt/ |
| | | pot | /pot/ | bid | /bıd/ |

**Table 3.2: Vocabulary items used for the computational simulations**

Words from this set of items were then selected at random (without replacement) to create sequences of between 2 and 4 words in length. Each sequence was separated by a boundary marker (an input and output vector consisting entirely of zeros). No attempt was made to capture higher-order regularities such as are involved in syntactic or constituent structure, although a set of sequences were excluded from the training set to allow testing of the network's generalisation performance.

***Architecture***

These training sequences were presented to a simple recurrent network of 6 inputs, 50 hidden units with copy-back connections to 50 context units and 20 output units. The network was trained to activate the lexical output units for all the words in the current

[6] Note that *pad* is not a word in the training vocabulary of the network.

sequence. Weights were updated by the standard back-propagation algorithm following the presentation of every input segment (learning rate= 0.02, no momentum, cross-entropy error measure - Hinton, 1989). The architecture of the network and a snapshot of the training regime is illustrated in Figure 3.2.

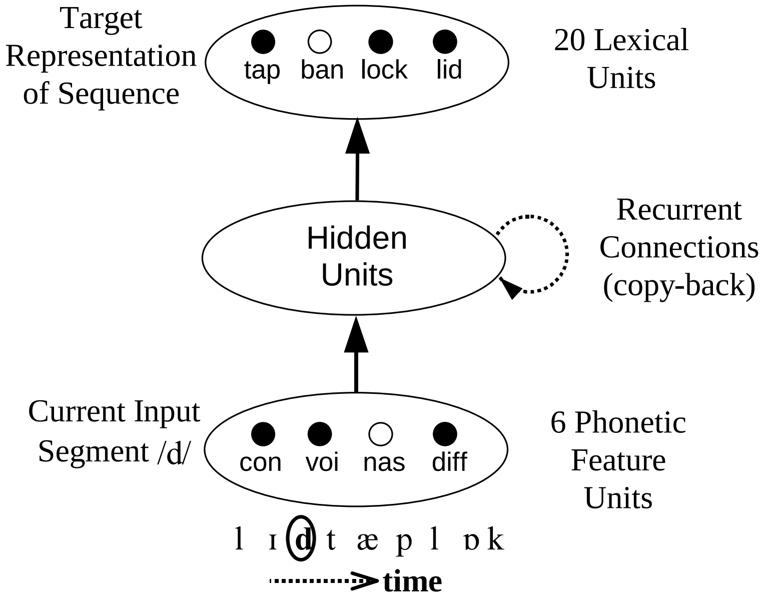


**Figure 3.2: A snapshot of the SRN during training on the segment /d/ during *"lid tap lock"*. Throughout each training sequence the target activation for the network is to activate a representation of all the words in that sequence, not just the current word.**

In preliminary simulations it was observed that changes to the bias weights for the output units (i.e. weights that set the activation threshold for these units) were considerably larger than those to weights connecting the output and hidden units. This is caused by the repeated weight updates with the same target pattern. Early on during each sequence, words occurring late in the sequence cannot be identified. In these cases, bias weights for units representing late occurring words are altered to increase the activation of these units irrespective of the current input. This makes it difficult for the network to learn what input segments correspond to these words. One solution to this problem is to decrease the overall learning rate such that changes to the bias weights cannot swamp updates to the weights connected to the hidden and input units. However this has the disadvantage of greatly increasing the number of epochs required to train the network.

An alternative solution is to disconnect the bias weights from the output units[7]. Since each lexical item occurs with equal frequency in the training set, the prior probability of each output unit being activated will be equal. Consequently removing the bias weights will have no effect on the performance of the fully trained network. Removing the bias weights allowed the use of larger learning rates in these simulations, speeding up the training of the network. All results reported subsequently come from simulations without bias weights to the output units. Training time apart, comparable results were obtained in simulations that included these bias weights.

Ten networks were trained using the architecture and training regime described above. Each network started with a different set of small random weights (initialised to between +/-0.1) and was trained on a different set of 500 000 randomly generated sequences. At this point, the output error measured on a set of test sequences held back during training had reached asymptote. Weights were fixed at their final values and the network was tested.

### 3.3.2. Results

Figure 3.3 shows the activation of target words for a test sequence averaged over the 10 fully trained networks. As can be seen in the graph, the network activates words as their constituent segments are presented at the input. Lexical units become partially activated in response to input that supports their identification (for example *lock* is partially activated at the onset of *lid*). Full activation is only observed when words are uniquely specified in the input. Once identified, lexical items remain active until the end of the sequence when output activations return to zero in preparation for subsequent sequences. This behaviour indicates that the model has learnt to lexically segment the speech stream in order to recognise individual words in connected speech.

[7] Thanks to Gary Cottrell for suggesting this.

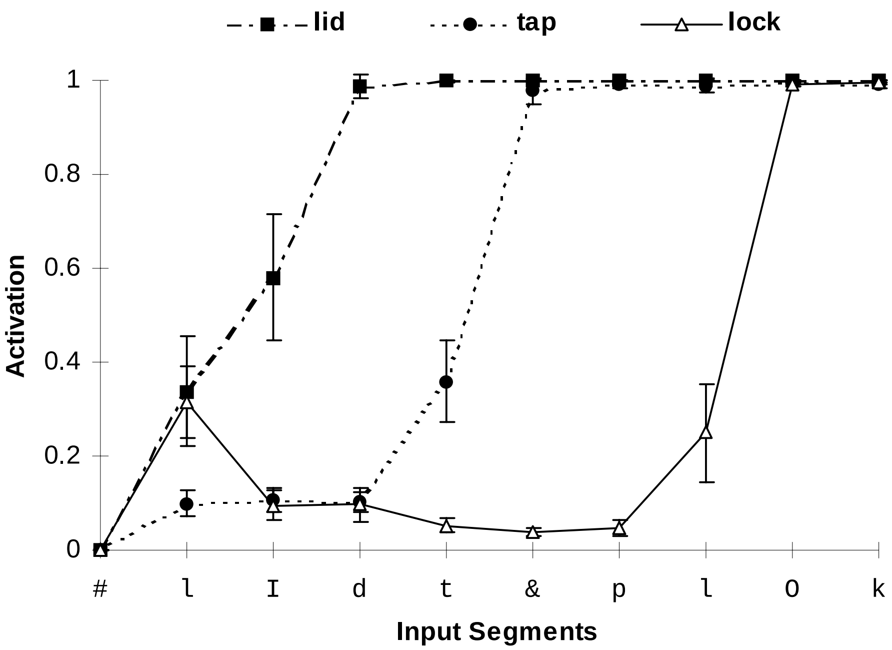


**Figure 3.3: Activation of lexical units during the sequence *"lid tap lock"* averaged across ten networks in Simulation 1. Each network activates words as they are presented in the input and preserves their activation until the end of the sequence. Error bars are 1 standard deviation.**

Since these networks were not provided with explicit cues to the location of word boundaries or with information about which segments make up individual lexical items, learning to identify individual words is a non-trivial task for this system. Unlike the network investigated by Norris (1990), the majority of active target units will not refer to the current word in the input. In cases where these networks are processing a word early on in a sequence it will therefore be impossible for the network to reduce error on these output units. Nonetheless, in spite of these irrelevant targets, the model successfully identifies individual words, through generalising the correspondences between different input sequences and the lexical units activated for those sequences. The strength of the network's generalisation is illustrated by the fact that the same activation profile is observed for test sequences that were not presented during training.

The effect of these irrelevant targets can however be seen in Figure 3.3 in the residual activation observed for words that have yet to be presented in the input. Such activation occurs for all lexical units (except those that are being or have been activated by the speech input) and represents the probability with which each lexical item could appear subsequently in the current sequence. Since there are only 20 items in the network's vocabulary any particular lexical item is fairly likely to occur and thus irrelevant units are activated to 0.1 when there are two remaining words in the sequence and to approximately

0.05 when only one word remains to be presented. In simulations with more realistic vocabularies these residual activations would be negligible since any lexical item is less likely (a priori) to appear in a particular sequence.

As shown by the error bars in Figure 3.3, there is some variability in the partial activations observed across the 10 simulations. This result reflects the different training sets used in each network. In cases where input is ambiguous, partial lexical activations will be biased towards items on which the network has been trained more frequently and more recently. This response to recent training suggests an account of the long term inhibitory effects of competition as shown in an auditory lexical decision experiment (Monsell & Hirsh, 1998). The auditory presentation of a cohort competitor, such as *bruise*, slows subsequent responses to the target *broom* even where more than a minute of unrelated trials separate the prime and target. These inhibitory effects have been interpreted as evidence for lateral inhibition between lexical items. Weight updates following the presentation of each word would provide an alternative account of this competition effect since they would boost the strength of weights involved in identifying the prime word and reduce the activation of weights that activated the competitor. Thus these results could be simulated without requiring direct inhibitory connections between lexical units.

***Partial activation for ambiguous input***

In the preceding discussion it has been suggested that the degree of activation of lexical units reflects the probability of a word having occurred in the input. This is confirmed by comparing the partial activation observed in different competitor environments shown in Figure 3.4. The left hand chart shows the pattern of activation observed for items with cohort competitors (in this case *lick* and *lid*). Output activations in these networks approximate the conditional probabilities of all the lexical candidates that match the current input (Servan-Schrieber, Cleeremans & McClelland, 1991). Thus at the onset of *lid* (where three candidates match the input) each competing word is activated to just over 0.3. On presentation of the second segment, when two candidates match, each item is activated to approximately 0.5. It is only at the offset of *lid* that full activation is obtained at the appropriate lexical unit. This result replicates Norris (1990) in showing how partial activation of cohort competitors can be simulated in a recurrent network as a consequence of the averaging of output activations for ambiguous inputs. However, whereas in Norris' network this result would be expected since the total activations across all the output units sum to one, in the simulations reported here the result illustrates that this averaging of activations only occurs between competing output units. Thus although the networks have been exposed to sequences in which both *lick* and *lid* are fully active, during the processing of a sequence in which only one or other word is present, the activation of competing lexical units sums to 1 and can be interpreted probabilistically.

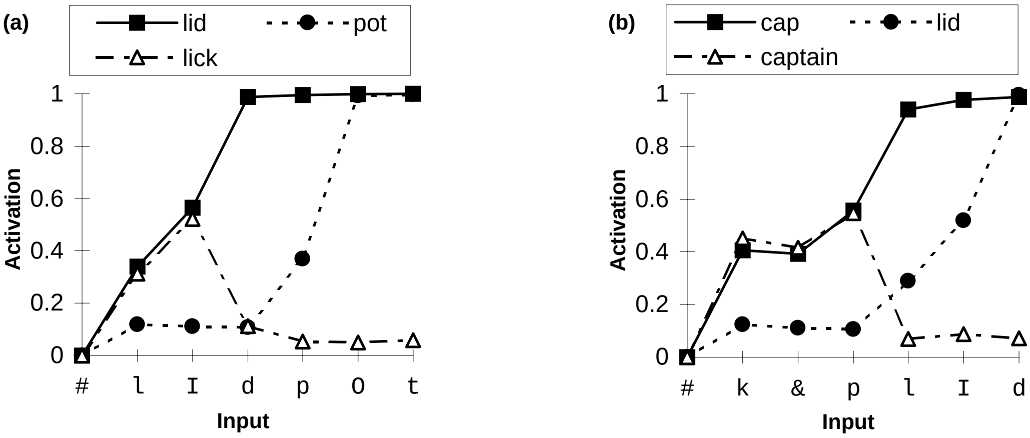


**Figure 3.4: Activation of cohort competitors and onset-embedded words in Simulation 1.**
**(a) Cohort competitors (*lid/lick*) during the sequence *"lid pot"***
**(b) Onset-Embedded words (*cap/captain*) during the sequence *"cap lid"***

In the psycholinguistic literature on spoken word recognition there is a large amount of empirical evidence that can be accounted for by a model in which activations are proportional to the conditional probability of the different lexical items that match the input. Such a model provides a natural account of the effect of word frequency on competition between cohort pairs like *road* and *robe* (Marslen-Wilson, 1990). In these cross-modal priming experiments, it was observed that high-frequency members of the cohort were more active for ambiguous stimuli (where the offset segment was cut off). This effect could be described as an reflecting competition between lexical units where the more frequent and hence more active candidate will dominate. However, in this simulation this effect is the result of the probabilistic behaviour of a system without direct competition between lexical units. More frequent lexical items are a more probable interpretation of ambiguous input and hence are more active during recognition. Results reported by Gaskell and Marslen-Wilson (submitted) further support this account by suggesting that the magnitude of semantic priming observed for ambiguous word fragments is directly proportional to the conditional probability of the prime word in that cohort environment.

### *Processing onset embedded words*

The pattern of activation observed for cohort pairs is repeated almost identically in Figure 3.4b for onset-embedded words. At the offset of the monosyllable, the two matching lexical items (*cap* and *captain*) are equally activated. It is only at the onset of the following word (the segment /l/ in *lid*) that disambiguating input is received (since the input will mismatch with *captain* at this point) and the networks are able to fully activate the lexical unit representing the word *cap*. Such behaviour indicates that the lexical competition network utilised in Shortlist (Norris, 1994) is not necessary to account for the recognition of onset-embedded words. A network in which the target of the recognition process is a representation of an entire sequence of words is also able to use following context to identify words that do not become unique until after their offset.

Interestingly the time-course of activation observed for onset-embedded words and longer competitors in these networks differs from that predicted by TRACE and Shortlist. In the current set of recurrent network simulations, lexical activations represent the conditional probability of each word given the current input. Consequently at the offset of an embedded word, where a short word and a long word are equally likely, recurrent networks will activate both words equally (see Figure 3.4b). This is in contrast to models incorporating direct intra-lexical competition which, because of greater inhibition for long words, predict greater activation for short word candidates at the offset of an embedded word. This difference between recurrent network and lexical competition accounts will be investigated experimentally in subsequent chapters.

Despite this difference in the time course of activation for embedded words, both recurrent network and lexical competition accounts still support an account of lexical segmentation in which mismatch with longer competitors plays an important role in the recognition of embedded words. Figure 3.5 therefore shows the networks' response in two cases where mismatch with longer lexical items is absent or delayed. The first example (Figure 3.5a) is where these networks are presented with a longer lexical item that contains a word embedded at its onset (for example *captain*). In this case the networks strongly activate the longer word and reject the embedded word.

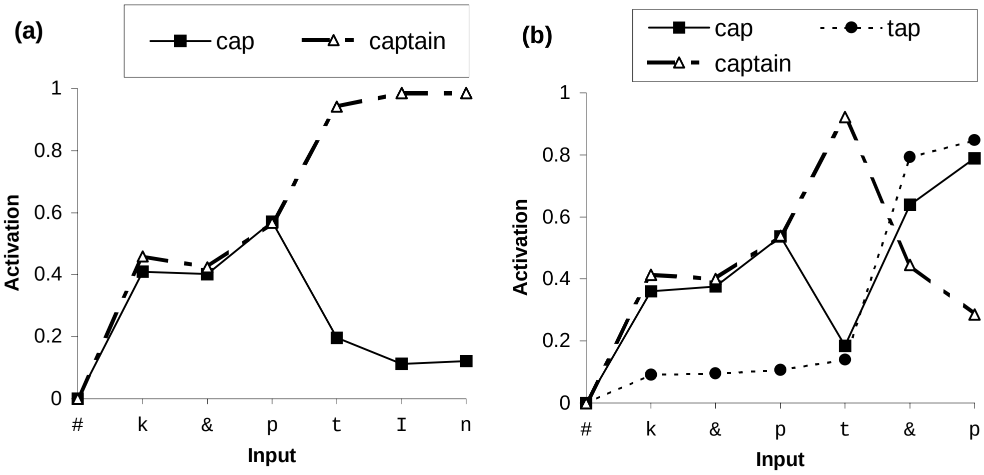


**Figure 3.5: Activation of embedded words and competitors for long word sequences and lexical garden-paths in Simulation 1.**

**(a) Bisyllables with embeddings (*captain/cap*) during *"captain"***

**(b) Lexical 'garden paths' (*cap/captain*) during *"cap tap"***

However, in the lexical garden-path in Figure 3.5b, an embedded word is followed by a continuation that matches the longer competitor. For example in the sequence *"cap tap"*, the competitor *captain* cannot be ruled out until the vowel of the second syllable. In this cases the recurrent networks are still able to revise lexical activations in response to the delayed mismatch between the speech stream and the longer competitor.

### *Processing offset-embedded words*

The final set of results shown for these networks concern the identification of words which contain another lexical item embedded at their offset. By the account proposed in the original cohort model (Marslen-Wilson and Welsh, 1978) – where only words sharing the same onset are jointly activated – it would not be predicted that these offset-embedded words (e.g. *lock* in *padlock*) would be activated during recognition. These recurrent networks show this pattern of performance, as illustrated in Figure 3.6. In contrast to the networks' performance for onset-embedded words, the model clearly rejects offset-embedded words during recognition, illustrated by the minimal activation of *lock* during *padlock* in Figure 3.6a. Similarly during presentation of a sequence such as *"lid lock"* – where the offset of the preceding syllable is identical to the syllable offset for the longer competitor – these recurrent networks do not activate the longer word.

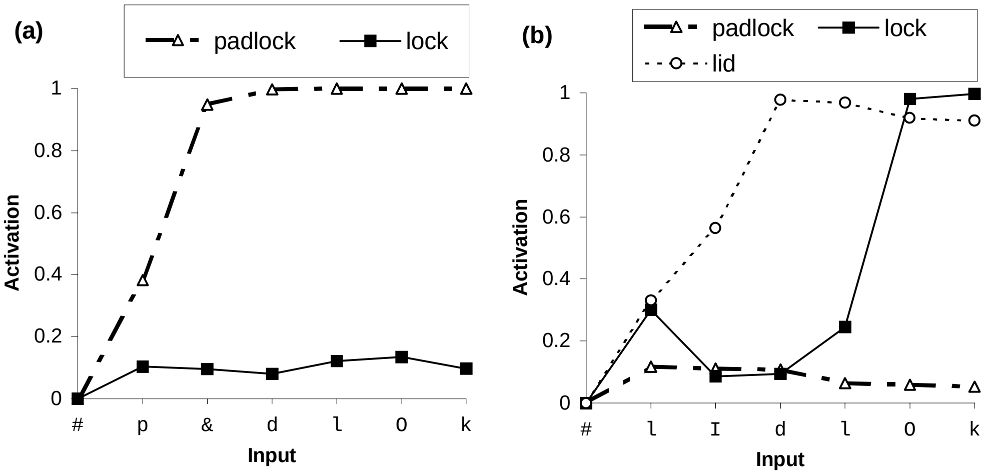


**Figure 3.6: Activation of offset-embedded words in Simulation 1:**
**(a) Bisyllables with offset embeddings (*padlock/lock*) during *"padlock"***
**(b) Offset embedded words (*lock/padlock*) during *"lid lock"***

Empirical evidence on the activation of offset-embedded words in connected speech is unclear at present. Shillcock (1990) reports obtaining significant priming from offset-embedded words to an associatively related target (e.g. *trombone* primes RIB, an associate of the embedded word *bone*) in English; a result that has been replicated in Dutch (Vroomen & de Gelder, 1997) and in single word presentation in English (Luce & Cluff, 1998). However, experiments by Gow and Gordon (1995) using stimuli in which both syllables make up a word (e.g. *window* composed of the two words *win* and *dough*) failed to replicate this finding. These stimuli would be expected to be more likely to support the offset-embedded interpretation since initial syllable matches a word. A series of experiments by Marslen-Wilson et al. (1994) using cross-modal repetition priming (which again might be expected to show increased priming compared to semantic priming) also failed to show priming for offset-embedded words. Given the apparent inconsistency of these findings and the lack of comparsion between the activation of correct and embedded word interpretations, no result reported so far would directly refute the pattern shown both by recurrent network and competition models – namely that offset-embedded words receive substantially less activation than the longer words in which they are embedded. These priming experiments will be reviewed in more detail in Chapter 4.

### 3.3.3. Discussion

The network described here has learnt to recognise words in connected speech without exposure to a pre-segmented training corpus. By generalising its experience of different sequences of phonemes in the input and different combinations of lexical units activated in the output, the network has learnt correspondences between the speech stream and lexical items from the language on which it was trained. Thus the network suggests an account of an important aspect of vocabulary acquisition – namely that correspondences between speech and meaning can be extracted through exposure to situations in which many words are heard and where many possible referents of those words are present in the environment.

In identifying words in connected speech the network implements a form of the maximal efficiency assumption in recognising words in sequences – progressively updating lexical activations as more input is presented. However, unlike some previous cohort-style accounts using recurrent networks (Norris, 1990; Gaskell and Marslen-Wilson, 1997) the network  is able to deal with ambiguous input, not only where the ambiguity is resolved within a word, but also where post-offset information is required for recognition – for instance in identifying onset-embedded words.

Furthermore these simulations do not rely on postulating an additional computational mechanism to implement direct intra-lexical competition (as in Shortlist – Norris, 1994). Nor does it require a training corpus that contains explicitly marked word boundaries (unlike Content & Sternon, 1994). It can therefore be claimed that the system is ‘learning’ to lexically segment connected speech. At least for this limited training set, correspondences between sequences of input and output activations (analogous to those found in the mapping from form to meaning) do provide a means by which a network could learn to identify individual words in connected speech. Further simulations are merited to investigate whether this method remains effective for more realistically sized vocabularies.

In claiming that these networks are learning lexical segmentation it is not intended that this is the only means by which segmentation can be learnt. The review of models of lexical segmentation in the previous chapter suggested that distributional analysis in self-supervised and unsupervised systems plays an important role in the discovery of boundaries between lexical units in connected speech. The goal of a second set of simulations was therefore to investigate how these distributional accounts of lexical segmentation might combine with the account of vocabulary acquisition proposed here.

## 3.4. Simulation 2 – Combining distributional and lexical accounts

The simulations that have been presented so far provide an account of how lexical acquisition could proceed without assuming that the system must be trained with one-to-one correspondences between sections of speech and lexical representations. However in the previous chapter, systems were described that can learn statistical properties of connected speech that can be used in detecting word boundaries. The ability of these networks to discover word boundaries in the unsegmented input seems to challenge an assumption made in Simulation 1 - that the speech stream is unsegmented prior to lexical acquisition. Further simulations were therefore carried out to investigate how these distributional and lexical accounts of lexical segmentation may be combined in vocabulary acquisition.

One means of encouraging a network to process distributional information is to use the prediction task described by Elman (1990). Training an SRN to output the identity of the input at the next time step has been shown to be an effective way of getting a network to represent the location of potential word boundaries (Cairns et al., 1997; Christiansen et al., 1998). Since the target output for the networks described in Simulation 1 is a representation that remains static during each sequence of words these networks may not make efficient use of the statistical regularities that exist in the training set to allow the detection of word boundaries. Adding the prediction task to the network may therefore help it to use distributional structure in learning to identify words in connected speech.

The approach taken in these simulations was to retrain the networks investigated previously, adding an additional set of output units predicting the input that will be presented to the network at the next time step. By comparing the training profile of networks with and without this prediction task the role of distributional analysis in vocabulary acquisition can be explored. These simulations will allow investigation of whether the prediction task, used in its simplest form, facilitates vocabulary acquisition in a recurrent network.

### 3.4.1. Method

Ten networks were trained using the architecture shown in Figure 3.7. These simulations used the same 10 sets of initial weights as in the first set of simulations, with additional weights connecting the hidden units and bias unit to a set of output units trained to predict the input features presented at the next time step. These networks were trained on 500 000 sequences from the same language used previously. Networks starting from the same initial weights were trained on the same randomly generated training sets. In this way ten pairs of networks with and without the prediction task can be compared. Each pair of networks started from the same initial weights and will be trained on the same set of sequences – a repeated measures comparison.

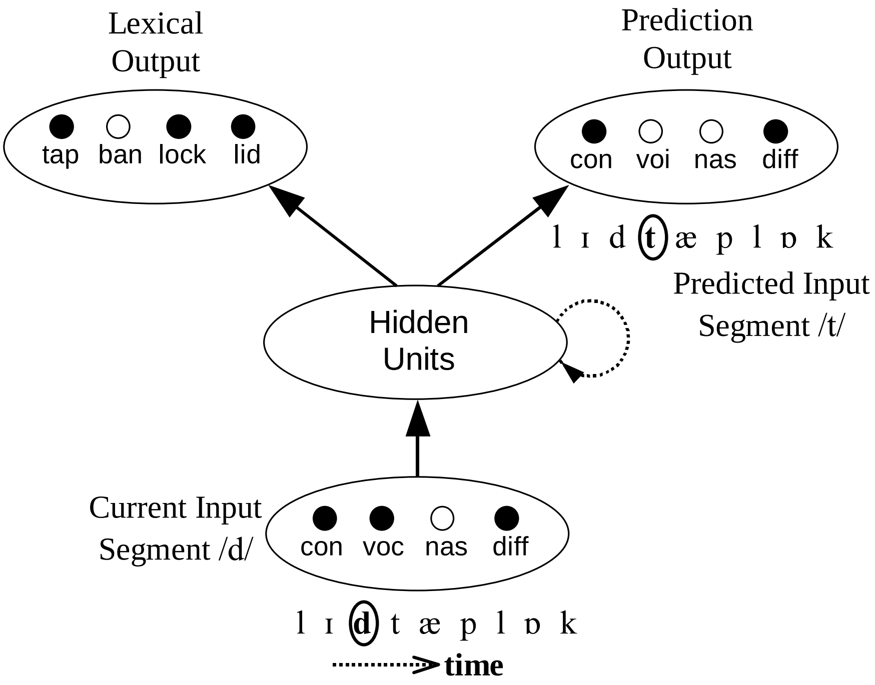


**Figure 3.7: A snapshot of the SRN trained in Simulation 2. The architecture and training regime are identical to that shown in Figure 3.2, except for the additional output units trained to predict the input at the following time step.**

### 3.4.2. Results

Every 25 000 sequences during training the networks' performance was evaluated on two distinct types of words – on the ten non-embedded monosyllables in the training set (as listed in Table 3.2) and on the two onset-embedded monosyllables in the training set. At the uniqueness point of each type of word the difference in activation between the lexical unit representing the target and the most active competitor was measured. For example, for the monosyllable *lid* the most active competitor was frequently *lick*. Consequently the difference in activation between *lick* and *lid* at segment /d/ in Figure 3.4a was one of the ten data points used to measure the average discrimination performance of each network. Similarly for embedded monosyllables the most active competitor at the uniqueness point would be likely to be the longer word. Thus, the difference in activation between *cap* and *captain* at the segment /l/ in Figure 3.4b would be measured. Since the onset-embedded words require following contexts to become unique, results were averaged over 5 consonants that could follow each embedded word, excluding the lexical-garden path sequences (Figure 3.5b) and continuations which duplicated the final segment (e.g. *cap put*). These results averaged over 10 networks with and without the prediction task at different points in training are presented in Figure 3.8.

As can be seen by comparing Figure 3.8a and b, both sets of networks learnt to discriminate the non-embedded monosyllables more rapidly than the embedded monosyllables. This is unsurprising since the networks must learn to use a variety of following contexts to identify the embedded words whereas the sequence of segments identifying a non-embedded word is invariant. This illustrates the effect of inconsistency of the input that the network must use to recognise embedded words. It is for this reason that these items are acquired more slowly by the network.

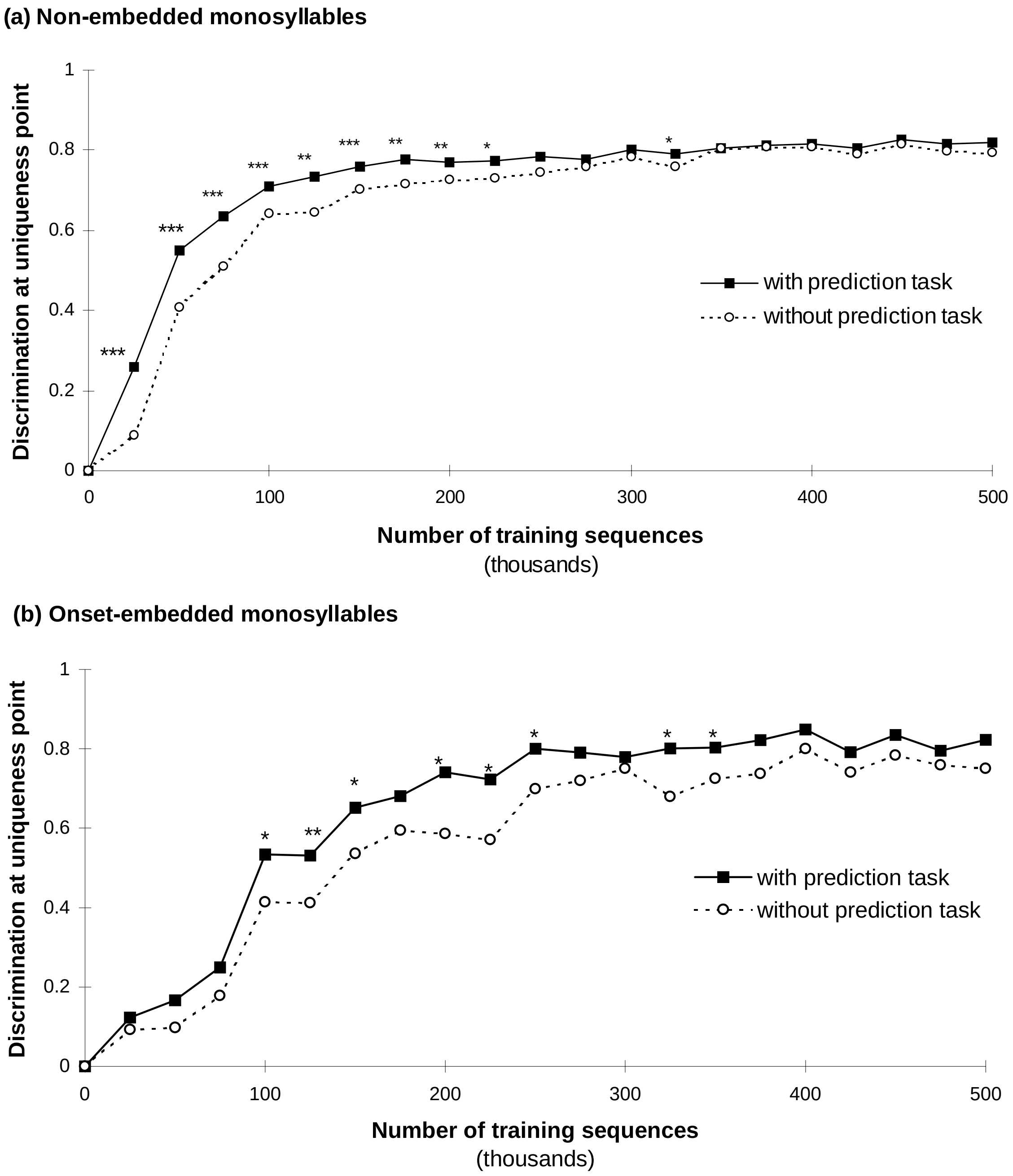


**Figure 3.8: Discrimination performance for networks trained with and without the prediction task. Performance measured at uniquness point for (a) Non-embedded monosyllables (b) Onset-embedded monosyllables. Paired t-tests[8] comparing networks with and without the prediction task *** p(1 tail)<0.001 ** p(1 tail)<0.01, * p(1 tail)<0.05**

[8] All tests were one-tailed paired t-tests, (df= 9) testing whether the discrimination performance of the 10 networks trained with the prediction task was significantly better than the performance of the identical network without the additional output task.

For both types of words however, networks that included the prediction task learnt the mapping significantly faster than the same network without this additional output. This can be seen in the results of paired t-tests on the networks performance at each point during training as marked in Figure 3.8. This speeded acquisition is most noticeable for the non-embedded monosyllables. Networks with the prediction task perform significantly better at discriminating monosyllabic words from competitors early on in training. These differences decrease as both networks reach asymptotic performance on the task; once fully trained there is no significant difference between the performance of the two sets of networks.

This effect is also present – though in a noisier form – for the onset-embedded words. During the period when the network is learning to identify embedded words, networks trained with the prediction task perform significantly better at the task of discriminating these words from their competitors. This effect appears to be weaker than for the non-embedded words partly as an artefact of the small scale of these simulations. Since these data are based on just the two embedded words in the network's training set, the results are susceptible to 'buffeting' by recent training, as described in connection with Figure 3.3 (see Bullinaria & Chater, 1996, for a more thorough discussion of artefacts associated with small scale models).

### 3.4.3. Discussion

These simulations demonstrate that networks trained to map a sequence of connected speech to a representation of all the words contained within that sequence can learn to lexically segment the speech stream and use following context appropriately in the recognition of onset-embedded words. Furthermore it has been shown that learning in such a network is facilitated by the use of an additional set of output units trained to carry out a prediction task that has been suggested in the developmental literature as an account of how infants discover the lexical segmentation in the speech stream. Thus the network provides a potential model of how lexical and distributional accounts of lexical segmentation may combine during acquisition.

It is clear that processes involved in learning the statistical structure of the input are not only beneficial in learning segmentation, but also assist a network in mapping the speech stream to meaningful units. Note, however, that the prediction task alone was considered inadequate as an account of lexical identification. This result therefore suggests that incorporating the prediction task helps the network develop appropriate internal representations that are then re-used in training the lexical output (see Clark & Thornton, 1997, for further discussion of this sort of ‘scaffolding’ process in connectionist learning).

Note that although both outputs were trained concurrently it is not the case that they learn at the same rate. Measuring output error separately for both sets of output units suggests that the prediction task is learnt more rapidly than the lexical output. The error curve has levelled off by the time the network has been trained on approximately 25000 sequences – long before the lexical output reaches asymptote. This is suggestive of the pattern observed in development. Preferential listening experiments suggested that children learn distributional or phonotactic aspects of their native language in the latter half of their first year whereas vocabulary acquisition proper does not commence until sometime during the second year (Jusczyk, 1997).

The networks described in this chapter provide a direct demonstration that distributional analysis carried out for the prediction task assists a network in learning to identify words in connected speech. Further investigation of these networks’ training profiles may therefore help to clarify the role of distributional analysis and pre-lexical segmentation cues in vocabulary acquisition. For instance it is unclear what aspects of distributional analysis are most valuable in bootstrapping vocabulary acquisition. Work by Christiansen et al. (1998) argues that a combination of cues (from phonotactics, metrical stress and marked utterance boundaries) is more effective than any single or paired cue. Future investigations could evaluate whether the same is also true for the architecture used here – though this would require extending the networks architecture and training regime to cope with more realistically structured training sets.

## 3.5. General discussion

Computational simulations have shown that the sequential recognition account of lexical segmentation that was reviewed in Chapter 2 can be implemented very simply in a recurrent neural network. However, these implementations have previously been incapable of identifying words that do not become unique before their offset – i.e. onset-embedded words. Although previous authors (Norris, 1994) have suggested that adding a secondary competition network to the output of a recurrent network is necessary to account for the identification of onset-embedded words, the simulations reported in this chapter have shown that no additional computational mechanisms are required. Merely extending the output representation on which the network is trained to include information on a sequence of words (rather than the single word used previously) is sufficient to enable a recurrent network model to recognise onset-embedded words.

Simulations reported in this chapter have also shown that networks trained in this way respond in a probabilistic fashion to temporarily ambiguous input. These models therefore retain the maximally-efficient recognition that was a strength of the original recurrent network simulations. As shown in Figure 3.4b, this produces a distinct activation profile for onset-embedded words than was observed in models incorporating direct inter-lexical competition. The short word bias that is required of lexical competition models in order to allow the identification of onset-embedded words is no longer observed in these recurrent networks. Consequently, one goal of the subsequent chapters of this thesis will be to evaluate and extend the available experimental evidence on the identification of onset-embedded words to determine which activation profile (short word bias or probabilistic activation) more accurately simulates the available empirical data on the time course of identification of onset-embedded words in connected speech.

### *Developmental accounts of segmentation and identification*

One important aspect of the recurrent network models used in this chapter is that the use of gradient descent learning algorithms to train the network suggests an account of the developmental processes involved in learning lexical segmentation and identification. The specific assumption made in these simulations is that infants do not acquire the mapping from speech to meaning by associating a representation of the form of a single word to a representation of the meaning of that word. Instead, the simulations presented here propose that the sound to meaning mapping is learnt by associating an entire sequences of sounds to a representation of the possible meaning of that sequence. The pairing of sequences of spoken input with possible interpretations of that speech allows the network to extract regularities between single lexical forms and their meanings from exposure to multiple, unsegmented sequences.

One challenge to this developmental claim – that lexical acquisition arises through learning a mapping from unsegmented speech to unsegmented meaning – comes from developmental data showing that prior to learning the meaning associated with lexical items, infants appear able to detect words in connected speech. This has been shown by preferential listening experiments in which 8 month old infants familiarised with an isolated word will subsequently listen longer to a sequence that contains a familiarised word (Jusczyk & Aslin, 1995). The converse result has also been shown – infants familiarised with words in sequences will listen longer to those words presented in isolation (see Jusczyk, 1999 for a review). These results provide evidence that infants are able to learn sequences of sounds that cohere as words, <u>prior</u> to acquiring the mapping from those sound sequences to meaning.

In Simulation 2 in the current chapter, it was shown that prior knowledge of the structure of speech sequences – as extracted using the prediction task – is shown to facilitate lexical learning. However, it is unclear whether this statistical knowledge of the speech stream is adequate to account for the apparently lexical knowledge of sequences in experiments reviewed by Jusczyk (1999). Simulations reported by Cairns et al (1997) for instance, failed to show lexical effects in networks trained on the prediction task alone. It is possible that neural network simulations of the prediction task are therefore insufficiently powerful to account for infants' word learning abilities prior to the acquisition of the form-meaning mapping. This statistical learning system may need to be bolstered by more powerful mechanims (such as the symbolic algorithms reviewed by Brent, 1999a) in order to account for distributional learning of word forms by infants.

Nonetheless, whatever the pre-lexical abilities of language learning infants, there is evidence that the structure of the adult lexicon is primarily determined by the nature of the mapping from form to meaning. For instance, in work investigating the processing of morphologically complex words, it is suggested that semantic factors are a major determinant of whether words are lexically represented in a decomposed form (Marslen-Wilson, et al., 1994). As reviewed in the previous chapter, where there is a transparent relationship between a complex word and its stem, the lexical representation for that item will be decomposed. Hence the lexical representation of *departure* will be derived from that of the stem *depart* while the semantically opaque relationship that exists between the stem *depart* and the word *department* would not permit this decomposition. Thus, it is unclear whether even highly sophisticated distributional learning systems could account for the lexical representation of derivational morphology. To the extent that the structure of lexical representation is affected by semantic factors, then accounts of lexical acquisition based on the extraction of form-meaning correspondences will be necessary. The development of connectionist accounts of morphological processing in which decomposition of morphologically complex words is an emergent property of a distributed form-to-meaning mapping (Gonnerman, Devlin, Anderson & Seidenberg, submitted) therefore lends support to the account of lexical acquisition that has been proposed here.

# 4. Investigating the recognition of embedded words

The presence of words embedded at the onset of longer words has been suggested to challenge certain models of lexical segmentation and identification. As reviewed in the previous two chapters, it has been argued that a lexical segmentation mechanism that uses the pre-offset recognition of words in connected speech to identify word boundaries would be disrupted by these lexical items. If embedded words require post-offset information for longer competitors to be ruled out, pre-offset identification would not be possible and the account of lexical segmentation proposed in sequential models of spoken word recognition would not be able to operate effectively.

Some authors propose that the temporary ambiguity of these words necessitates models of word recognition that incorporate direct competition between lexical units (McQueen, Cutler, Briscoe & Norris, 1995). Lexical competition allows the identification of onset-embedded words since it enables mismatch which rules out longer competitors to boost the activation of the embedded word (McClelland & Elman, 1986; Norris, 1994).

However, in the previous two chapters, two different strands of evidence have been presented that question the validity of this inference from the presence of embedded words to models of spoken word recognition that employ lexical-level competition. Given the theoretical importance of onset-embedded words in distinguishing between alternative accounts of lexical segmentation and spoken word recognition, experimental evidence is required to support this argument.

Two assumptions are involved in this argument from embedded words to lexical competition, both of which make specific predictions regarding the time course of identification of words in connected speech. Consequently, experimental investigations can be used to evaluate whether the conclusion of McQueen et al. (1995) – that lexical competition is a necessary property of models of spoken word recognition – is valid.

The first assumption is that ambiguities created by onset-embedded words are resolved by incorporating direct competition between lexical items. In its strongest form this argument could be interpreted as suggesting that models without competition are incapable of identifying onset-embedded words. This strong form has been ruled out by simulations reported by Content & Sternon (1994) and by the recurrent network models investigated in the previous chapter. As described in Chapter 3, networks in which the target of the recognition processes is a representation of an entire sequence provide a natural account of the identification of onset-embedded words without incorporating direct, inhibitory links between lexical units. Following contexts that rule out longer competitors are used to rule in embedded words, allowing their identification.

However, it may be more reasonable to reinterpret the argument presented by McQueen et al. as stating that the time course of identification of embedded words in connected speech more closely matches the predictions of lexical competition accounts than models that lack direct competition between lexical items. In previous chapters, lexical competition models were described that predict a short word bias during identification. That is, at the offset of a sequence of phonemes making up an embedded word, competition based models such as TRACE and Shortlist predict greater activation of units representing short words than long words.

Conversely, the recurrent network simulations described in Chapter 3 display probabilistic behaviour in the resolution of ambiguities during recognition. Thus, these networks predict that short embedded words and longer competitors will be equally active following a fragment of speech that matches both a short word and the onset of a longer word (all other things being equal). This difference between the two accounts can be tested in experiments on the time course of identification of words in connected speech. Specifically, lexical competition models predict increased activations for short words in a case where two otherwise equally plausible lexical items are active during identification, while the recurrent network account would predict that there will be no bias towards either short or long words during identification.

In discussing this apparent discrepancy between models with and without direct lexical competition, it is apparent that both models predict that onset-embedded words will be ambiguous with longer competitors during identification. It is only on the basis of complete ambiguity between embedded words and longer competitors that pre-offset identification of onset-embedded words would not be possible.

As described in the review of the acoustic-phonetics literature in Chapter 2, however, differences in segments at word onsets and duration differences between syllables in short and long words may provide acoustic cues to distinguish embedded words from longer competitors. Any pre-lexical acoustic cue that helps distinguish onset-embedded words from longer lexical items would substantially reduce the ambiguity created by embedded words. These cues would weaken the claim that onset-embedded words produce ambiguities that can only be resolved through the use of lexical competition between word candidates.

From an experimental perspective, acoustic cues to distinguish short and long words would predict that an onset-embedded word (e.g. *cap*) should be activated more strongly by a fragment containing that word than by speech containing a longer lexical item (e.g. *captain*). Conversely, longer words (*captain*) should be activated more strongly by a matching fragment of speech than by a fragment containing an embedded word (*cap*).

Given the conflicting predictions regarding the time course of activation of embedded words in connected speech, the opening section of this chapter will review the relevant experimental literature. This review will focus on whether experiments make comparisons between embedded words and longer competitors that would be required to detect the presence of discriminatory acoustic cues and also whether biases towards short word hypotheses show up in the results of these experiments.

## 4.1. Review of previous experiments

In investigating the time course of activation of words in connected speech, researchers come up against methodological problems caused by the temporal nature of speech. For instance, in order to use reaction time as a dependent measure for a behavioural task, response times need to be measured from an appropriate position in the speech stream. In investigations of visual word recognition the time taken to process a stimulus can uncontroversially be measured from the onset of the visually presented word. However, in spoken word recognition the time that stimulus items take to be presented may differ between different words. Consequently, measurement of response time relative to the onset or the offset of a word may introduce an experimental confound through differences between stimuli in the time at which information needed to make a response becomes available in the speech stream.

This difficulty in controlling both the amount of sensory information available to participants and the amount of time provided for processing of the speech stimuli has led to research focusing separately on these two aspects of the recognition process. For instance, the gating technique provides information regarding the amount of sensory information required for the identification of words; whereas tasks such as auditory lexical decision or word-spotting provide an indication of the amount of time required for processing stimuli that are assumed to be matched for the rate at which sensory information becomes available.

These issues will be prominent in reviewing previous experimental investigations on the time course of identification of onset-embedded words. Although providing valuable information on the nature of the recognition process for embedded words, there is little data comparing the lexical activation of embedded words and longer competitors at relevant positions in the speech stream. The differential predictions derived from the two classes of computational models and from acoustic-phonetic analysis of spoken words are specific to the activation of lexical hypotheses at particular points in the speech stream. The results of previous experiments may therefore only falsify the predictions of different accounts where they address the processing of stimuli at specific positions in the speech signal.

### 4.1.1. Gating

Gating is a frequently used task in experimental psycholinguistics in which speech is presented to subjects in fragments or gates of progressively increasing duration. Following each gate, subjects are generally asked to write down their best guess as to the identity of the word (or words) that they can hear, along with a rating reflecting their confidence in the response that they have given. By recording subjects' responses and confidence ratings at gates stepping through a word (usually starting from word onset), the gating task can be used to provide measures reflecting the activation and identification of competing lexical hypotheses as acoustic information accumulates. For an overview of research using the gating task see Grosjean (1996).

Dependent measures provided by gating can be expressed in terms of the amount of sensory information required for activation of a given lexical item. This is usually measured by the *isolation point* of a stimulus – the point at which subjects give a correct response and then do not change their mind at subsequent gates. An alternative statistic derived from gating measures the amount of input that is required for subjects to confidently recognise lexical items. This is measured by *total acceptance point*, or *recognition point*, usually defined as the point at which confidence ratings reach a predetermined criterion. Both measures are argued to reflect the amount of sensory information required for lexical access.

Note however that on the criteria that were discussed earlier, the standard gating task (with multiple, successive presentations of individual items and un-timed written responses) will not provide any information about the processing time required for lexical access or identification. Indeed this failure to control the amount of processing that can be done on sections of speech may introduce responses biases or otherwise distort the results obtained in gating. For instance, experiments by Cotton and Grosjean (1984) suggest that, through participants perseverating with previous responses, the repeated presentation scheme may produce an overly conservative estimate of the amount of sensory input required for identification. On a more positive note, work by Tyler and Wessels (1985) suggests that spoken responses made under time pressure match reasonably well to those obtained when subjects write their responses.

Experiments using the gating task support accounts in which the recognition of onset-embedded words is delayed until after their acoustic offset. Grosjean (1985) gated through test words measuring isolation points and recognition points for low-frequency monosyllables and frequency-matched bisyllables in minimal sentence contexts. Grosjean found that many monosyllabic words were not isolated or recognised until after their acoustic offset. For instance the isolation point for the word *bun* in the sentence "*I saw the **bun** in the store*" came at the offset of the following word in the sentence – approximately 150ms after the offset of the word.

Although this result may suggest that acoustic cues to rule out long words are not present in these stimuli, it is unclear whether this conclusion can be drawn in the absence of direct comparisons of matched short and long word stimuli. Although some incorrect responses in the Grosjean study were longer words that contained the target (for example, responding *bunny* for the word *bun*), there was no comparison of responses to stimuli containing this longer word. Therefore, these experiments would be insensitive to effects produced by acoustic differences between short and long words. Similarly, although the presence of long word responses to short word stimuli may be taken as evidence that responses were not entirely biased towards short word responses, it is unclear how to evaluate short word biases without investigation of responses to long words which contain onset-embeddings.

A gating experiment carried out by Bard and colleagues (Bard, Shillcock & Altmann, 1988) evaluated the extent of delayed recognition for more naturalistic speech stimuli. When gating through samples of connected speech a word at a time they found that listeners typically failed to identify some 20% of words (mostly closed-class items) until after their acoustic offset. However, since stimuli were presented to participants a word at a time in this study, speech was explicitly segmented during presentation. Consequently, questions regarding the ambiguity of words embedded at the onset of longer words are not addressed. Furthermore, since short word stimuli would be cut off at their acoustic offset (making the length of the word explicit), the results cannot evaluate short biases during lexical identification.

Gating experiments have provided valuable information on the relationship between the acoustic signal and lexical access in connected speech. However, there has, thus far, been no systematic investigation of the recognition of embedded words in connected speech where competition between monosyllables and the longer words in which they are embedded has been controlled for. Consequently, the utility of the acoustic cues to word length that was described in Chapter 2 remains unclear. Similarly, without direct comparisons of responses to short and long target words, it is unclear whether short word biases are present in these experiments. Further gating experiments with materials designed to investigate the recognition of onset embedded words are therefore necessary.

### 4.1.2. Word-spotting and auditory lexical decision

The word-spotting task, as reviewed by McQueen (1996), appears tailor-made for the investigation of lexical segmentation. The task requires subjects to listen to (usually bisyllabic) nonword strings such as /mɪntəf/ and press a button if they detect a monosyllabic word (in this case *mint*) embedded at either the onset or the offset of the string. Since subjects are not told the identity of the word that they are trying to detect, the task resembles the problem that listeners face in trying to segment words in connected speech – where lexical items will be embedded in a longer stream of speech.

However, despite this resemblance between the word-spotting task and recognition in connected speech, the slow reaction times and high error rates suggest that this task is a difficult one for subjects to carry out. In some cases as many as 70% of trials result in an error where subjects fail to make a response to a stimulus containing an embedded word. Perhaps a more appropriate interpretation of this task is as a go/no-go version of the lexical decision task[1] in which subjects must decide the lexical status of all possible segmentations of a bisyllabic non-word into two words. In some cases, participants are told the location of the embedded word beforehand (i.e. whether the word is at the start or end of the nonword) thereby reducing the number of alternative segmentations to consider. However, making a response in word-spotting task is still likely to require multiple segmentations of the stimulus, followed by a lexical decision on each potential segmentation.

A necessary assumption in order to interpret reaction time data obtained with word-spotting is that it is the process of segmentation that produces differences in reaction time for different conditions. However, since comparisons of auditory lexical decision responses can be affected by choosing an inappropriate position from which to measure response times (see Goldinger, 1996b for further discussion) there may be an additional confound from the lexical decision component of the task. Furthermore, the task only provides simple measures of processing time (RT and error rate) and therefore ignores the temporal properties of the stimulus that subjects are responding to. This may make it difficult to interpret results in terms of properties of specific sections of the stimuli.

Experiments carried out by Cutler and Norris (1988) used the word-spotting task to investigate whether the lexical stress of a subsequent syllable affected the detection of a word embedded at the onset of a bisyllabic non-word. For CVCC words (such as *mint*) detection latencies were slower where the word was embedded in a bisyllable with a stressed first and second syllable (stimuli such as /mɪnteɪv/) than in words with a weak or unstressed second syllable (stimuli such as /mɪntəf/). Cutler and Norris interpret this effect as indicating that stressed syllables are used as a cue to the segmentation of connected speech and consequently that stimuli such as /mɪnteɪv/ are segmented into words as [mɪn][teɪv], producing slower latencies for the detection of the word *mint*.

This finding suggests that information coming in after the offset of a word assists recognition. However, with respect to potential acoustic cues to word length or word boundaries, this study is unable to provide any information about what sections of the speech stream played a role in aiding recognition. For instance there is a potential

[1] Thanks to Billi Randall for discussion of this interpretation of the word-spotting task.

confounding factor in the stimuli, whereby items that have a stressed second syllable and start with a voiceless stop will be aspirated in stressed syllables. This is of particular relevance, since Christie (1974) reports that the aspiration of voiceless stops in syllable initial position provides a strong cue to the detection of word boundaries. Thus it is unclear whether effects reported by Cutler and Norris (1988) as being caused by metrical stress of syllables of these stimuli reflect the operation of a metrical segmentation mechanism or instead arise through the introduction of allophonic variation that provides an acoustic cue to a word boundary.

Other studies have used word-spotting to show that competition from other lexical items has an inhibitory effect on the identification of monosyllabic words in longer strings. For example, studies by McQueen, Norris and Cutler (1994) in English and Vroomen, van Zon and de Gelder (1996) in Dutch, show that response times were significantly slower for detecting the word *mess* in the sequence /dəmɛs/ which is part of the word *domestic* than in the matched sequence /nəmɛs/ which can not be continued to form a word. This effect, they suggest, results from competition between the word *domestic* and the embedded word *mess*. Similar results were also obtained for words embedded at the onset of a lexical item; participants found it harder to detect the word *sack* in the sequence /sækrəf/ (part of the word *sacrifice*) than in the non-word sequence /sækrək/ – though this effect was only apparent by increased error rates, not by slower response times.

These findings provide evidence for effects of competition between lexical candidates that do not share word boundaries. This result is therefore cited in support of models of lexical segmentation that incorporate inhibitory connections between lexical items. However, effects of the lexical status of continuations of embedded words would also be predicted by the recurrent network models described in the previous chapter (see Figure 3.5 for example). It is therefore unclear that these results can mediate between different theories of lexical segmentation.

Furthermore, these inhibitory effects of competition have not been replicated in experiments using lexical decision. For instance, in experiments reported by Luce and Lyons (1999) it was found that lexical decision latencies to words that contained an embedded word were <u>faster</u> than to matched words that did not contain an onset-embedded word. Although these results suggest that onset-embedded words are activated during recognition – since the activation of embedded words facilitates lexical decision responses – these findings are contrary to the predictions of competition based models.

Word-spotting and lexical decision tasks have provided evidence that onset-embedded words are activated during the perception of longer words and that following context can influence their identification. However, such results fall short of the systematic comparison of the activation of onset-embedded words and longer competitors that would be required to arbitrate between lexical competition and recurrent network accounts. Furthermore, since these lexical decision and word-spotting experiments used short, bisyllabic stimuli they may underestimate the role of acoustic differences between embedded words and longer competitors during identification. Acoustic cues to word boundaries, such as the duration differences that were described in Chapter 2, may require more extended spoken contexts in order to be processed effectively. It is therefore possible that effects of these acoustic cues will only be apparent where full sentences can be used as experimental stimuli.

### 4.1.3. Cross-modal priming

One method of assessing the activation of competing interpretations of words within spoken sentences is through the cross-modal priming of lexical decision responses. As pioneered by Swinney, Onifer, Prather and Hirshkowitz (1979) this task has been used to assess the on-line activation of the different meanings of homophonous words like *bank*. By comparing lexical decision RTs to words that are related to the different meanings of *bank* (for example the target words RIVER and MONEY) following either the ambiguous test word or an unrelated control prime, Swinney was able to assess the degree to which different meanings were activated. Comparing the priming effect obtained at different positions in the speech stream allows investigation of the time course with which contextually appropriate meanings of homophonous words are selected.

In the literature on spoken word recognition, cross-modal priming has also been used to investigate the lexical access process for words that have clear meanings, but which may be temporarily ambiguous in the speech stream, such as cohort competitors like *cabin* and *cabbage* (Gaskell & Marslen-Wilson, in press; Marslen-Wilson, 1990; Zwitserlood, 1989; Zwitserlood & Schriefers, 1995). In priming experiments target words are commonly semantically and/or associatively related to different meanings of the target (Moss, Ostrin, Tyler & Marslen-Wilson, 1995; Shelton & Martin, 1992; Tanenhaus, Burgess &

Seidenberg, 1988). Where alternative interpretations are orthographically distinct lexical items it is possible simply to repeat the prime word as the target – see Tabossi (1996) and Zwitserlood (1996) for a review of different variants of the cross-modal priming task and Gaskell & Marslen-Wilson (submitted) for a direct comparison of cross-modal semantic and repetition priming of cohort competitors.

Experiments using cross-modal priming have demonstrated that words embedded at the offset of other words are activated during the recognition of connected speech. This finding has been inferred from the significant priming of words related to these offset-embedded words. For instance, Shillcock (1990) demonstrated significant priming of the target word RIB by sentences containing the word *trombone* (via the embedded word *bone*). This has been confirmed using single word presentations (Luce & Cluff, 1998; Vroomen & de Gelder, 1997) - though experiments using repetition priming have failed to replicate this finding (Marslen-Wilson et al., 1994). These results have been taken as evidence that non-aligned lexical hypotheses are activated during connected speech – a finding that would challenge accounts of lexical identification (such as sequential recognition accounts) in which only words that start at a known word onset are activated during recognition.

None of these experiments, however, measured the activation of words related to the longer word as well as the embedded word. Consequently, it is unclear whether significant priming indicates that the perceptual system fails to distinguish between the embedded word and its longer competitor (a finding no current account of spoken word recognition would predict) or merely that participants in these experiments become aware of the relationship between prime and target by some post-access strategic process (Shelton & Martin, 1992).

One study that did compare activations of both appropriate and inappropriate segmentations of potentially ambiguous stimuli was carried out by Gow and Gordon (1995). They compared the priming of associatively related targets (FLOWER or KISS) from phonemically identical sequences such as *tulips* and *two lips*. Results demonstrated priming of the target KISS from two word stimuli (*two lips*) though not from single word stimuli (*tulips*). Conversely, targets (such as FLOWER) related to the long word were primed by both single word (*tulips*) and two word (*two lips*) stimuli.

By comparison with the results of Shillcock (1990), this failure to find priming between the offset-embedded word in the auditory prime *tulips* and the target KISS is surprising. Most accounts of lexical segmentation predict that stimuli in which both syllables are words would produce more mis-segmentations than words such as *trombone* used in the Shillcock study. One interpretation of the discrepancy between these findings is that in experiments where both appropriate and inappropriate interpretations of the prime stimuli are probed (as was the case for the Gow and Gordon study), priming effects are more resistant to effects of strategic expectations by participants.

The failure to observe priming of offset-embedded words in the Gow and Gordon (1995) study is interpreted as evidence for sensitivity to acoustic cues that mark word onsets. However, since prime sentences in their experiments continued after the presentation of the visual targets, it is unclear whether information in the 700ms of following context that participants heard whilst making a response may also play a role in the priming effects obtained in these experiments. By allowing prime stimuli to continue after the presentation of the target, these studies fail to control the amount of information in the speech stream that can be processed by participants. Hence conclusions regarding the importance of one particular section of the speech stream may be questioned. Furthermore, since these experiments only measured the activation of words embedded at the offset of a longer word, they do not provide constraining evidence regarding the important issue introduced earlier in the chapter of whether the recognition system is biased towards short word hypotheses in the identification of onset-embedded words.

Experiments by Tabossi, Burani & Scott (1995) in Italian also investigated ambiguities created by words being embedded at the onset of longer words. They found equal priming for associates of the word *visite* (visit) from sequences containing that word and from sequences where *visite* was formed by sections of two adjacent words (as in the sequence *visi tediati* (faces bored)). This effect was also found in a subsequent experiment where an allophonic cue to the presence of a word boundary was present in these two word stimuli. These results confirm the findings of Gow and Gordon (1995) that lexical items made by combining two adjacent words are accessed in connected speech.

The results obtained by Tabossi and colleagues suggest that whatever acoustic cues to word boundaries may be present in Italian – and they may be more weakly marked than in English (Bertinetto, 1981) – do not allow the system to distinguish words created from concatenating two adjacent lexical items from a single lexical item. Priming of meanings related to words created from concatenated lexical items was even obtained in the case where an allophonic cue to a word boundary was present in their stimuli. However since, as in other studies, Tabossi et al. (1995) did not investigate the identification of short words that are embedded in longer lexical items – such as *visi* (faces) in *visite* (visit) it is unclear if alternative explanations of these results based on strategic effects can be ruled out.

Several studies have demonstrated ambiguity created by the absence of explicitly marked word boundaries in connected speech. However, the only study (Gow & Gordon, 1995) to measure the severity of this ambiguity (by comparing whether listeners are able to distinguish appropriate from inappropriate segmentations) found rather less ambiguity than other results might have predicted. However, this study focussed on words embedded at the offset of a longer word rather than the onset-embedded words that are more critical for the theoretical accounts under discussion here.

Consequently, there remains a conflict between work describing acoustic differences that might provide a means by which to discriminate short and long words and models that assume that onset-embedded words are ambiguous at their offset. This conflict remains unresolved since experiments have not compared the activation of short and long words that share the same onset during connected speech. For the same reason, it is not possible to draw strong conclusions regarding the differential predictions of recurrent network and lexical competition accounts regarding short word biases during the identification of onset-embedded words. Without comparison of the activation of short and long words during the processing of stimuli containing either short or long words it is unclear how a bias towards short word interpretations could be established.

## 4.2. Experimental design

In the various experiments reviewed here, a variety of techniques were used to investigate the processing of monosyllables embedded in longer words and of longer words that contain onset-embeddings. However, none of these studies have adequately investigated the time course with which embedded words are recognised in the speech stream. Consequently, although there is evidence showing the activation of longer competitors during the recognition of embedded words (and vice versa), there are few firm conclusions about how these competing lexical hypotheses are resolved on-line and whether acoustic cues that distinguish short from long words play a role in the identification of embedded words.

In order to compare the time course of identification of onset-embedded words and longer competitors in connected speech it is necessary to use methods in which sentences can be presented to participants. This rules out tasks such as word-spotting in which only bisyllabic stimuli can be used. Furthermore, to track the activation of alternative lexical hypotheses across precisely measured sections of speech, the standard form of the cross-modal priming task, in which stimuli continue after the presentation of the visual target, is also unsuitable. For this reason all the experiments reported in this thesis used sentence fragments, with stimuli being cut off at positions of interest in the speech stream. By comparing interpretations of short and long stimuli at specific points in the stimuli, it is possible to evaluate the extent to which different sources of information in the speech stream contribute to the identification of onset-embedded words and longer competitors. Concerns may be raised that by cutting off speech a cue to the location of a word boundary is provided (in the silence that follows the offset of the gated speech). However, since stimuli will be cut-off during a word as well as after its offset, participants who attempted to use such a cue would find it unhelpful.

In the standard form of the gating task, participants are presented with progressively longer fragments of speech and have to write down the words that they can identify at each gate. Comparing responses to stimuli containing short and long words allows investigation of the extent to which onset-embedded words create ambiguity between short and long lexical candidates. Investigating how listeners' interpretations change across gates enables measurement of how the recognition of these words is affected by different sources of information that are available in the speech stream.

Concerns have been raised about the effect of successive presentations of the same stimuli and the off-line nature of responses in gating (Cotton & Grosjean, 1984; Tyler & Wessels, 1985; Walley, Michela & Wood, 1995). Consequently results obtained in a gating study will be compared with experiments in which cross-modal priming is used to provide an on-line measure of lexical activation at each gate (Gaskell & Marslen-Wilson, 1997; Zwitserlood, 1989; Zwitserlood & Schriefers, 1995). Since the competing interpretations of embedded words are distinct lexical items, repetition priming of lexical decision can be used to provide a measure of the activation of the prime that avoids possible confounds that could be produced by differences in semantic or associative relatedness.

In order to conclude that listeners are able to detect subtle acoustic differences (such as syllable duration) between short and long word stimuli, any confounding factors must first be ruled out. The initial series of experiments reported in this thesis will therefore use stimuli that maximised the potential ambiguity between short and long words. This was achieved by using *lexical garden-paths*, stimuli in which speech coming after the offset of an embedded word continues to match a longer competitor – for example, the sequence *cap tucked* in which the onset of the following word matches the onset of the second syllable of the competitor *captain*. In this way, even allowing for co-articulation, syllables which can either be a monosyllabic word or the start of a longer word will be as acoustically similar as possible (except for the acoustic differences that were described in Chapter 2).

The activation of short and long words for these lexical garden-path sequences was compared with matched sentences including longer lexical items that contained these embedded words at their onset. It should therefore be possible to rule out strategic accounts of the priming effects observed and allow investigation of whether the on-line activation of words in connected speech is biased towards short words, as predicted by lexical competition accounts of spoken word recognition.

## 4.2.1. Stimuli

***Short and long word pairs***

Starting from the CELEX lexical database (Baayen, Pipenbrook & Guilikers, 1995) bisyllabic words were selected which had a morphologically unrelated monosyllable embedded at their onset (e.g. *captain*, containing the embedded word *cap*). Only bisyllables with a metrically stressed first syllable were chosen and in all cases the monosyllabic word exactly matched the syllabification of the longer word. Pairs such as *cat* and *cattle* were excluded, since by the maximal-onset principle (Selkirk, 1984) *cattle* would be syllabified as [kæ][tl̩] with a boundary within the embedded word. The monosyllables chosen all had at least three letters and consisted of three or more phonological segments.

Items were rejected if they were not of the same syntactic class, or if either word was orthographically unusual (such as the pair *pizza* and *peat*). However, items were not required to be fully orthographically embedded (pairs such as *track* and *tractor* were included as well as *captain* and *cap*[2]). A further criterion was that at the offset of the monosyllable there should be a limited number of longer items that contain the embedded word (this excluded items like *con* embedded in *concrete* where *con*- as a prefix is found in over 200 words). The mean number of words in which the monosyllables were embedded was 12 items (maximum, 43; minimum, 1). In all cases the long word was the most frequent word in this group. To avoid biases towards either short or long words in our test stimuli, pairs were rejected if they did not occur with approximately equal frequency in the language. Across the set of 40 pairs of words that were used in the experiments a paired t-test showed that there were no significant differences in the frequency of the pairs of short and long words (mean frequency short words = 35/million, long words = 25/million, t(39)= 1.07, p>.1).

***Test sentences***

Given that one goal of these experiments was to test for effects of acoustic cues (such as greater syllable duration in monosyllabic words) that may only be contrastive by comparison with prior context, items were placed approximately in the middle of test sentences. Each sentence contained an average of 6 syllables of neutral preceding context (range 3 to 11 syllables) so that subjects would be able to use ongoing prosodic cues that were present in these stimuli. Cloze tests were carried out on these sentence contexts to ensure neither of the target words were predictable. Each test sentence had several words after the test item, so that listeners would not be able to use prosodic cues to the end of a sentence as potential evidence of a short rather than a long word being present. No major clause boundaries followed the short test words to avoid possible intonation differences

[2] Auditory priming experiments by Jakimik, Cole and Rudnicky (1985) report differences in priming from a bisyllables to an onset-embedded monosyllable depending on whether the monosyllable was spelt in the same or different way. However, these effects were observed at an SOA of two seconds, much longer than the typical SOA used in cross-modal experiments. This suggests that strategic effects may have contributed to these results.

that might distinguish them from their longer competitors (Christophe, Guasti, Nespor, Dupoux & Ooyen, 1997).

In order to create as much ambiguity in these stimuli as possible and provide the most stringent test of the claim that there are acoustic cues that distinguish between short and long words in connected speech, it is necessary to exclude acoustic differences in the embedded syllables caused by co-articulation from following segments. Continuations for the short word stimuli were therefore chosen that started with the same onset segment or segments as the second syllable of the longer word. An example pair of sentences from the set of 40 used in the experiments are shown below (with target words emphasised):

(Short word) The soldier saluted the flag with his **cap** tucked under his arm.

(Long word) The soldier saluted the flag with his **captain** looking on.

The set of 40 sentence pairs shown in Appendix A were recorded by the author onto digital audio tape (DAT) in a sound-proof booth. Each pair of sentences was recorded successively to help ensure that the sentences were produced with near-identical intonation patterns and without prosodic breaks after the monosyllabic words. These recordings were then passed through an anti-aliasing filter and digitised at a sampling rate of 22khz using a DT2821 sound-card attached to a Dell PC.

### 4.2.2. Acoustic analysis and alignment points

In order to determine whether listeners are sensitive to duration differences between syllables in monosyllabic and bisyllabic words, it is important to ascertain whether these and other possible acoustic differences are present in these stimuli. Furthermore, given the intention to compare interpretations of stimuli that contain embedded words and longer competitors, it is necessary to make these contrasts between stimuli containing equivalent acoustic-phonetic information. Consequently, *alignment points* (hereafter *AP*) were set up at phonetically equivalent positions in each sentence. Measuring acoustic differences and differences in participants' interpretations with respect to these alignment points helps ensure that results reflect cues to the location of word boundaries and are not artefacts caused by information from subsequent segments or syllables in either set of stimuli.

The start and the end points of each sentence were marked using the BLISS speech editing system (Mertus, 1989). Additional markers were placed at the onset of the target word – a point at which each pair of sentences should be as identical as possible. This similarity was confirmed by listening to the sentence onsets and by visual inspection of the speech wave and fundamental frequency (F0) contours for a selection of the test items. Acoustic analysis of the duration of the word immediately preceding this marker (usually an article such as *the*) showed no reliable differences in duration (short word duration = 97ms, long word duration = 98ms; t(39)= 0.070; p>.1).

| Cursor positions | Measure | Short word stimulus | Long word stimulus | Difference |
|---|---|---|---|---|
| *onset* – $AP_1$ | Duration (ms) | 291 | 243 | * * |
| | Voicing time (ms) | 184 | 165 | * * |
| | F0 (hz[a]) | 112 | 113 | ns |
| | mean RMS[a] | 2476 | 2657 | (* ) |
| $AP_1 – AP_2$ | Duration (ms) | 79 | 77 | ns |
| $AP_2 – AP_3$ | Duration (ms) | 42 | 44 | ns |

**Table 4.1: Alignment points and acoustic measurements for stimuli in experiment 1. [a]F0 and RMS energy for voiced section of syllable only. Statistical significance: ns p>.1, (*) p<.1, ** p<.01**

The second alignment point ($AP_2$) was placed following the onset segment (or segments) of the second syllable. A paired t-test showed that there were no significant differences in the duration of onsets which were word initial in the short word stimuli and word medial in the long word stimuli (t(39)= 0.42, p>.1). This contrasts with the stimuli used by Gow and Gordon (1995), as well as with other findings in acoustic phonetics showing the significantly greater duration of word-initial segments (Klatt, 1976). This difference in the acoustic properties of our stimuli may be attributable either to the absence of prosodic boundaries before the onset-segments in our stimuli or to acoustic differences being obscured by subsequent phonetic differences.

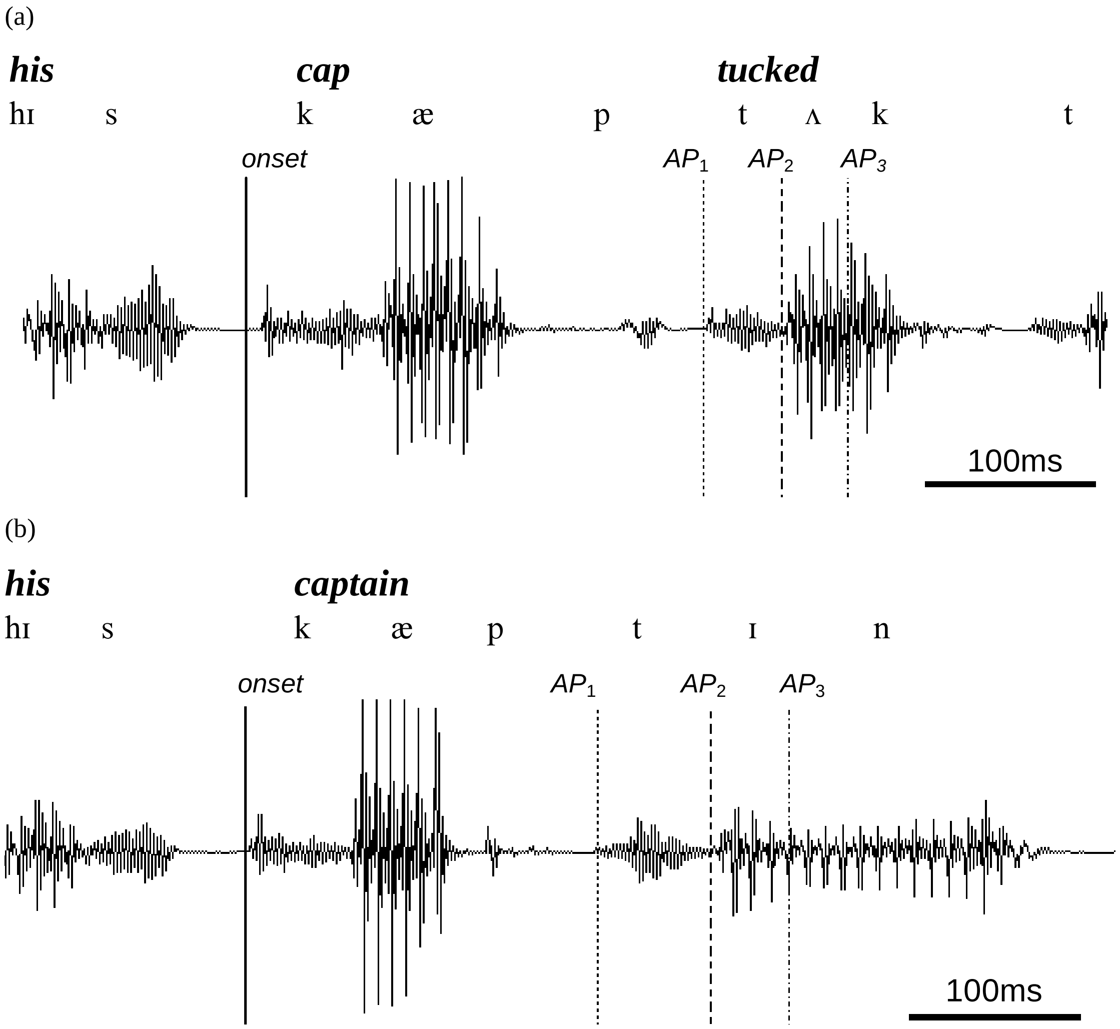


**Figure 4.1: Speech waveforms and alignment points for the stimuli in experiment 1. *onset* – onset of target word, *$AP_1$* – offset of target word, *$AP_2$* – onset of second syllable, *$AP_3$* – vowel of second syllable. Stimulus items are:**

**(a)** *"The soldier saluted the flag with **his cap tucked** under his arm."*
**(b)** *"The soldier saluted the flag with **his captain** looking on."*

The first alignment point ($AP_1$) was placed at the offset of the first syllable of the prime word (at the end of the closure of the final segment). Measurements of the acoustic duration of the first syllable (from the onset of the target word to $AP_1$) shows the expected difference in acoustic duration in monosyllabic and bisyllabic words, as shown in Table 4.1. This 48 ms difference in duration was highly significant across the 40 test items (t(39)= 9.35, p<.01). Further analysis confirmed that a large proportion of this difference could be accounted for by differences in the duration of the vowel. Analysis of the duration of the voiced portion of the target syllable showed reliable differences in duration (t(39)= 3.11, p<.01). However, this difference apart, by the criteria described by

Fear, Cutler & Butterfield (1995) there was no major difference in the amount of stress applied to these syllables. There were no significant differences in the fundamental frequency of the two syllables (t(39)= 0.53, p>.1) and measurements of the acoustic energy in the vowel of these syllables showed a minor difference in mean RMS energy. Syllables at the onset of long words had greater energy, suggesting they were more strongly stressed than the equivalent syllable as a monosyllable though this effect was marginal (t(39)= 2.01, p<.1).

The third alignment point ($AP_3$) marks the earliest location where the stimuli are expected to contain different phonemes. This marker was placed 4 pitch periods into the vowel of the second syllable, for instance 52ms into the vowel [ʌ] of *cap tucked*. Again, there were no overall differences in the duration of this section of vowel (t(39)= 1.35, p>.1). An example of the position of these alignment points is illustrated in Figure 4.1.

## 4.3. Experiment 1 – Gating

The first experiment carried out to test these hypotheses used the gating task reviewed previously. This method was used to assess the role of the acoustic differences described in

Table 4.1 in identifying stimuli that are potentially ambiguous between a bisyllabic word and an onset-embedded monosyllable. For instance, if listeners are sensitive to duration differences in the syllable /kæp/ for words like *cap* and *captain* it would be expected that responses to the pairs of stimuli will diverge at or before $AP_1$ – the offset of the first syllable. However, if sub-phonemic cues in the production of segments that are word onsets in short word stimuli and word medial in long word stimuli are of greater importance (as suggested by Gow and Gordon (1995) and Nakatani and Dukes (1977)) then responses will diverge at $AP_2$. Finally, if recognition requires phonemic mismatch between short and long stimuli then listeners would be unable to distinguish these paired stimuli prior to $AP_3$; this being the earliest point at which our experimental stimuli differ phonemically.

### 4.3.1. Method

***Participants***

Twenty-four English speakers from the Birkbeck Speech and Language subject pool were tested. Most were University of London students, all were aged between 18 and 45 and were paid for their participation. All were native speakers of British English and had normal hearing and no history of language impairment.

***Design and materials***

Experiment 1 used the standard gating method in which participants make written responses to successively longer fragments of recorded speech. For the stimulus sentences used, where more than one word needs to be identified, one set of dependent variables will be the word or words identified by participants following each fragment or gate. In addition, participants were asked to rate the confidence of their responses using a 9 point scale ranging from 1 (guess) to 9 (confident).

The independent variable was whether the sentence fragments played to the participants contained a short or a long word. Each sentence in the experiment was played out as 10 successive fragments with the entire sentence being presented from the start to a cut-off point. Cut-off points were the three alignment points described previously and shown in Figure 4.1, as well as two initial gates 50 and 100ms before $AP_1$ and five gates (designated gates 6 to 10) placed 50, 100, 200, 300 and 400ms after $AP_3$. In all cases, it was expected that the gate 400ms after $AP_3$ would contain sufficient information to enable participants to identify the word following the target item.

The 40 pairs of test sentences (containing a short or long word) were pseudo-randomly divided into two experimental versions such that each version contained only one member of each stimulus pair. An additional 20 sentences were added to each version; four were used as practice items to acquaint participants with the task and the remaining 16 items were fillers to distract from the large number of embedded words in the test sentences.

***Procedure***

Participants were tested in groups of between 2 and 4, sitting in booths in a quiet room. They were provided with answer books containing the onset of each sentence up to (but not including) the target word and were instructed to identify the word or words that continued each sentence based on the speech they heard at each gate. Participants were instructed to make a response for every fragment of speech that they heard and to write down as many words as they could hear in each continuation. They were also asked to provide a confidence rating on a 9 point scale with 1 being a guess and 9 representing confidence.

Sentences were played from a PC equipped with a DT2821 soundcard through closed-ear headphones. Fragments were played at 6 second intervals, with an extra 2 second being provided at gates after $AP_3$ to allow participants more time to write down words coming after the target item. The 56 test and filler sentences were divided into four blocks, each lasting approximately 20 minutes with a five minute break being given between blocks and a 10 minute break at the half way point. The whole testing session, including the practice items, lasted approximately two hours.

### 4.3.2. Results and discussion

Data from two participants (one from each version) were rejected for failing to comply with the instructions to make a response for each fragment they heard. The remaining 8800 responses (22 subjects, 40 items, 10 responses/item) were coded for the identity of the target word and subsequent words along with the confidence rating and analysed to investigate the proportion of responses (by participants and by items) that matched either of the target words.

All participants produced correct responses for the majority of the test items by the final gate. However, there were three items, (*ban*, *bran* and *win*) for which the short word stimuli were not recognised by 50% of participants at the final gate. Consequently these items (and their corresponding bisyllables, *bandage*, *brandy* and *winter*) were not analysed.

The proportion of responses at different gates that matched either the short or long target words are shown in Figure 4.2. As can be seen in the graph, at early gates (up to and including the offset of the first syllable at $AP_1$), the majority of participants' responses match the short target word (e.g. CAP). Even at the first gate, 100ms before the offset of the embedded word ($AP_1$) subjects hear enough of the target word to identify the first syllable.

These results also show that, following the large number of short word responses at early gates the proportion of responses that match the short word target decreases at $AP_2$. Since the isolation point as conventionally calculated from gating experiments is the average gate at which subjects produce the correct response and then don't change their response at subsequent gates, this U-shaped pattern of short word responses will produce a bi-modal distribution of isolation points for the short word stimuli. For approximately 65% of items and/or participants, isolation points will be at a gate after $AP_2$, while the remaining 35% of isolation points (where responses don't change subsequently) will be substantially earlier. Consequently, isolation points will be bi-modally distributed with measures of central tendency being unrepresentative of the behaviour of any given participant or item. Statistical analysis will therefore use the proportion of responses that matched the target words at each gate as the dependent measure.

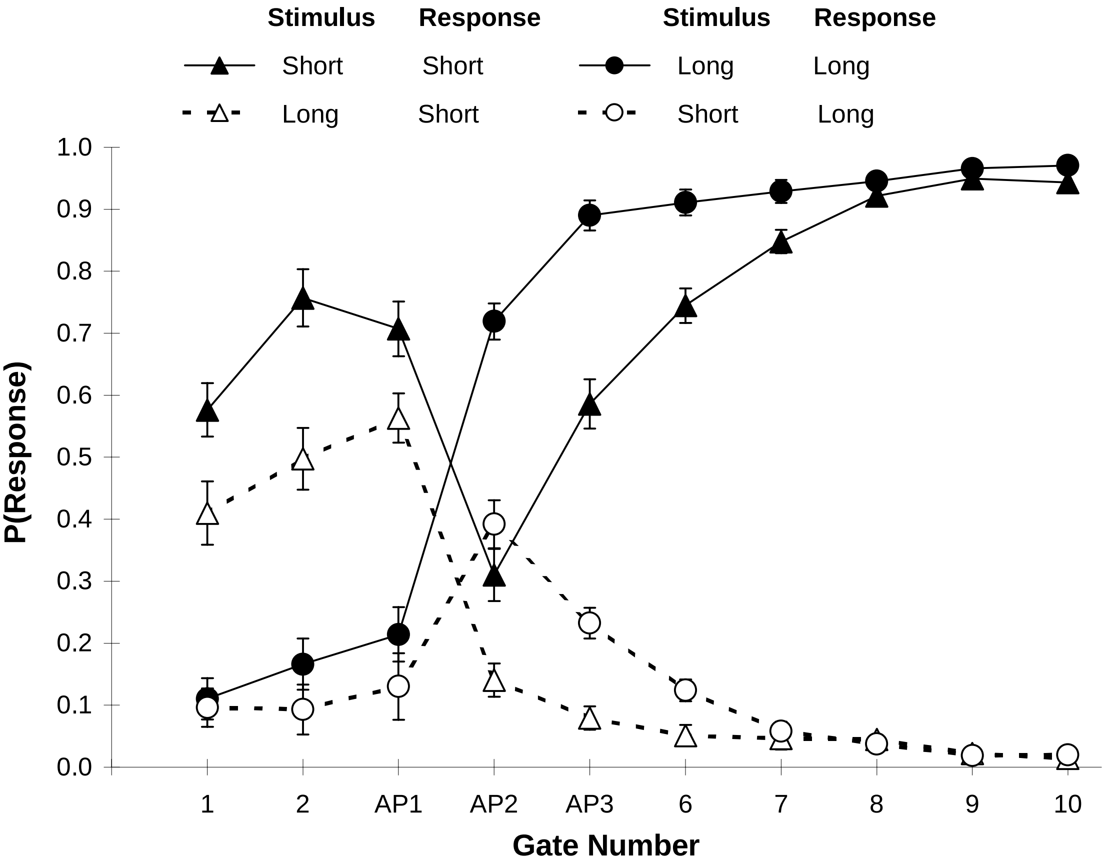


**Figure 4.2: Experiment 1 – Gating. Proportion of responses matching short and long target words for stimuli containing short and long words. Error bars are 1 standard error.**

### *Acoustic cues to word length*

Although at early gates there is an overall bias towards short word responses, there are differences in the proportion of short word responses made at early gates depending on which of the pair of stimuli participants were hearing (as shown in Figure 4.2). ANOVA on the proportion of short word responses across the three gates up to $AP_1$, using the repeated measures factors of stimulus type (short or long word) and gate number (gate 1, 2 or $AP_1$) shows that significantly more short word responses were made to short word stimuli than to long word stimuli ($\underline{F}_1$[1,20]= 60.32, p<.001; $\underline{F}_2$[1,35]= 26.86, p<.001). There was also a significant effect of gate ($\underline{F}_1$[2,40]= 30.37, p<.001; $\underline{F}_2$[2,70]= 9.60, p<.001) reflecting the greater number of short word responses at the later gates and an interaction between stimulus type and gate significant by participants and not items ($\underline{F}_1$[2,40]= 5.81, p<.01; $\underline{F}_2$[2,70]= 2.35, p>.1).

Parallel effects were found in the analyses of long word responses. Again over the first three gates there were significantly more long word response to long word stimuli than to short word stimuli ($\underline{F}_1$[1,20]= 7.34, p<.05; $\underline{F}_2$[1,35]= 4.69, p<.05). There was also a significant effect of gate, reflecting the increasing number of responses matching either target word across these three gates ($\underline{F}_1$[2,40]= 14.40; p<.001; $\underline{F}_2$[2,70]= 8.39, p<.001) and an interaction between stimulus type and gate – though this was of only marginal significance by items ($\underline{F}_1$[2,40]= 6.48, p<.01; $\underline{F}_2$[2,70]= 2.82, p<.1).

These effects of stimulus type suggest that subjects are able to use acoustic differences to discriminate the initial syllables of short and long words. The significant effects of gate show the sensitivity of the gating task to the arrival of new acoustic information, while interactions between stimulus type and gate suggest that acoustic information that allows subjects to discriminate between short and long words becomes more available throughout these gates.

Similar effects can also be observed in the confidence ratings. However, given the small number of long word responses at the first three gates, there were insufficient data points to analyse ratings from these responses. Ratings data for the first three gates were averaged for short word responses and are shown in Table 4.2. Analysis of these confidence ratings confirm the effect of stimulus type and gate shown in the analyses of word responses. Participants produced significantly higher confidence ratings for short word responses to short word stimuli than to long word stimuli ($\underline{F}_1$[1,20]= 14.56, p<.001; $\underline{F}_2$[1,27]= 9.86, p<.01). There was also a significant effect of gate number ($\underline{F}_1$[2,40]= 69.25, p<.001; $\underline{F}_2$[2,54]= 106.70, p<.001) and no significant interaction between these variables.

| | Confidence Ratings | | |
|---|---|---|---|
| Stimulus Type | Gate 1 | Gate 2 | $AP_1$ |
| Short Word | 3.21 | 4.73 | 5.82 |
| Long Word | 2.96 | 3.91 | 5.07 |

**Table 4.2: Mean confidence ratings for short word responses to short and long word stimuli at initial gates. 1 = guess, 9 = confident**

### *Delayed recognition and response biases*

Despite these differences, the recognition of embedded words still appears to be delayed compared to the identification of the longer words in which they are embedded. It is only at gate 8 that there is no significant difference between the proportion of correct responses given to short words and long word stimuli (t(36)= 0.96, p>.1). This delay in recognition appears to result from competition from longer interpretations, since at $AP_2$ (the onset of the second syllable) participants gave many more long word responses to the short word stimuli than at any previous gate. It is only when there is clear mismatch between the short word stimuli and the long target words (for instance the phonemic differences between *cap tucked* and *captain*) at $AP_3$ and beyond that subjects are able to revise these hypotheses and identify the short words. This result confirms the role of information coming after the offset of a word in identifying embedded words as suggested by the results of gating (Grosjean, 1985) and word-spotting experiments (Cutler & Norris, 1988).

However, it is also necessary to consider possible effects of response biases. In off-line gating experiments, participants may be biased towards producing the shortest single word that accounts for all the speech segments that they can hear in the current fragment (Tyler, 1984). Such a bias could account for two properties of this data. Firstly, it would explain the predominance of short word responses at the initial three gates. Since participants are hearing some or all of a single syllable such as [kæp], they will be inclined to produce monosyllabic words in response, even in cases where the acoustic duration of the syllable might suggest that it was more likely to come from a bisyllabic word. This bias may lead to under-estimating the effectiveness of the cue provided by syllable duration, since participants will produce fewer bisyllabic responses at early gates.

The second aspect of the results that could be interpreted in terms of this single word bias is the large increase in long word responses at $AP_2$. A bias towards responses that account for as much speech as possible would encourage participants to respond with a longer word (e.g. *captain*) for stimuli such as [kæpt]. Response biases might therefore increase the number of long word responses at $AP_2$ and hence exaggerate the amount of competition between short and long hypotheses – producing the delayed recognition observed for onset embedded words.

In summary, the results of Experiment 1 suggest that listeners are sensitive to acoustic cues that distinguish between syllables of short and long words. The significant differences before the offset of the first syllable confirm that subjects are able to discriminate short and long words before they diverge phonemically. This challenges the assumptions of models in which the onset of embedded words are assumed to be indistinguishable from the initial syllables of longer words.

Despite these acoustic differences, participants display an overall preference towards short word interpretations at early gates. At later gates, where continuations match longer words, long word interpretations are preferred. This appears consistent with models that incorporate lexical-level competition – with short and long words competing during identification even where they do not share word boundaries. Furthermore, the overall bias towards short word responses at early gates is as predicted by lexical competition models.

However, given the questions raised earlier about the possible role of response biases in determining gating responses, it is important to use more on-line methods to measure the activation of embedded words and longer competitors in connected speech. With this gating experiment as background, the next chapter therefore reports a series of experiments using an on-line task (cross-modal priming) to probe the activation of onset-embedded words and longer competitors during connected speech.

# 5. Cross-modal priming experiments with embedded words

Cross-modal priming of lexical decision is a well-established method for probing the activation of competing interpretations of lexically-ambiguous spoken sequences (Zwitserlood, 1989; Gow & Gordon, 1995; Tabossi, Burani & Scott, 1995; Gaskell & Marslen-Wilson, 1996). In constructing materials for use in priming tasks, a variety of different relations between prime and target have been utilised. Target words that are both semantically and associatively related to the prime are commonly used. For example, the pair *cat* and DOG[1] not only have a similar meaning, but are also associatively related since a majority of participants in a free-association task produce the word DOG for the cue word *cat* (Moss & Older, 1996).

However, since associated pairs can produce priming in the absence of any semantic relationship between prime and target (such as between the pair *pillar* and SOCIETY Moss, Hare, Day & Tyler, 1994) associative priming may reflect form-based associations between words that frequently co-occur. Thus associative priming need not require access to meaning in the same way as pure semantic priming appears to (see Moss et al., 1994 and Plaut, 1995, for networks that model the separate contribution of associative and semantic relationships). Recent research addressing issues in semantic representation and processing (e.g. Moss, Ostrin, Tyler & Marslen-Wilson, 1995; McRae, de Sa, & Seidenberg, 1997) has therefore used pure semantic priming without any associative relation between prime and target.

Cross-modal semantic (non-associative) priming may, however, be too weak to be used to assess the activation of competing interpretations of sentences (though see Moss & Marslen-Wilson, 1993). Since the goal of the experiments reported here is to assess the activation of competing interpretations of embedded words, the constraint of having to find pairs of words which both have strong associates would restrict the number of items

[1] Throughout this thesis I will follow the convention of listing prime stimuli in *italics* and target stimuli in CAPITALS.

that could be used. Consequently the priming experiments reported in this chapter all used repetition priming – where the visual target is identical to the auditory prime.

Questions have been raised about the susceptibility of repetition priming to form-based effects. However, these effects may be minimised by including non-word foils where an equivalent degree of phonological overlap is paired with a non-word target (e.g. *stumble*–STUMB). With appropriately constructed filler items, priming is not observed between phonologically related pairs such as *pillow*–PILL in cross-modal repetition priming with single word primes, though priming is observed for pairs that are morphologically and semantically related like *darkness*–DARK (Marslen-Wilson, Tyler, Waksler & Older, 1994). These results are interpreted as evidence that repetition priming is sensitive to lexical-level effects and thus may be used to assess the lexical activation of competing interpretations of ambiguous sequences. A further advantage of the use of repetition priming is that it removes a potential source of variance produced by differences in the strength of the semantic or associative relationships between primes and targets in different conditions.

By varying the point in the prime sentence at which the visual target is presented and by cutting off the spoken prime at the probe position, cross-modal priming allows the time course of lexical activation to be measured at different points during an utterance while controlling how much speech the listeners have heard (Gaskell & Marslen-Wilson, 1996; Zwitserlood, 1989). By using the same stimuli and cut-off points as in Experiment 1, results can be compared to those obtained in gating. However, since cross-modal priming produces an on-line measure of the lexical activation of competing interpretations, the results should be less affected by the response biases that were evident in the gating data.

A series of cross-modal priming experiments was therefore carried out, each investigating the activation of competing lexical hypotheses at one point in the speech stream. The initial experiment used stimuli cut-off at the offset of the first syllable of the critical words – alignment point 1 ($AP_1$) in Experiment 1.

## 5.1. Experiment 2a – $AP_1$

Results obtained in the gating task indicate that listeners were strongly biased towards short word interpretations of stimuli presented up to $AP_1$. On this basis, strong and significant priming of short words would be expected at this probe position, while priming of long words may be weaker or non-significant. The results of this experiment also suggest that acoustic differences between segments and syllables in short and long words can be used by listeners to assist in discriminating monosyllabic words from the onset of longer words in which they are embedded.

However, the off-line response task used in gating may be susceptible to systematic biases that could distort the data in unforeseen ways. Consequently, follow-up experiments using cross-modal repetition priming will provide converging evidence on the time course of identification of onset embedded words and longer competitors. This data will provide valuable evidence regarding the ambiguity of syllables in short and long words.

### 5.1.1. Method

***Participants***

Seventy four paid participants from the Birkbeck Centre for Speech and Language subject pool were tested on this experiment. None of the participants had taken part in Experiment 1.

***Design and materials***

The same 40 pairs of test sentences were used as in Experiment 1. In all cases, these were presented up to the marker placed at the offset of the initial syllable of the test words ($AP_1$ – see Section 4.2.2 for further details). To provide a baseline measure, lexical decision response times following an unrelated control prime were compared to responses following the test sentences. The control prime sentences were identical to the test sentences in all but the word at the probe position which was replaced with a contextually appropriate but unrelated prime; either a monosyllabic word matched in frequency to the short test word, or a bisyllable matched to the long word target (see Table 5.1).

| Prime Type | Prime Stimulus | Short Target | Long Target |
|---|---|---|---|
| Short Test | *The soldier saluted the flag with his* ***cap***$^{a}$ *tucked under his arm.* | CAP | CAPTAIN |
| Long Test | *The soldier saluted the flag with his* ***cap***$^{a}$***tain*** *looking on.* | CAP | CAPTAIN |
| Short Control | *The soldier saluted the flag with his* ***palm***$^{a}$ *facing forwards.* | CAP | CAPTAIN |
| Long Control | *The soldier saluted the flag with his* ***rif***$^{a}$***le*** *by his side.* | CAP | CAPTAIN |

**Table 5.1: Prime and target stimuli for experiment 2a. a = probe position for Experiment 2a.**

This produced an experimental design with four prime types (two test prime and two control prime conditions). The additional two sets of 40 control prime sentences were recorded at the same time as the original test items from the gating experiment to minimise possible differences in voice quality or prosodic structure across the four sentences for each item. The four test and control primes were paired with either of the two targets, producing a 40 item, eight condition experiment as shown for an example set of four sentences in Table 5.1. The 320 test trials were rotated into eight experimental versions, such that each subject heard only one version of each sentence. Test version was included as a variable in subsequent analyses to reduce estimates of random error. In the analysis of participant means this referred to the test version to which each participant was assigned. In the items analysis this referred to the number of the item group sharing the same assignment of conditions to test versions.

In addition to the 40 test items in each experimental version 80 filler sentences were interspersed with the test items. Of these fillers, 20 were followed by a word target that was phonologically and semantically unrelated to the prime sentence. A further 20 filler sentences were followed by a non-word target that was phonologically similar to the word at the probe position in the prime sentences – these were added to discourage participants from associating phonological overlap between prime and target with a ‘yes’ response. The remaining 40 fillers were followed by non-word targets that were unrelated to the prime stimulus.

Also included in the experiment were 20 practice items and 10 lead-in items to allow participants to settle into each experimental block. This produced 150 sentences in each version (50% followed by word targets and 50% by non-words). Of the 150 sentences in the experiment, 20 were test sentences where a phonologically related word target followed the auditory prime. This produced a relatedness proportion of 13% over the entire test set for a given participant. To encourage participants to attend to the auditory prime sentences a recognition test on some of the filler sentences was given at the end of the experimental sessions.

### ***Procedure***

Participants were tested on one of the eight experimental versions in groups of one to four seated in booths in a quiet room. They were warned that they would be given a memory test on the auditory stimuli after the main experiment, but also informed that they should not attempt to rehearse or memorise the sentences. Each version of the experiment was split into four sessions, first a block of 20 practice sentences, after which subjects were given feedback on their lexical decision reaction times and errors. This was followed by two test blocks, each starting with five lead-in items for which no data was recorded, followed by 60 sentences, with a short interval between blocks. A pencil and paper recognition test was given after the end of the second test block.

Each trial started with a sentence fragment being played over headphones. At the first alignment point (or an equivalent position in the control sentences and at a range of positions in the filler sentences) the speech was cut off and a word presented for 200ms on the monitor in front of each participant. Participants were required to press the ‘yes’ button with their dominant hand if the target was a real English word or the ‘no’ button if it was not. Reaction times were measured from the onset of the target word (corresponding to the offset of the prime stimulus) with a 3 second time out. Following the presentation of each target (and the participant’s response), there was a short pause before the start of the next trial when the procedure was repeated. Each test session, including practice items, lasted approximately 25 minutes.

At the end of the lexical decision experiment, participants were given a sheet listing 25 sentences, 12 of which were filler sentences from the experiment. Participants were instructed to circle any sentences which seemed familiar, even if they had not heard all of the sentence. There was no time limit on this task, though most completed it within 5 minutes.

### 5.1.2. Results

Of the 74 participants, 9 were excluded for slow or error-prone lexical decision responses (excluding those participants whose mean test and control RT was greater than 750ms and/or produced more than 12.5% errors). One test target (BRAN) produced a large number of errors (over 30%) and consistently slow reaction times (over 750ms) and was therefore removed from the analysis, along with the matched bisyllable (BRANDY). Also excluded were 4 outlying responses over 1200ms each coming from a different participant in response to a different item. These discarded response times and response times from trials in which participants responded incorrectly were treated as missing data points and played no part in the following analyses. Following these exclusions, mean reaction times and error rates in each condition were as shown in Table 5.2.

| | | Short Target (CAP) | | Long Target (CAPTAIN) | |
|---|---|---|---|---|---|
| Prime Type | Prime Word | RT (ms) | Error (%) | RT (ms) | Error (%) |
| Short Test | *cap* | 485 | 3.2 | 539 | 4.8 |
| Long Test | *captain* | 501 | 2.8 | 528 | 6.4 |
| Short Control | *palm* | 512 | 3.8 | 561 | 4.3 |
| Long Control | *rifle* | 512 | 2.8 | 557 | 6.0 |

**Table 5.2: Experiment 2a. Mean response times and error rates by prime and target type.**

A three-way repeated measures ANOVA was carried out to investigate effects of the length of the word from which the prime syllable was taken (short/long), prime type (test/control) and length of the target word (short/long). An additional between-groups factor of version or item group (in the participants and items analysis respectively) was included to reduce estimates of random variation, though effects involving this factor will not be reported. ANOVA on response times showed a main effect of prime type (test/control) ($\underline{F}_1$[1,57]= 33.34, p<.001; $\underline{F}_2$[1,31]= 26.33, p<.001) indicating significantly faster responses following related test primes. There was also a significant main effect of the length of the target word ($\underline{F}_1$[1,57]= 80.98, p<.001; $\underline{F}_2$[1,31]= 30.48, p<.001), with significantly faster lexical decision responses to shorter target words. There were no main effects involving the number of syllables in the primes.

Even though prime sentences were cut off at the offset of a single syllable such as [kæp], there was no difference in the amount of priming for words that exactly matched the prime (*cap*) compared to longer items in which the prime was embedded (*captain*) (see Figure 5.1). There was no interaction between prime type and target length ($\underline{F}_1$[1,57]= 1.09, p>.1; $\underline{F}_2$[1,31]= 1.48, p>.1). This is a very different pattern to that observed in the gating experiment. At $AP_1$ in Experiment 1 subjects produced many more short word than long word responses. The lack of an equivalent effect here shows that cross-modal priming may be less susceptible to the single word bias that was observed in Experiment 1, suggesting that these results provide a purer measure of lexical activation than those obtained in gating.

Perhaps the most crucial result in this experiment is the significant interaction between the number of syllables in the prime and the number of syllables in the target ($\underline{F}_1$[1,57]= 7.89, p<.01; $\underline{F}_2$[1,31]= 5.55, p<.05). Lexical decision responses were faster when the number of syllables in the target matched the number of syllables in the prime. Inspection of the condition means shown in Table 5.2 suggests that this effect is only to be found for the test prime condition. The three-way interaction between prime type, prime syllables and target syllables was significant by items but not by participants ($\underline{F}_1$[1,57]= 2.15,p>.1; $\underline{F}_2$[1,31]= 5.03, p<.05).

The critical interaction between prime length and target length was examined by carrying out pairwise comparisons between response times following test and control primes. To simplify these contrasts, data was collapsed over the two control prime conditions since these had been found not to differ. Following the guidelines provided by Toothaker (1991), pairwise comparisons were carried out using one-way repeated-measures ANOVAs including the non-repeated factors of version and item group in the participants and items analysis, respectively. Since only four comparisons are required for this experiment (comparing responses to each target following test and control primes) no correction is required to control family-wise error. The magnitude of these differences and the significance of these pairwise comparisons is plotted in Figure 5.1.

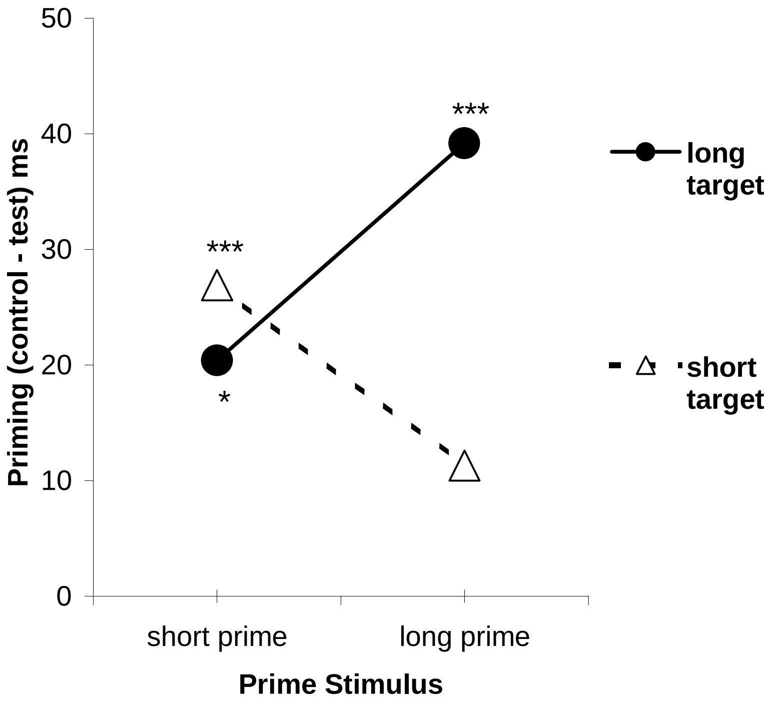


**Figure 5.1: Experiment 2a – Magnitude and significance of repetition priming by prime and target type. *** priming p<.001, * priming p<.05**

Strongest priming in this experiment was observed where the prime syllable comes from the same word as the target. The initial syllable of a long word significantly speeded responses to a long word target ($\underline{F}_1$[1,57]= 20.09, p<.001; $\underline{F}_2$[1,31]= 24.30, p<.001) but not to a short word target ($\underline{F}_1$[1,57]= 2.61, p>.1; $\underline{F}_2$[1,31]= 2.33, p>.1). Short word primes significantly facilitate responses to a short word target ($\underline{F}_1$[1,57]= 15.13, p<.001; $\underline{F}_2$[1,31]= 23.23, p<.001) with numerically weaker, though still significant, priming of responses to long words ($\underline{F}_1$[1,57]= 6.57, p<.05; $\underline{F}_2$[1,31]= 4.31, p<.05).

The interaction between prime and target length is confirmed by an analysis using the difference between response times following test and control primes as the dependent variable with the length of the prime word and the length of the target word as independent variables. This confirmed that there was no overall difference in priming of short or long targets ($\underline{F}_1<1$; $\underline{F}_2[1,31]= 1.71$, $p>.1$) or from short or long primes ($\underline{F}_1<1$; $\underline{F}_2<1$). There was however a significant crossover interaction between the factors of target length and prime length ($\underline{F}_1[1,57]= 9.14$, $p<.01$; $\underline{F}_2[1,31]= 8.76$, $p<.01$).

Error rates were arcsine transformed to stabilise variances (Winer, 1971) and then entered into the same three-way ANOVA as the response time data. There was a marginal main effect of target length ($\underline{F}_1[1,57]= 7.2$, $p<.01$; $\underline{F}_2[1,31]= 3.99$, $p<.1$) reflecting fewer lexical decision errors for short targets and a trend towards a two way interaction between prime length and target length ($\underline{F}_1[1,57]= 3.17$, $p<.1$; $\underline{F}_2[1,31]= 2.48$, $p>.1$). No other effects approached significance in this analysis.

## 5.1.3. Discussion

The first important result that emerges from these analyses is that cross-modal repetition priming suggests that the lexical activation of competing hypotheses is less biased towards short words than was suggested by the results of Experiment 1. In the gating study the majority of responses to stimuli presented up to $AP_1$ matched the short word. However, no such preference for short word interpretations is shown in the priming data. This result presents problems for lexical competition models such as TRACE, which predict an initial bias towards short word interpretations during the identification of onset-embedded words.

The second important point, confirming the pattern of results in gating, is that these results provide evidence that the perceptual system is sensitive to acoustic differences between syllables from short and long words. The cross-over interaction shown in Figure 5.1 demonstrates that significantly greater priming is found where the prime syllable comes from the same word as the target. Recall that this difference is observed for prime stimuli cut off at $AP_1$ – where participants can only hear the first syllable of the test words.

In previous discussions of lexical accounts of word segmentation (McQueen, Cutler, Briscoe & Norris, 1995) it has been argued that the presence of large numbers of onset embedded words places an important constraint on the structure of the spoken word recognition system. This argument, based on an assumption that word recognition proceeds from a phonemic or syllabic representation of the speech input, is that since onset-embedded words are ambiguous at their offset, following context must be used for their identification. This necessitates models of spoken word recognition that incorporate lexical-level competition. The experimental results reported so far suggest that the perceptual system is able to use sub-phonemic cues to distinguish onset-embedded words from the start of longer words in which they are embedded. To the extent that this is generally the case, processes such as delayed recognition and mechanisms of lexical competition may play a less important role in spoken word recognition than would be argued in accounts based on a purely phonemic analysis of the speech stream.

In the light of this finding another result from the gating study can now be re-examined – namely the evidence for competition between short and long words after $AP_1$ (the offset of the first syllable). Given the absence of biases towards short word interpretations in the cross-modal priming results at $AP_1$, it might be expected that effects of competition at later probe positions – which could also be attributed to response biases in gating - would be reduced in cross-modal priming. In three further experiments the cross-modal priming task was used to examine the lexical activation of potentially competing short and long target words at three further probe positions.

## 5.2. Experiment 2b-d

These subsequent experiments were set up in a similar fashion to Experiment 2a except that the probe position was advanced through the stimuli across the three experiments. In this way it is possible to track the activation of competing interpretations as increasing amounts of speech is presented to participants. One further difference between these experiments and Experiment 2a is that, since there were no significant differences between the two control primes used in Experiment 2a, only one control prime condition was used for each target type. This reduces the number of experimental versions required to test the lexical activation of each target word at a given probe position.

The probe positions used in these experiments were set up to facilitate comparisons with the gating study. Experiment 2b and c tested responses at the two later alignment points ($AP_2$ and $AP_3$ respectively) used in Experiment 1. The probe position in Experiment 2b ($AP_2$) marks the point at which the continuation of the embedded word becomes available to listeners (for example, the /t/ in *captain* and *cap tucked*). In the gating experiment this information produces a marked increase in the proportion of long word responses to both short and long word stimuli. Consequently, it might be predicted that there would be an increase in the amount of priming observed for long word stimuli in Experiment 2b. Given the results obtained at $AP_2$ in Experiment 1, this increase may be observed for both short and long word prime stimuli. However, given the potential for acoustic cues that mark word onsets (Gow & Gordon, 1995; Nakatani & Dukes, 1977) the magnitude of this effect may depend on whether these segments occur at the onset of a word or not.

The third alignment point ($AP_3$), tested in Experiment 2c, is placed in the vowel of the second syllable. $AP_3$ marks the point where the stimuli used in these experiments diverge phonemically. In Experiment 1 there is a corresponding divergence in the responses made at this point in the stimuli. For long word stimuli listeners continue to produce more long word responses, whereas for the short word stimuli, short word responses increase between $AP_2$ and $AP_3$. Consequently, less ambiguity would be predicted in the priming effects at this probe position. However, since the cross-modal priming results in Experiment 2a already suggest less ambiguity than was predicted from gating, changes at this probe position may be less marked.

The final probe position ($AP_4$), used in Experiment 2d, is placed 100ms after $AP_3$ (equivalent to gate 7 in Experiment 1). This was chosen to be a place at which the majority of responses in the gating experiment correctly identified the target word for both types of stimuli. Consequently, no facilitation of targets that do not match the prime words would be predicted at this probe position.

### *Participants*

Across the three experiments, 181 paid participants from the same population used previously were tested (54 on Experiment 2b, 72 on Experiment 2c, 55 on Experiment 2d[2]). None of these had taken part in any of the previous experiments.

[2] Differences between the number of participants tested in these experiments reflect differences in the amount of prior experience groups had had with the lexical decision task, and hence how many participants were rejected for slow and error prone responses.

### *Design and materials*

The design and materials used in these three experiments were identical to the previous experiments using the same sets of 40 items. The prime stimuli were presented up to $AP_2$ in Experiment 2b, up to $AP_3$ in Experiment 2c and up to a point 100ms after $AP_3$ in Experiment 2d. As in the previous experiment short and long target words were visually presented at the point at which the speech was cut off, with participants making a yes/no lexical decision response to the target word.

| Prime Type | Prime Stimulus | Short Target | Long Target |
|---|---|---|---|
| Short Test | *The soldier saluted the flag with his* ***cap t$^{b}$u$^{c}$ck$^{d}$ed*** *under his arm* | CAP | CAPTAIN |
| Long Test | *The soldier saluted the flag with his* ***capt$^{b}$ai$^{c}$n$^{d}$*** *looking on* | CAP | CAPTAIN |
| Short Control | *The soldier saluted the flag with his* ***palm*** *facing forwards* | CAP | – |
| Long Control | *The soldier saluted the flag with his* ***rifle*** *by his side* | – | CAPTAIN |

**Table 5.3: Prime stimuli and target stimuli for experiments 2b-d with approximate probe positions marked for the test stimuli.**

The only significant divergence from Experiment 2a was in the number of control prime conditions used. Previously there were two separate control prime conditions, each matched in length and frequency to one of the pair of target words. In Experiment 2a both control primes were used with each target word. Since no significant differences were found between these two control primes, the design of each experiment was reduced to include only one control prime for each target type. These were chosen to be matched in length and frequency to each target (i.e., short control primes were used for short targets and long control primes were used for long targets). This produced three experiments each with six conditions (three prime types and two target types) as shown in Table 5.3.

As in Experiment 2a, related non-word fillers were added to ensure that form overlap between prime and target was not associated with a ‘yes’ response. The same set of 20 items were used in Experiment 2b, with an additional seven related non-word fillers used in Experiment 2c and 2d. Unrelated trials were also added to each experimental version, 20 with word targets and 40 with non-word targets for Experiment 2b. An additional 35 unrelated trials were added in Experiments 2c and 2d, 21 with word targets, 14 with non-word targets. This produced experimental versions where the overall proportion of trials that included a related test item was 18% in Experiment 2b and just over 14% in Experiment 2c and 2d.

### *Procedure*

The procedure for each of Experiment 2b-d was identical to that used previously except that test stimuli were presented up to the end of the onset segments of the second syllable of the test words ($AP_2$) in Experiment 2b, up to the vowel of the second syllable ($AP_3$) in Experiment 2c and up to a point 100ms after $AP_3$ in Experiment 2d ($AP_4$). Targets were visually presented at the offset of the auditory prime. See Section 4.2.2 in Chapter 4 for further details of the alignment points used.

### *Analysis*

Results from this series of three experiments were analysed following data trimming as carried out for Experiment 2a. Participants were rejected for slow or error prone responses (mean test and control RT greater than 750ms and/or more than 12.5% errors on responses to test words). As previously, the target word BRAN produced consistently slow and error prone responses and was removed (along with the bisyllable BRANDY) from the analysis of all three experiments. Outlying response times were removed by excluding data-points above a response time cut-off set by examination of an RT histogram for each experiment.

Our goal in these experiments was to use the magnitude of priming as a measure of the lexical activation of competing interpretations. Consequently, analyses of overall reaction times and error rates are relegated to Appendix B in order to focus on statistical analyses that directly investigate the priming effects obtained in these experiments. Pairwise comparisons of response times and error rates following test and control primes were used to evaluate the magnitude and significance of priming effects. As in Experiment 2a these comparisons use repeated measures ANOVAs including a between groups factor of version (in analysis by participants) and item-group (in analysis by items). Comparisons between priming effects obtained for different primes and targets will be made using RT differences following test and control primes as the dependent measure (see Monsell & Hirsh (1998) for a similar approach to the statistical analysis of priming experiments).

## 5.2.1. Results of Experiment 2b – $AP_2$

Of the 54 participants tested, five were rejected for slow and/or error prone lexical decision responses. Two outlying data-points over 1400ms were also removed. Mean response times and error rates for the remaining subjects and items are given in Table 5.4. Analysis of variance on this data is reported in Appendix B.

| | | Short Target (CAP) | | Long Target (CAPTAIN) | |
|---|---|---|---|---|---|
| Prime Type | Prime Word | RT (ms) | Error (%) | RT (ms) | Error (%) |
| Short Test | *cap* | 549 | 2.2 | 584 | 4.2 |
| Long Test | *captain* | 556 | 2.6 | 552 | 1.5 |
| Control | *palm/rifle* | 570 | 3.0 | 607 | 7.5 |

**Table 5.4: Experiment 2b. Mean response times and error rates by prime and target type.**

Pairwise comparisons of response times following test and control primes showed that short word targets were significantly facilitated by short primes ($F_1$[1,43]= 5.92, p<.05; $F_2$[1,33]= 4.46, p<.05) with no significant facilitation by long primes ($F_1$[1,43]= 2.53, p>.1; $F_2$[1,33]= 3.92, p<.1). Conversely, long word targets were significantly primed both by long word prime stimuli items ($F_1$[1,43]= 22.01, p<.001; $F_2$[1,33]= 26.93, p<.001) and by short word stimuli ($F_1$[1,43]= 4.35, p<.05; $F_2$[1,33]= 4.12, p<.05). The magnitude and significance of these priming effects is illustrated in Figure 5.2.

Analysis on test-control difference scores with factors of prime type (short vs long prime stimuli) and target type (short vs long target words) showed more priming by long word primes than short word primes, as indicated by a significant main effect of prime type in this analysis ($F_1$[1,43]= 5.73, p<.05; $F_2$[1,33]= 5.77, p<.05). There was also a marginally significant main effect of target type ($F_1$[1,43]= 3.54, p<.1; $F_2$[1,33]= 3.49, p<.1) reflecting a tendency for long word targets to be primed more strongly overall than short words. The interaction between prime and target type (see Figure 5.2) was also significant ($F_1$[1,43]= 7.63, p<.01; $F_2$[1,33]= 7.17, p<.05) suggesting that, despite the conflicting information coming from continuations of the short word stimuli, greater facilitation is observed where the prime stimulus is identical to the target.

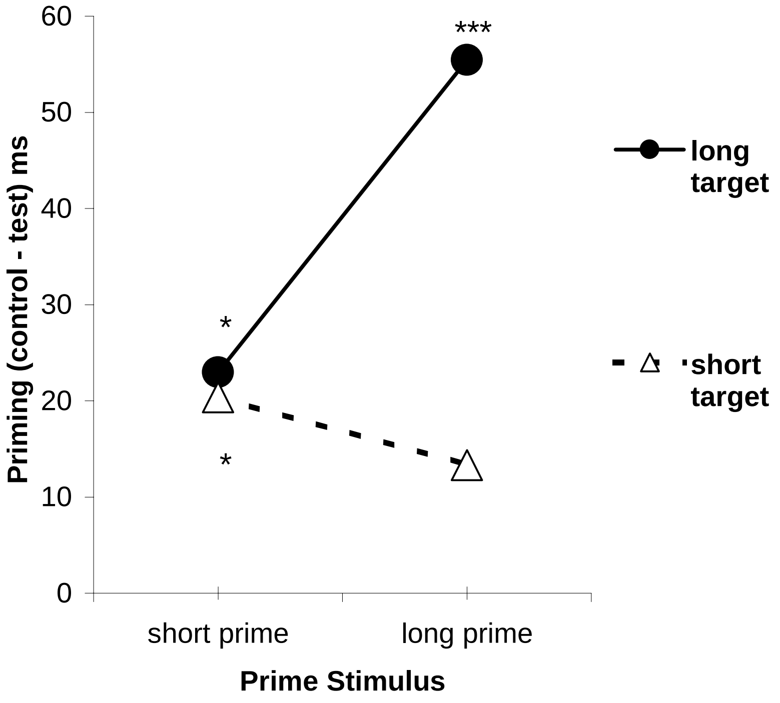


**Figure 5.2: Experiment 2b – Magnitude and significance of repetition priming by prime and target type. *** priming p<.001, * priming p<.05**

Pairwise comparisons of error rates also showed significant facilitation of responses to long words. Participants made significantly fewer errors to long targets when they were preceded by a long word prime ($F_1$[1,43]= 19.66, p<.001; $F_2$[1,33]= 9.92, p<.01) compared to error rates following control primes. There was also a marginal reduction in error rate when bisyllabic targets followed monosyllabic primes in the items analysis ($F_1$[1,43]= 2,49, p>.1; $F_2$[1,33]= 3.56, p<.1). There were no significant differences in error rates to monosyllabic targets following either short or long control primes (all p>.1).

At the probe position tested in this experiment, participants started to hear the onset of the syllable following an embedded word (the /t/ segment in stimuli such as *cap tucked* or *captain*). Priming effects at this probe position suggest that this information plays a role in the identification of the long word stimuli – indicated by significantly greater priming from long word stimuli and for long word targets. This pattern of results for short word primes suggests that these stimuli are more ambiguous than long word primes at this probe position. This question of how garden-path continuations of embedded words affect the activation of short and long competitors will be considered further in subsequent experiments.

### 5.2.2. Results of Experiment 2c – $AP_3$

Out of 72 participants tested, 14 were discarded for slow or error prone responses by the same criteria used previously. Data from an additional participant whose mean response times were more than two standard deviations faster than any other participant were also removed. Also excluded were 12 individual outlying responses slower than 1350ms. Mean response times and error rates following these exclusions are shown in Table 5.5 with ANOVAs on these data in Appendix B.

| | | Short Target (CAP) | | Long Target (CAPTAIN) | |
|---|---|---|---|---|---|
| Prime Type | Prime Word | RT (ms) | Error (%) | RT (ms) | Error (%) |
| Short Test | *cap* | 521 | 3.2 | 565 | 4.2 |
| Long Test | *captain* | 548 | 2.0 | 536 | 3.3 |
| Control | *palm/rifle* | 552 | 3.1 | 593 | 8.3 |

**Table 5.5: Experiment 2c. Mean response times and error rates by prime and target type.**

Pairwise analysis of priming effects at this probe position indicates relatively little change between this and the previous probe position, although there is an increasingly clear separation of the priming effects elicited by short and long primes (see Figure 5.3). Short primes significantly speed responses to short word targets ($F_1$[1,51]= 14.47, p<.001; $F_2$[1,33]= 10.80, p<.01) with numerically similar but statistically weaker effects for long targets ($F_1$[1,51]= 5.36, p<.05; $F_2$[1,33]= 3.33, p<.1). This indicates that there is still some ambiguity present in these short word stimuli – both competing interpretations can be primed at this probe position. For the long word primes there is strong priming of long word targets ($F_1$[1,51]= 36.40, p<.001; $F_2$[1,33]= 24.27, p<.001) but no evidence of significant facilitation of short word targets ($F_1$<1; $F_2$<1). Long word primes are clearly less ambiguous than short word primes at this probe position.

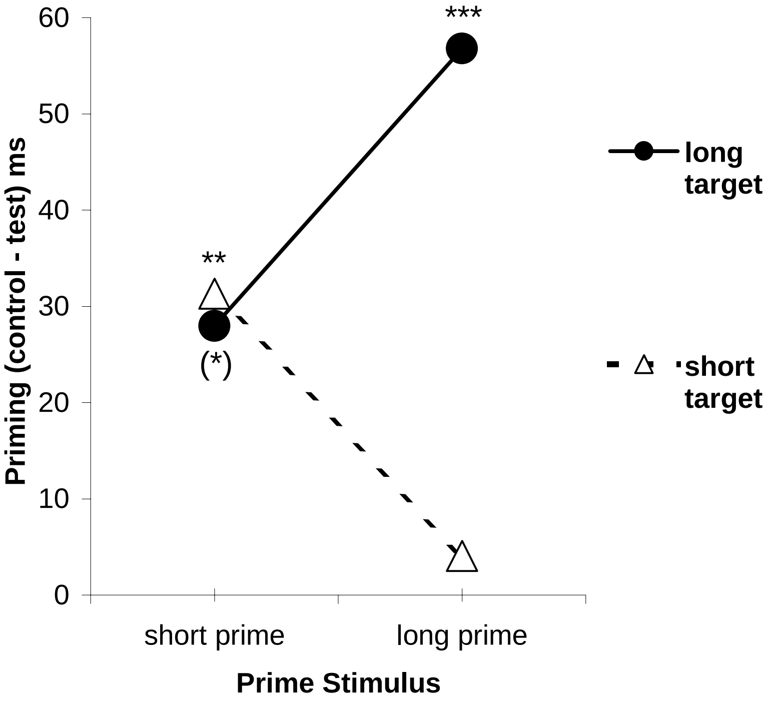


**Figure 5.3: Experiment 2c - Magnitude and significance of repetition priming by prime and target type. *** priming p<.001, ** priming p<.01, (*) priming p<.1**

A similar pattern is obtained in comparisons of error rates for long word targets following test and control primes. Participants made significantly fewer errors to long words after hearing long test primes than following control primes ($F_1$[1,51]= 7.92, p<0.01; $F_2$[1,33]= 10.91, p<0.01). Error rates were also reduced for long words following short word primes compared to controls ($F_1$[1,51]= 4.84, p<0.05; $F_2$[1,33]= 4.21 p<0.05). No significant differences in error rates were found for short word targets.

Analysis of response time differences between test and control primes shows a pattern more similar to that obtained in Experiment 2b than would be predicted on the basis of the gating data. The magnitude of priming showed a significant interaction between prime and target type ($\underline{F}_1$[1,51]= 15.84, p<.001; $\underline{F}_2$[1,33]= 19.57, p<.001) such that greater facilitation was observed where prime and targets matched. As in Experiment 2b there was a main effect of target type though this was non-significant by items ($\underline{F}_1$[1,51]= 4.81, p<.05; $\underline{F}_2$[1,33]= 2.58, p>.1). The main effect of prime type was non-significant ($\underline{F}_1$<1; $\underline{F}_2$<1).

This pattern of results, where long word stimuli only prime long word targets, again suggests that the perceptual system can distinguish bisyllabic from onset-embedded monosyllables at this probe position. This is consistent with the high proportion of correct responses to long word stimuli at this alignment point in the gating experiment. More surprising is the continuing ambiguity of the short word stimuli, with significant priming of long as well as short word targets. At $AP_3$ in Experiment 1, more participants responded with the short target word than the long target word. Given that the prime stimuli differ phonemically at $AP_3$ (in the vowel of the second syllable) it would have been expected that a clear preference for short word interpretations of short word stimuli would be observed at this probe position.

This discrepancy between the results obtained in gating and cross-modal priming may simply reflect the greater time available to subjects for the processing of stimuli in the gating task. However, it is necessary to rule out the possibility that there is some systematic difference in the measures of lexical activation obtained for short and long words in the cross-modal priming experiments. It is therefore of interest to compare the priming of short and long words from short word stimuli at a probe position where it was expected that these stimuli would be unambiguous at $AP_4$ - 100ms beyond $AP_3$ (equivalent to gate 7 in Experiment 1).

### 5.2.3. Results of Experiment 2d – $AP_4$

Out of 55 participants tested on the six versions of this experiment, eight were rejected for slow and/or error prone responses. Also removed were two data-points over 1400ms. Mean response times and error rates are shown in Table 5.6 with ANOVAs by prime and target type reported in Appendix B.

| | | **Short Target**<br>(CAP) | | **Long Target**<br>(CAPTAIN) | |
|---|---|---|---|---|---|
| **Prime**<br>**Type** | **Prime**<br>**Word** | **RT**<br>(ms) | **Error**<br>(%) | **RT**<br>(ms) | **Error**<br>(%) |
| Short Test | *cap* | 510 | 3.2 | 577 | 7.9 |
| Long Test | *captain* | 543 | 5.9 | 540 | 4.3 |
| Control | *palm/rifle* | 541 | 5.7 | 586 | 6.8 |

**Table 5.6: Experiment 2d. Mean response times and error rates by prime and target type.**

Priming effects were analysed by planned comparisons illustrated in Figure 5.4 showing that responses to short targets were significantly speeded by short primes ($F_1$[1,41]= 17.47, p<.001; $F_2$[1,33]= 11.65, p<.01) but not by long primes ($F_1$<1; $F_2$<1). Similarly long targets were significantly primed by long primes ($F_1$[1,41]= 18.05, p<.001; $F_2$[1,33]= 17.03, p<.001) but not by short primes ($F_1$[1,41]= 1.13, p>.1; $F_2$[1,33]= 1.83, p>.1). For this experiment, comparison of error rates following test and control primes did not shown any significant differences.

As can be seen in Figure 5.4, there is a cross-over interaction of priming effects by prime and target length in this experiment. ANOVAs using the difference between response times following test and control primes as the dependent variable showed no main effects of either prime length ($\underline{F}_1$<1; $\underline{F}_2$<1) or of target length ($\underline{F}_1$[1,41]= 1.32, p>.1; $\underline{F}_2$[1,33]= 1.63, p>.1) with a highly significant interaction between these factors ($\underline{F}_1$[1,41]= 27.59, p<.001; $\underline{F}_2$[1,33]= 30.08, p<.001).

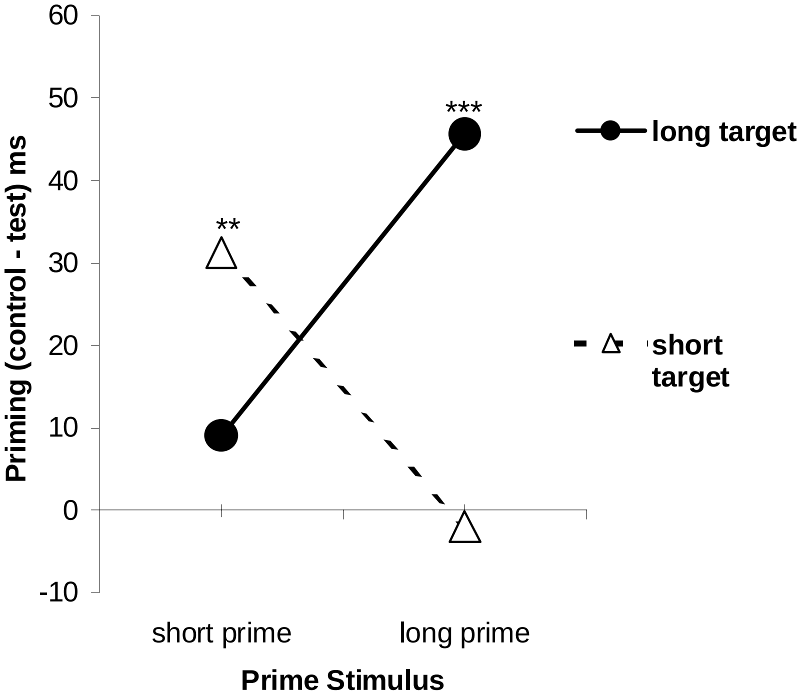


**Figure 5.4: Experiment 2d – Magnitude and significance of repetition priming by prime and target type. *** priming p<.001, ** priming p<.01**

The results obtained in experiment 2d indicate that reliable priming is only observed where prime and target stimuli are identical – irrespective of whether the prime is a short or a long word. Although priming effects at earlier probe positions suggested that short embedded words had been ruled out as interpretations of long word stimuli, it is only at this probe position that there is sufficient mismatch between short stimuli and long words for the perceptual system to rule out longer lexical hypotheses. Since stimuli presented up to $AP_3$ incorporated sufficient mismatch to allow a majority of participants to successfully identify short word stimuli in Experiment 1, it appears that the gating task is more sensitive to effects of mismatch between short word stimuli and long word candidates than cross-modal priming. Cross-modal priming required an additional 100ms of mismatching information in the speech presented after $AP_3$ to produce unambiguous priming of short words in the absence of any facilitation of long word targets.

## 5.2.4. Combined analysis of Experiments 2a-d

In order to investigate whether priming effects changed across the four probe positions, an analysis was carried out on data combined from all four experiments. To aid this comparison, reaction times were normalised by participants (dividing individual RTs by the mean response time for all (non-filler) targets for that participant and multiplying by the mean response time over all participants in the four experiments). Differences between responses following test and control primes were calculated for these normalised RTs and are shown in Figure 5.5. These difference scores were also entered into a three-way ANOVA with the factors of prime type, target type and probe position. Probe position is coded as a within groups factor in the items analyses and a between groups factor in the analysis by participants.

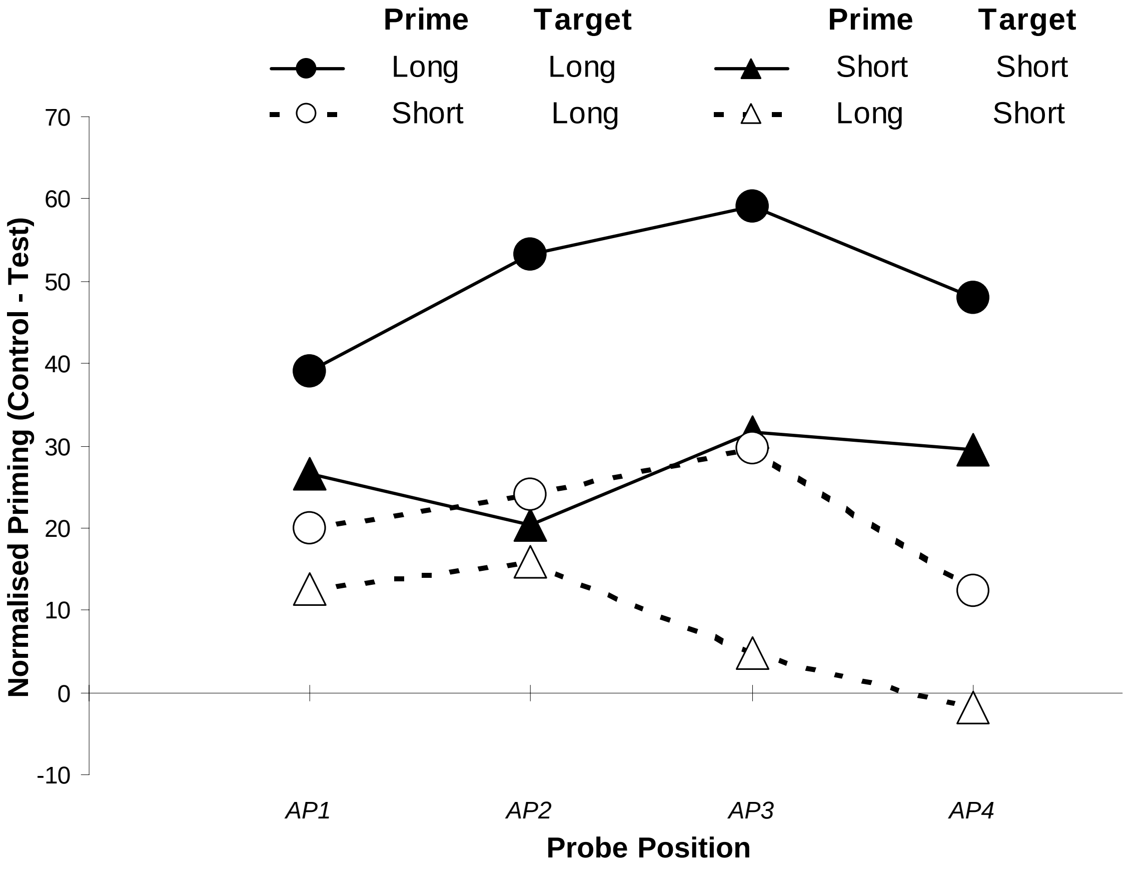


**Figure 5.5: Normalised magnitude of priming in Experiments 2a-d by prime and target type.**

There was a main effect of target type ($F_1$[1,214]= 7.44, p<.01; $F_2$[1,38]= 6.53, p<.05) indicating that overall more priming was observed for long targets. One possible explanation, which will be explored in later experiments is that short word stimuli with continuations that match long targets increased the activation of long words, especially at later probe positions (*AP*2, *AP*3 and *AP*4). However, the effect of target length did not interact with probe position ($F_1$<1; $F_2$<1).

The combined analysis also revealed a significant interaction between prime and target length ($F_1$[1,214]= 51.47, p<.001; $F_2$[1,38]= 56.87, p<.001) reflecting the pattern observed in each of the experiments (see Figure 5.1 to 5.4) for priming to be strongest between prime and targets of the same length. This effect did not interact with probe position ($F_1$[3,214]= 1.23, p>.1; $F_2$[3,114]= 1.52, p>.1) indicating that information to distinguish short from long stimuli was present at all four probe positions.

## 5.3. General discussion

The overall results of Experiment 2 primarily confirm one of the main results of the gating study; namely that the perceptual system can distinguish between monosyllabic words and the onset of a bisyllable even at the offset of the first syllable ($AP_1$). One striking consequence is that in none of the four priming experiments is significant priming of short word targets from long word primes observed. It can therefore be concluded that differences in the acoustic form of syllables from short and long words can directly affect relative levels of lexical activation for the short and long target words. This indicates that the presence of embedded words in our long word stimuli is not producing significant levels of ambiguity at the levels of lexical representation tapped into by the repetition priming task.

A second finding of these experiments is that, unlike gating, cross-modal priming does not produce a bias towards short word interpretations of lexical items that contain an onset-embedded word. This result appears to challenge the account of lexical segmentation and lexical access provided by models such as TRACE that through their implementation of lexical competition predict greater activation of lexical units representing short words.

With reference to the two empirical predictions invoked in arguments from the presence of onset-embedded words to the necessity of lexical competition models, the experiments reported in this chapter present a considerable challenge on both of these points. The implications of these results will be discussed in turn.

### 5.3.1. Acoustic cues to word length

Our results indicate that some acoustic difference between short and long word stimuli allows the perceptual system to discriminate between onset-embedded words and the start of longer competitors. As reviewed in Chapter 2, a variety of acoustic cues have been proposed that might be able to account for this result. The two most strongly attested cues are qualitative changes in the initial segments of words compared to segments that are in the middle or at the offset of words (Lehiste, 1960; Nakatani & Dukes, 1977), and differences in segment and syllable duration between monosyllables and longer words (Klatt, 1976; Lehiste, 1972; Nakatani & Schaffer, 1978).

The results of experiments carried out by Gow and Gordon (1995) have been used to argue that qualitative differences in onset segments are used to distinguish otherwise ambiguous stimuli such as *tulips* and *two lips*. Gow and Gordon conclude that these onset segments contribute more strongly to the processes of lexical access and segmentation than other sections of a word (see Gow, Melvold & Manuel (1996) for more details of this ‘Good Start’ model). However, without directly manipulating these cues while controlling other aspects of the stimuli, it is unsafe to conclude that these onset-cues (and not other differences in their stimuli) are responsible for their results.

Similar caution is also required in drawing conclusions from the results of the current experiments. The stimuli used have not been directly manipulated to include only controlled acoustic differences. However, since the methods used to present speech to participants do allow control of which sections of the stimuli can be heard in a particular experimental condition, it is possible to determine when in the speech stream the relevant acoustic differences can be found. Since the participants in these experiments were able to rule out embedded words when hearing the onset of a longer word – when a marked word onset had not been presented and would not be expected – we can conclude that the onset cues described by Gow and Gordon (1995) are unlikely to be responsible for the pattern of results obtained in these experiments. It is also worth noting that differences in the duration of word onsets were not statistically reliable for the stimuli used in these experiments.

The acoustic cue that is most likely to be available to participants is the duration difference observed between syllables in short and long test words. If listeners are able to detect these differences in syllable duration and use them as a cue to the location of word boundaries the early discrimination of short and long words that was observed in our experiments might result. There is already evidence suggesting that listeners have ready access to duration information in other aspects of speech perception. Experiments have demonstrated that changes in syllable duration can induce changes in voice-onset time boundaries for the perception of voiced and voiceless stop consonants (Miller & Lieberman, 1979; Miller, Green & Reeves, 1986; Volatis & Miller, 1992; Kessinger & Blumenstein, 1998) while rate dependant information has been shown to be important in the perception of time compressed speech (Foulke & Sticht, 1969; Dupoux & Green, 1997; Pallier, Sebastian-Galles, Dupoux, Christophe & Mehler, 1998) Follow-up experiments that directly manipulate the duration of segments and syllables in short and long words are required in order to establish that duration, and not some previously unconsidered cue, is responsible for the discrimination of stimuli in short and long words.

### 5.3.2. Lexical competition and short word biases

Having taken these acoustic differences into account, a further conclusion that is supported by these results is that cross-modal priming does not show the same bias towards short word hypotheses as was observed at early gates in Experiment 1. Indeed, the combined analysis of the four experiments shows a significant main effect of target type indicating that more priming was observed for long targets than for short word targets. Consequently, the time course of identification predicted by lexical competition models (i.e., that where two words are activated in the speech stream, short words will be more strongly activated) was not supported by the cross-modal priming experiments reported here. None of the four cross-modal priming experiments reported in this chapter showed greater priming for short words than for long words.

Although these results may favour recurrent network accounts that predict approximately equal activation for competing words irrespective of length, without directly simulating the pattern of results produced in these experiments it is not possible to interpret this data as supporting models without direct inter-lexical competition. Consequently, in the next chapter the recurrent network model developed in Chapter 3 was extended to account for the processing of acoustic cues to word length. Only by directly simulating the time course of identification of embedded words and longer competitors can the results of these priming experiments be considered to support recurrent network accounts.

# 6. Acoustic cues to word length in recurrent networks

Simulations described in Chapter 3 demonstrated that a recurrent network trained to activate a representation of all the words in a sequence provides an account of the delayed recognition of onset-embedded words. Following training, the network activates all lexical units that match the current input, with the degree of activation representing the probability of each word given the current input. Where the input matches both a complete word and the start of a longer word the network will activate each item equally. Only where following context rules out the longer lexical item will the network fully activate the embedded word.

This behavioural profile (illustrated in Figure 3.4 and 3.5) contrasts with that predicted by models that incorporate lexical-level competition. Lexical competition models such as TRACE (McClelland & Elman, 1986) produce a bias towards short word interpretations at the offset of an embedded word. This is due to the greater number of competitors that are present for longer words. This short word bias benefits embedded words since it allows them to inhibit longer competitors and thus be recognised more easily.

As was described in the previous chapters, both of these behavioural profiles are inadequate as an account of the time course with which embedded words are identified. The results of gating and cross-modal priming experiments demonstrate that short and long words can be distinguished before the offset of the first syllable of the embedded word. Therefore, since neither the lexical competition models nor the recurrent network simulations in Chapter 3 incorporate any input cue that would serve to distinguish between short and long words, both accounts are at present insufficient to account for the experimental data presented in Chapters 4 and 5.

An important goal of the modelling reported in this chapter is to produce a computational account of these experimental data. At present, recurrent network accounts of spoken word recognition are incomplete with respect to the processing of cues to word length and word boundaries. Simulating the effect of these cues on the identification of embedded words not only allows evaluation of whether recurrent networks are sufficient as an account of the processing of onset-embedded words, but will also allow testable predictions for future experiments to be generated from the model.

## 6.1. Acoustic cues to word length

The current hypothesis was that differences in segment and syllable duration provide the acoustic cue to word boundaries required to account for the experimental data. As described in the review of the acoustic-phonetics literature presented in Chapter 2, the increased duration of syllables in monosyllabic words has been reliably reported (Klatt, 1976; Lehiste, 1972) and there is experimental data supporting the use of duration as a cue to the perception of word boundaries (Nakatani & Schaffer, 1978). Furthermore, as shown in Table 4.1 in Chapter 4, differences in duration between syllables such as /kæp/ in words like *cap* and *captain* are present in the stimuli used in the current series of experiments.

Sensitivity to differences in syllable duration would therefore provide an account of listeners' ability to discriminate between short and long words before the onset of the following word – as shown by differences between responses to short and long word stimuli at $AP_1$ both in gating (Experiment 1) and in cross-modal priming (Experiment 2a). Thus, although the critical experiments directly manipulating duration have yet to be carried out, current evidence suggests that the discrimination of short and long words may involve the detection of differences in segment and syllable duration.

### 6.1.1. The perception of duration differences

An important constraint in modelling the perception of changes to segment and syllable duration in the recognition of onset-embedded words is that duration can not be used as a deterministic cue to whether a word is monosyllabic or bisyllabic. Despite the significant differences between the duration of the syllables in short and long word stimuli (p<.001 by a paired t-test), as shown by the histogram in Figure 6.1 there is considerable overlap in the distribution of durations for syllables in short and long words.

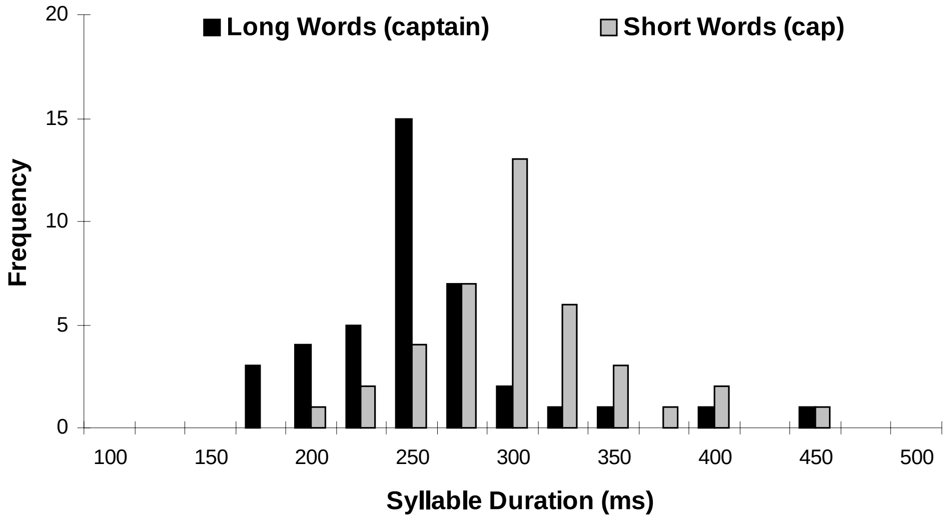


**Figure 6.1: Histogram of the duration of target syllables in short and long test items**

Such overlap will be even more marked in comparisons of naturally occurring speech. Significant differences in syllable duration in short and long words are only found where additional sources of variance can be controlled for. These may be caused by measuring syllables produced by different speakers, at different positions in an utterance, and containing different constituent segments and associated stress (Klatt, 1976). Without compensating for these additional sources of variance it will not be possible to use syllable duration to distinguish short from long words (see for instance the discussion between Crystal and House (1990) and Anderson and Port (1994) regarding the reliability of duration as a cue to word boundaries in English).

From a computational perspective, this implies that it will not be possible to partition syllables into those coming from short and long words by setting a simple duration threshold. Instead, some additional process will be required to adapt to the ongoing speech stream so that information from the spoken context can be used to set an appropriate boundary for distinguishing syllables in short and long words. This process is required to allow the system to compensate for other sources of variation in syllable duration in order to identify syllables as coming from a short or a long word.

In modelling this adaptive process, the simulations reported in this chapter focus on an extreme form of ambiguity – the case where two syllables with identical durations in fact come from words of different length. Syllables such as these can still be used to distinguish between short and long words if they are produced in spoken contexts that lead the listener to expect a syllable of a particular duration. For instance, if a sequence is produced at a fast speech rate, a syllable from a monosyllabic word would be produced with a relatively short duration. A syllable with an identical duration but produced in a slower sentence, on the other hand, would be more likely to have come from a bisyllabic word.

Such cases – embedded syllables in short and long words occurring with the same duration – only arose for a small number of the experimental stimuli. This is due to the test syllables being controlled for many potential sources of variation in syllable duration. These tightly controlled stimuli therefore provided less sources of variation that could lead to syllables in short and long words being produced with identical durations. However, for more naturally produced speech, in which these sources of variation are uncontrolled it is likely that the duration of syllables in short and long words will need to be disambiguated by the contexts in which they are produced.

In simulating the processing of duration as a cue to the placement of word boundaries, the models developed in this chapter used, as a test case, stimuli in which identical durations were produced for syllables in short and long words. This will allow investigation of the processes by which compensation for contextual changes in syllable duration are applied during recognition. These simulations focus on just one of the possible variables that can affect duration in such a way as to make syllables ambiguous – a change in the overall rate at which speech is produced in sequences.

## 6.2. Simulation 3 – Processing syllable duration in fast and slow sequences

The adaptive processing of acoustic input is an important aspect of the perception of connected speech. A system that can compensate for differences in the acoustic properties of speech produced by different speakers at different rates is also required to account for the perception of time-compressed speech (Foulke & Sticht, 1969; Dupoux & Green, 1997; Pallier, Sebastian-Galles, Dupoux, Christophe, & Mehler, 1998). The simulations reported here will investigate whether the recurrent network architecture that was described in Chapter 3 is able to adaptively process an input cue analogous to changes in duration in different rate sequences. This will require the network to process identical words differently, depending on the rate at which the preceding words in an utterance are presented.

### 6.2.1. Representing duration information in connectionist networks

In order to investigate the processing of duration in recurrent networks a decision must be made about how duration is to be represented. Since these simulations use an artificial language coded as discrete segments there is considerable freedom to represent duration in a form that makes the network's task as easy as possible. However, it is also important that the representation should not make unrealistic assumptions about the information available in the speech stream.

Previous computational models have coded for duration information by duplicating segments over many time steps in the input (Abu-Bakar & Chater, 1995; Gupta & Mozer, 1993). This may incorporate the unrealistic assumption that longer segments will have identical spectral properties to shorter segments but extended over more time steps. This simplifying assumption may help processing in the network since it creates greater similarity between identical sequences presented at different rates. However, coding for duration in this way is also computationally expensive since it requires extended training and testing sets to include these duplicated segments. Furthermore, in order to incorporate effects of preceding context, the network must encode information over more intervening segments. It was consequently decided to use a more computationally tractable coding scheme for segment and syllable durations.

The method of coding for duration that is most tractable for the network is to use input units that (separate from all other inputs) provide information about the duration of the current segment or syllable. This representation assumes that duration can be coded in an equivalent way to any other aspect of the speech input. Indeed, from the networks' point of view, these additional units could represent any source of information that helps distinguish between short and long words – not just differences in duration. Since these network investigations are intended to simulate the results of Experiment 2 (where duration differences were only one of several possible acoustic cues that could distinguish between short and long words) such a representation may be easier to justify in this context. In the remains of this chapter, however, the input representation will be set using an assumption that these units are coding for duration. This will allow us to make clear predictions regarding the computational mechanisms that may be required in simulating the affect of duration cues on lexical activation.

The question then becomes how to represent duration at this input. Exploratory simulations added a single input unit which represented duration by its activation (i.e. high activation for long syllables, low activation for short syllables). However, these simulations proved unsuccessful since the small changes in activation produced in this single unit were swamped by the binary inputs representing phonetic features.

Consequently a binary representation of duration over a block of three units was used in the network. These provided for three duration codes along with an additional input unit that was active whenever duration information was inappropriate for the current input segment (i.e. the unit was set to zero in the gaps between sequences and in certain syllable positions in later simulations). These codes allow a context-dependent representation of the duration of syllables to be used. This provides a simple simulation of the overlapping distributions of syllable duration that may result from changes in the overall rate at which a sequence is generated.

| **Syllable** | | | **Speech Rate** | |
|---|---|---|---|---|
| **Duration** | **Code** | **Network Input** | **Fast** | **Slow** |
| no duration | 0 | 0 0 0 | - | - |
| short | 1 | 1 1 1 | bisyllable | - |
| medium | 2 | 0 1 1 | monosyllable | bisyllable |
| long | 3 | 0 0 1 | - | monosyllable |

**Table 6.1: Network representation of syllable duration in Simulations 3 and 4.**

The activation of the three input units in coding for three values of duration (plus the ‘no duration’ input) is shown in Table 6.1. As can be seen, the longest syllable duration (code 3) is only used for monosyllabic words, while the shortest duration (code 1) is only used to represent bisyllables. However, the intermediate duration (code 2) can be used for either a monosyllabic or a bisyllabic word depending on the rate at which the sequence is being produced. Thus, for a sequence produced at a ‘slow rate’, the longest duration (code

3) will be reserved for monosyllables and code 2 will represent bisyllables. For sequences produced at the ‘fast rate’ code 2 will represent monosyllables and code 1 will be used to represent the duration of syllables in bisyllabic words. Thus code 2 will be ambiguous; depending on the overall ‘rate’ at which the sequence is presented; syllables coded with this duration can either come from short (monosyllabic) or long (bisyllabic) words.

For each sequence of words generated in the training set a speech rate was selected at random. This determined which two of the three duration codes would be presented to the network for that sequence of words. Since syllables coded as duration 2 are ambiguous, the networks will need to use the duration associated with previous words in the current sequence to determine the overall speech rate for that sequence. For onset-embedded words, where the identity of segments does not distinguish between short and long words, the network should be able to use duration to contribute to the identification of a word with an ambiguous syllable (such as /kæp/ in *cap* and *captain*). The network’s sensitivity to the duration cue was tested by comparing the activation of an onset-embedded word occurring at the start of a sequence (where duration code 2 will be ambiguous and will not provide any cue to discriminate short from long words) with the same input occurring as the second word of a sequence (where prior context should allow the network to determine the overall ‘rate’ of that sequence, and disambiguate the duration code).

### 6.2.2. Network architecture and training set

The architecture used for the initial simulations was identical to that used in Simulation 2 reported in Chapter 3 (see Figure 3.7) except for the addition of 3 input units and prediction outputs to represent the duration codes shown in Table 6.1. These simulations therefore used a simple recurrent network with 9 input units, 50 hidden units copied back to 50 context units and 29 output units. As in the simulations reported previously, the lexical output units had no bias weights although these were retained for the auto-encoder outputs. The architecture of the network and a snapshot during training is shown in Figure 6.2.

Training sets for these simulations were constructed using the 20 word vocabulary shown in Table 3.2 coded using the distributed phonetic feature representation shown in Table 3.1. One difference between this and previous simulations was that for each sequence of between 2 and 4 words one of the two speech rates described in the previous section was chosen at random. This determines which two of the three duration values are used to code for short and long words in that sequence. In the initial set of simulations, all segments in all syllables were coded for duration. This training regime is illustrated in Figure 6.2 where the network is being trained on the sequence "*lid topknot*" at the fast rate. The monosyllable *lid* is therefore presented with a duration of 2 while the prediction output for the first segment of the bisyllable *topknot* is being trained to predict a syllable with duration 1.

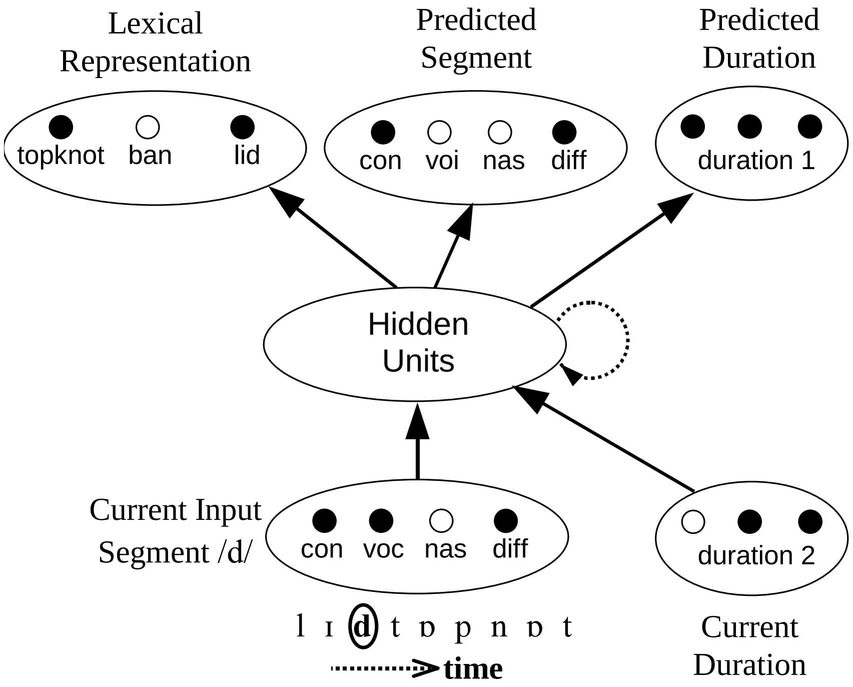


**Figure 6.2: A snapshot of the SRN during training for the sequence "*lid topknot*" at the fast rate in simulation 3. Three additional input units and prediction units represent the duration of the current syllable and the predicted duration of the subsequent segment.**

Preliminary simulations showed that the network was too reliant on duration information. For instance, since there was only one bisyllabic word (*topknot*) that began with the segment /t/, for sequences where the unambiguous duration code was used (code 1), the effective uniqueness point for the network was at the initial segment of the word. This inappropriate performance is apparent because of the small number of vocabulary items in the network's training set. However, this unrealistic aspect of the simulation is caused by the network using duration to rule out segmentally appropriate hypotheses. This suggests that the duration code is too reliable to simulate the properties of the speech stream, reflecting the discrepancy between simulations in which only one variable affects syllable duration and the more realistic case in which multiple, unreliable variables interact to determine the duration of segments and syllables in the input. In order to approximate the statistical properties of connected speech more closely, the duration cue was made less reliable in subsequent simulations. This was achieved by replacing 20% of the words coded with unambiguous durations (codes 1 and 3) with the ambiguous duration (code 2).

### 6.2.3. Results

Ten networks were trained using different random seeds for generating the training sequences and different sets of random initial weights. Lexical activations for onset-embedded words and competitors (both presented with the ambiguous duration code 2) were recorded and averaged over 10 fully trained networks. Results shown in Figure 6.3 compare the activation of onset-embedded words and longer competitors for sequences where the initial syllable of both items is coded as duration 2 (and could therefore come from either a short embedded word or a longer competitor).

The results shown in Figure 6.3a and Figure 6.3c essentially replicate the pattern of performance shown in simulations without duration information. Without preceding context the network is unable to identify the underlying speech rate for these sequences. Consequently it processes the input segments representing the embedded words identically – regardless of whether the embedded syllable comes from a short or a long word. In these circumstances the network is forced to use following context to resolve this ambiguity. Since these networks receive less training on inputs requiring the use of following context than the simulations reported in Chapter 3, their performance is marginally worse than the networks trained without duration cues in Simulation 1. However, as can be seen by comparing these graphs with Figure 3.4b and Figure 3.5a, the activation profile of both sets of networks is qualitatively similar.

By comparison, Figure 6.3b and 6.3d show that the network where preceding contexts are available, the network can use the duration of the preceding word to determine the rate at which the current sequence is being produced. Consequently, although the input for the syllable /kæp/ is identical in both cases (duration code 2), these networks can determine whether the ambiguous duration is more likely to have come from a short or a long word and increase the activation of the appropriate lexical unit accordingly. Both *cap* and *captain* are presented with a single word of preceding context, with an unambiguous value of duration in each case. This difference in duration allows the networks to determine whether the target syllable is faster or slower than the unambiguous preceding word allowing them to process an otherwise ambiguous input correctly. These networks therefore show increased activation of the appropriate lexical units before the offset of the embedded word.

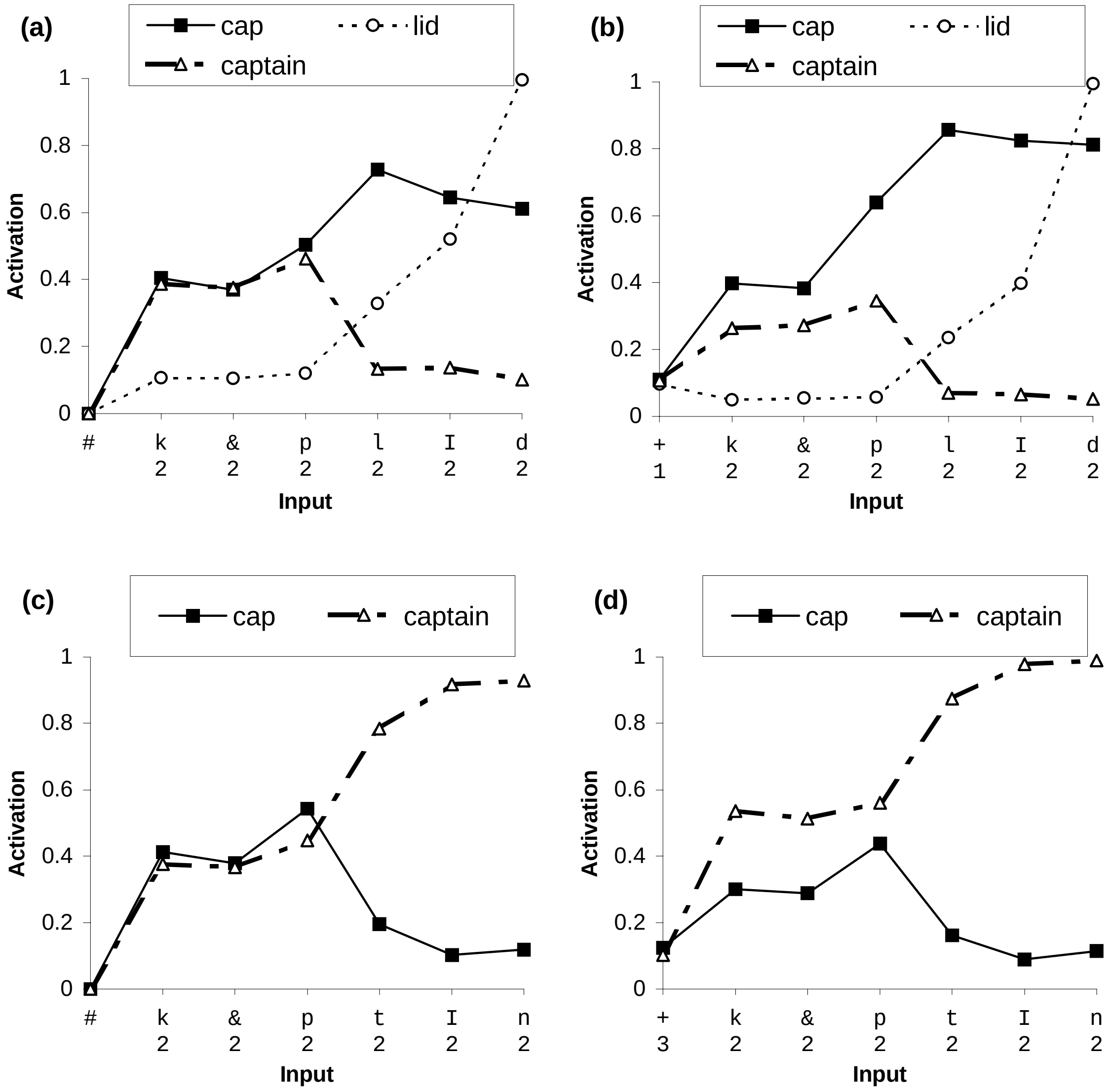


**Figure 6.3: Activation of an onset-embedded word (*cap*) and competitor (*captain*) averaged over ten networks in Simulation 3. Input segments are presented with the duration codes shown in Table 6.1. Embedded words occur either as the first word in a sequence (Figures a and c, preceded by a sequence boundary marked #) or as the second word in a sequence (Figures b and d, preceded by another word marked + with an associated syllable duration).**

**Example sequences are:**

**(a) *"cap lid"*** (fast rate) **(b)** bisyllable + ***"cap lid"*** (fast rate)
**(c) *"captain"*** (slow rate) **(d)** monosyllable + ***"captain"*** (slow rate)

This difference between the activation profile of onset-embedded words at the start or in the middle of a sequence, makes the prediction that if syllable durations overlap in short and long words, then the ambiguity created by onset-embedded words will be greater when embedded words are presented in isolation or at the onset of a sentence. Without information from prior context by which to determine speech rate it may not be possible for the recognition system to account for contextual variation in order to use syllable duration as a cue to the length of the target word. Since the majority of experiments looking at word identification have used single word stimuli this may explain the absence of duration effects in the literature reviewed in Chapter 4. Further experimental work should aim to investigate whether – as observed in this simulation – differences in the duration of preceding syllables can alter the perception of an otherwise ambiguous embedded word (as illustrated in Figure 6.3b and d).

However, one aspect of this model may be inappropriate as a simulation of the processing of duration cues to word length. Since duration information is presented at all positions in a syllable it is possible for the model to use changes in the duration of adjacent segments as a cue to the location of a word boundary. Given the importance of transitional information in the processing of simple recurrent networks (Elman, 1990; Servan-Schreiber, Cleeremans, & McClelland, 1991) it is likely that this information plays a role in the network's use of duration as a cue to word length. This transitional information constitutes an unrealistic aspect of the model since these transitions carry information that would not be present in connected speech where only certain speech segments may change with syllable duration.

To take a concrete example, both of the input sequences shown in Figure 6.3b and Figure 6.3d used an unambiguous duration prior to the embedded word. Consequently, there was a change in duration at the onset of the target word that was distinctive to sequences in which the target word was a short word (Figure 6.3b) or a long word (Figure 6.3d). It is possible that this change in duration (rather than duration per se) provides a cue to the length of the target word.

By comparison the test sequences shown in Figure 6.4 illustrate the processing of sequences where two successive words have an ambiguous duration value. Where both words are presented with duration code 2, there will be no change in duration at the transition between words to provide a cue to the length of the target word. Instead, the network must use the duration of the first word in combination with its lexical identity to determine the rate at which the sequence is being presented.

The most common case in the networks' training sets where two successive words are presented with an ambiguous duration is where two monosyllables are presented in a sequence at the fast rate. The networks' processing of this input is illustrated in Figure 6.4a. As can be seen the network still succeeds in using the duration of the word to increase the activation of the monosyllable over its longer competitor before the acoustic offset of the word.

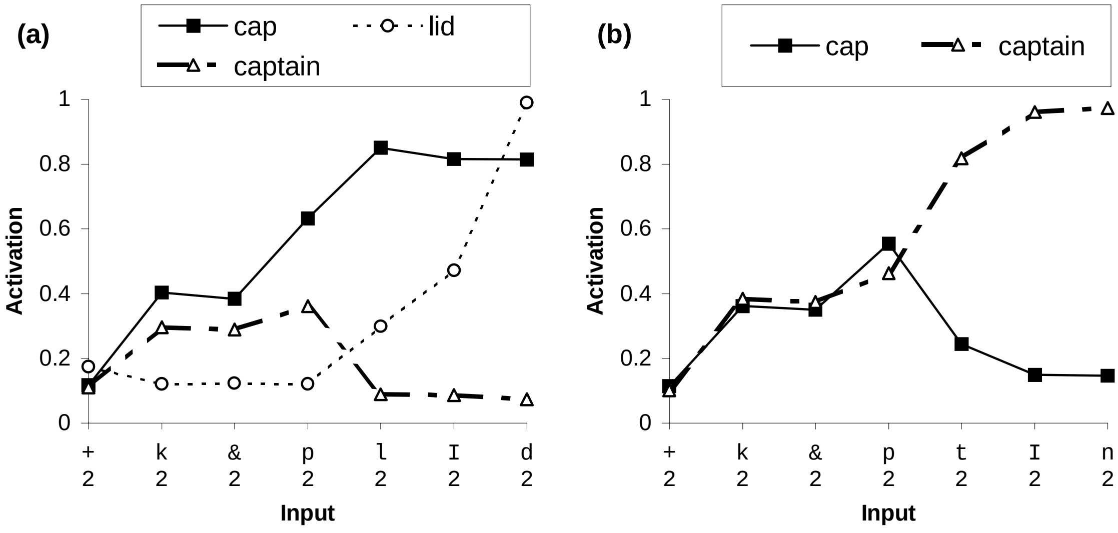


**Figure 6.4: Activation of an onset-embedded word (*cap*) and competitor (*captain*) in Simulation 3. All syllables in these test sequences were presented with an ambiguous duration.**

**Example sequences are:**
**(a)** monosyllable + ***"cap lid"*** (fast rate) **(b)** bisyllable + ***"captain"*** (slow rate)

A similar pattern of two successive ambiguous duration codes can also occur where two bisyllabic words are presented consecutively in a sequence at the slow rate. This case will occur less often in the training set (since there are fewer bisyllabic words than monosyllabic words in the language). The network's processing of this input is illustrated in Figure 6.4b. As can be seen in this graph the networks no longer favour the appropriate lexical item at the offset of the embedded word. Comparing Figure 6.4b with Figure 6.3d shows that the networks' activation profile in identifying bisyllables containing an embedded word is only altered by the duration of the previous word where there is a change in duration at the word boundary. If the lexical identity of the previous word as well as its duration must be used to establish the rate at which the sequence is presented (where the duration associated with both words in a sequence are ambiguous) the networks is not always able to use prior context in processing ambiguous input.

### 6.2.4. Discussion

These networks are clearly able to process input that includes a representation analogous to duration. Although not shown here, the network's responses to unambiguous duration codes are clear and categorical. However, as was described in the conclusions of Chapter 5, because of the amount of variation in segment and syllable duration (both between and within speakers) it is likely that duration will seldom provide an unambiguous or absolute cue to the number of syllables in a word. Consequently, in modelling sensitivity to duration cues, this simulations has focused on a test case in which the input to the network would be ambiguous without prior context (i.e. it could come from either a monosyllable or a bisyllable). In this case embedded words can only be disambiguated where prior context is used to detect the underlying 'rate' of the sequence. Since the overall rate at which a sentence is produced has been shown to effect voice onset time (VOT) boundaries for the perception of stop consonants (Wayland, Miller and Volaitis, 1994), it might be expected that equivalent results would be observed in the perception of monosyllabic and bisyllabic words. However, further experiments are required to demonstrate that the same adaptive properties are observed in the use of syllable duration as a cue to word length and word boundaries.

As discussed in conjunction with Figure 6.3 there is clear evidence that the network is able to use duration as a cue (relative to preceding context) to determine whether an embedded word is more likely to have come from a short word or a long word. However detailed investigation suggests that the networks' use of this information is more efficient for stimuli in which the previous word has an unambiguous duration. This suggests that it is the transition between different duration syllables that carries the most salient information for the network. It is therefore necessary to establish that the network can make use of duration information where a more appropriate representation of the speech stream is provided.

## 6.3. Simulation 4 – Representing duration from vowel to vowel

It has been shown in Simulation 3 that the networks investigated here are capable of using inputs corresponding to duration as a relative cue. However the input representation used for this network provides a more salient duration cue than might be found in the speech stream. Most importantly, the input representation assumes that duration information is present in all of the segments that make up a word. Although prior work has shown that changes in speech rate will manifest themselves across an entire syllable (for example VOT will change for a word-initial stop consonant depending on overall syllable length, Miller (1979)), it is unlikely that these differences in rate could be detected from the onset of a syllable. Indeed there is a discrepancy between the position at which changes arise as a result of altered syllable durations and the point at which syllable duration can be detected in order to compensate for these changes. This has been a focus of recent modelling work on the effects of speech rate on phonetic categorisation (Abu-Bakar & Chater, 1995).

Effects of syllable duration on phonetic categorisation suggest that the overall duration of a syllable can not be established until late on during its presentation. Consequently transitional information that was responsible for the processing of duration cues in Simulation 3 is unlikely to be available in real speech. A more realistic assumption would be that duration information is only available late on during the processing of a syllable. Since the majority of variation in syllable duration is caused by changes in the vowel (Klatt, 1976) a further set of simulations were therefore carried out in which duration input is only presented for vowel segments.

### 6.3.1. Network architecture and training set

The network architecture, input and output representations and vocabulary remained the same for this simulation as in Simulation 3. The only change made was that the duration input was only activated for the vowel of each syllable. By providing duration information in vowels only, the network is no longer able to detect changes in syllable duration at word boundaries. To make use of prior context in processing ambiguous values of duration, the network will have to retain information about the preceding syllable across at least two intervening, unmarked segments. In order not to penalise the network for continuing to activate the duration output after the vowel segment, no error was propagated back from the duration units at the prediction output unless the predicted segment was a vowel.

Results of initial simulations carried out using this training regime were disappointing – the network showed no ability to use prior context in processing syllables with ambiguous durations. Graphs equivalent to those in Figure 6.3 showed that the pattern of activation in these networks was identical, irrespective of whether ambiguous (embedded) syllables occurred at the start or as the second word of a sequence. This finding confirms that in Simulation 3 it is transitional information between syllables that enable the networks reported previously to display sensitivity to prior context in processing the duration information. Where duration information is separated by unmarked segments, networks were unable to use prior context in processing the ambiguous input.

These results are reminiscent of those reported by Elman (1993) showing that it is difficult to train SRNs to process long distance dependencies where unrelated information is presented at intervening time steps. This is a consequence of the SRN having limited access to representations of states at previous time steps. Unless relevant aspects of the prior input are represented in the hidden units at the time step before it needs to be used then the network will be blind to prior context. Since two unmarked time steps separated the duration inputs for successive syllables, these networks did not retain a representation of the duration of the previous syllable that would allow them to use prior context.

One solution that has been used in training networks faced with this problem is to use fully recurrent networks in which error is propagated back over copies of the system’s internal representations extending over more than one time-step (Rumelhart, Hinton, & Williams, 1986). Such an approach is not without its problems, however, since information represented several time steps previously will have a decreasing influence on the error gradients that drive the learning algorithm. Another approach used to encourage the network to retain an internal representation of the necessary information was described by Maskara and Noetzel (1993). This requires the network to output a copy of the hidden unit activations from the previous time step. Such an approach is reported to be successful in learning centre-embedded sequences that SRNs find difficult.

The approach taken in this thesis was to extend the prediction task that was used previously. As in simulations reported by Shillcock, Levy and Chater (1991) and Gaskell, Hare and Marslen-Wilson (1995) output units were added that not only represented the following segment and duration to be presented in the input, but also the current and previous input. In order to activate a representation of the identity and duration of segments presented at preceding time steps the network must retain an appropriate representation at the hidden units. This internal representation helps ensure that the network has access to segment information and syllable duration from the previous word – assisting the learning of the duration cue.

Aside from the additional output units for this extended output task, all other aspects of the network were identical to those reported previously. A simple recurrent network was used with 9 inputs, 50 hidden units copied back to 50 context units, and 47 output units – 20 lexical units and 9 output units (6 for phonetic features, 3 for duration) for each of the 3 prediction outputs. The network was trained on the same vocabulary used before with duration represented over three units using the coding scheme shown in Table 6.1. However, in this simulation, duration information was only presented at the input for the vowel of each syllable. Similarly, error was only propagated back from output units representing the duration of vowel segments.

### 6.3.2. Results and discussion

Ten networks were trained for 500 000 sequences using this architecture and training regime. All results reported below are the average of ten training runs, each having different initial weights and randomly generated training sequences. In contrast to preliminary simulations that did not include output units representing the current and previous input, these networks were successful in utilising prior context to identify the onset-embedded words in the training set. In all four combinations of preceding word and ambiguous target syllable shown in Figure 6.5, the model produces increased activation of the correct lexical item at the offset of the syllable forming an embedded word.

Therefore, incorporating additional output units representing the current and previous input segments enabled these networks to retain duration information from the vowel of one syllable to the vowel of the next. By retaining this additional information over several intervening time steps the network was able to use duration cues in cases where the previous word, as well as the current word, is presented with an ambiguous syllable duration – unlike the networks described in Simulation 3.

The addition of output units that encourage the network to retain a more complete representation of the prior context appears to increase the networks ability to retain lexical information about the preceding word. This allows the system to use not only the duration of the preceding word but also its lexical identity to detect the rate at which a sequence is presented. Hence these networks can process ambiguous target syllables more effectively than the networks in Simulation 3.

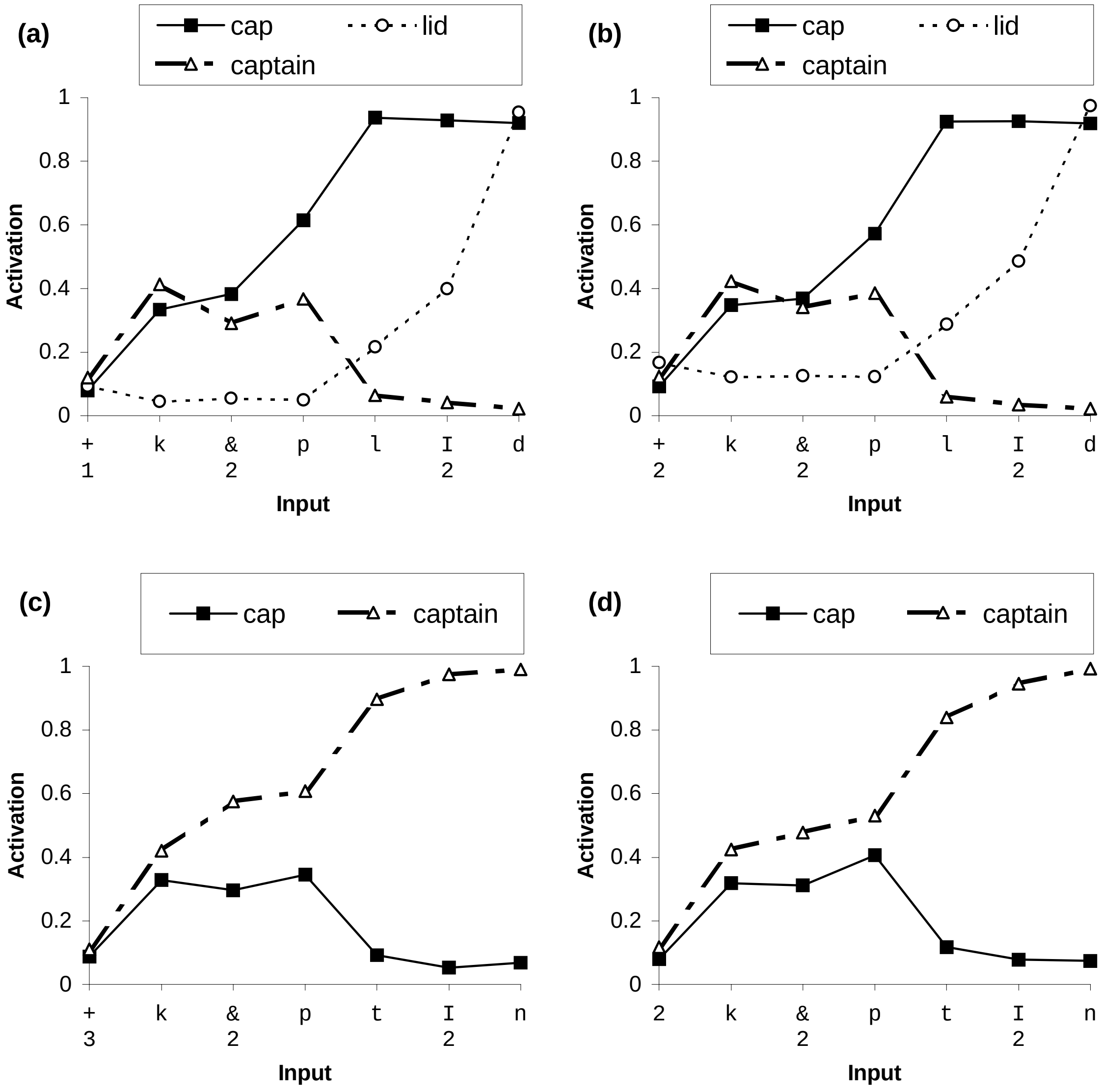


**Figure 6.5: Activation of onset-embedded words (*cap*) and competitors (*captain*) with ambiguous durations in Simulation 4. Target syllables occur as the second word of a sequence, preceded either by a word with an unambiguous duration (a and c) or an ambiguous duration (b and d)**

**Example sequences are:**

**(a)** bisyllable + "***cap lid*”** (fast rate)
**(b)** monosyllable + "***cap lid*”** (fast rate)
**(c)** monosyllable + ***"captain*”** (slow rate)
**(d)** bisyllable + ***"captain*”** (slow rate)

It is of interest that this property was only observed where additional output tasks were used to force the network to represent previous input at the hidden layer. In all the simulations reported so far the network must still activate a representation of the identity of the preceding word at the lexical units. However, it appears that in order for the networks to use the identity of previous words this information needs to be represented in a particular form at the hidden units – merely activating the appropriate lexical unit is insufficient. These additional output tasks therefore play a valuable role in structuring the networks' internal representations to enable them to use duration information in identifying lexical items. A similar argument was also applied to the faster learning observed in Simulation 2 reported in Chapter 3; comparing networks with and without the prediction task suggests that the additional task helps structure the networks' internal representations of the input to assist in learning the lexical identification task.

The networks in Simulation 4 are therefore capable of using duration information to disambiguate onset-embedded words in sequences in which the detection of durations associated with short and long words must be adjusted by consideration of the rate at which preceding words in a sequence were presented. Adaptive processing such as this is likely to be required in order to use the duration of syllables in connected speech as a cue to the location of word boundaries. It therefore seems appropriate to compare the activation profile observed in the networks reported in Simulation 4 to the results of the experiments reported in the previous chapters.

## 6.4. Simulating experimental data

Given concerns over the role of response biases in gating, cross-modal priming is likely to provide a more transparent measure of lexical activation than gating data. A further assumption in simulating priming data is that the magnitude of priming is directly proportional to the activation of the relevant lexical output unit in the network. However, since priming effects are not equivalent to lexical activations in a network it is inappropriate to transform the activation scale into priming units. Comparisons between experiments and simulations will be facilitated, however, by plotting priming and simulation data side-by-side and by using the same scale in all four sets of experimental data and simulation results.

### 6.4.1. Method

An important goal in simulating the priming data will be to ensure that statistical analyses of the networks' activations produce the same results as those obtained for behavioural data. Since there are only two onset embedded words and two longer competitors in the network's vocabulary it will not be possible to do analyses over different items. Consequently, statistical analysis will focus on analysing whether a pattern of results is reliable across all ten networks trained, treating each network as a single subject. Experimental evidence suggested that there were two sources of information that are relevant to the identification of onset-embedded words and longer competitors. We will describe how these sources of information were simulated in turn.

***Duration cues to word boundaries***

The first source of information involved in the identification of onset-embedded words is the acoustic difference between syllables of monosyllabic and bi-syllabic words. Simulations reported in this chapter represent this acoustic cue as a difference in the activation of a group of units that code for syllable duration. As discussed previously, this may be an unrealistic representation of the speech stream. However, this simulation captures the adaptive nature of the acoustic processing of duration cues to word boundaries: specifically that sensitivity to the duration cue requires a comparison of the current syllable with that of preceding words in the sequence. In order to incorporate this property into the simulation, the critical syllables of the test stimuli were presented with an ambiguous duration (i.e. both monosyllabic and bisyllabic words used duration code 2 from Table 6.1). Since the stimuli used in the priming experiments have several syllables of preceding context, the test sequences for the network were presented with one word of preceding context to allow the duration input to be disambiguated.

The test sequences contained a mixture of preceding contexts with both monosyllabic and bisyllabic words. Preceding contexts for the onset-embedded words will exclude the longer words in which they are embedded and vice-versa. In all cases, however, the target syllable will be of an ambiguous duration (code 2). These target words are more appropriate in simulating the results of the priming experiments since prior duration is required to process the duration of target syllables in these test stimuli correctly.

### *Continuations of short word stimuli*

The second source of information that plays a role in the identification of onset-embedded words and longer competitors is the segment that follows the offset of the embedded word. In our test stimuli this continuation matched a longer word. Priming results for these lexical garden-path stimuli suggested that longer lexical items continued to be activated. It is only at later probe positions where there is mismatch between these continuations of short word stimuli and longer lexical items that significant priming of embedded words in the absence of priming of longer competitors was obtained. Thus it was argued  that lexical garden-path stimuli – such as the sequence *cap tucked* – played an important role in producing the activation profile shown for short word stimuli in cross-modal priming.

To simulate the results of these priming experiments, embedded words were therefore placed in lexical garden-path contexts equivalent to those used in the experimental stimuli. These lexical garden paths were generated such that a monosyllabic word matching the onset of the longer lexical item followed the embedded word (i.e. sequences used were of the form *cap tap* rather than *cap topknot*). This will ensure that changes in the duration of subsequent input do not provide an additional cue that following segments come from a separate word.

### *Probe positions and alignment points*

In comparing the time course of activation of short and long words in these sequences, it is important to ensure not only that the stimuli are appropriate for comparison with experimental data, but also that the probe positions match those that were used in the experiments. Since the input to the network is divided into discrete segments, information in the input can be specified more precisely than was possible for the speech used in the experiments. However, in the initial comparisons of experiment and model, it is simplest to assume that the information assumed to be available at each alignment point in the experimental stimuli is available to the network.

It is assumed that the stimuli up to $AP_1$ contain information about the identity of the first syllable but no information about following segments. The section of speech between $AP_1$ and $AP_2$ was assumed to contain information relating to the identity of the onset of the following word, but no information that mismatches with the longer lexical item. After

$AP_2$, the stimuli containing short and long words diverge phonemically as information in the vowel (up to $AP_3$) and offset (up to $AP_4$) of the second syllable is presented to the network.

Thus, for one of the two pairs of test sequences in the network (c*aptain* and *cap tap*, rather than *bandit* and *ban dock*), $AP_1$ was following the segments /kæp/ with the vowel being presented with duration 2. Thus, this syllable will be identical for stimuli containing short and long words. However, it is expected that the network will be able to use prior context in disambiguating this input. As was intended for the experimental stimuli, $AP_2$ includes the initial segment of the following word e.g. /kæpt/. It is only at $AP_3$, following the input /kæptɪ/ for *captain* and /kæptæ/ for *cap tap* that test stimuli will diverge phonemically. The final probe position $AP_4$ is assumed to be at the offset of both words /kæptɪn/ and /kæptæp/ respectively – a point some time after there is mismatch between the two sets of stimuli. In the following sections the experimental and simulation results for each probe position will be described. In analysing the results of the network simulations, stimulus type corresponds to whether the test sequences contains a short or a long word – equivalent to the prime stimulus factor in the experimental data. Lexical unit refers to whether the activation of a short or long word unit is being measured – analogous to the long or short target word in the priming experiments.

### 6.4.2. Experiment 2a – $AP_1$

The main result obtained in this cross-modal priming experiment was that at the offset of a syllable there is sufficient information for listeners to distinguish an embedded word from the onset of a longer competitor. As shown in Figure 6.6a, there was a cross-over interaction between the length of the prime and target such that greatest priming is observed in conditions for which the prime stimuli contains a word of the same length as the target. This pattern is shown by a significant interaction between prime and target length in the difference score analysis with no main effect of either prime or target length.

As can be seen in Figure 6.6b a similar pattern of results is shown for the activation of short and long lexical units in the network. Analysis of variance on the data obtained in the ten trained networks confirms the presence of a significant interaction between stimulus type and lexical unit (F[1,9]= 45.39, p<.001) with no main effect of stimulus type

(F<1) and a marginal effect of lexical unit (F[1,9]= 4.54, p<.1) suggesting increased activation for short lexical units.

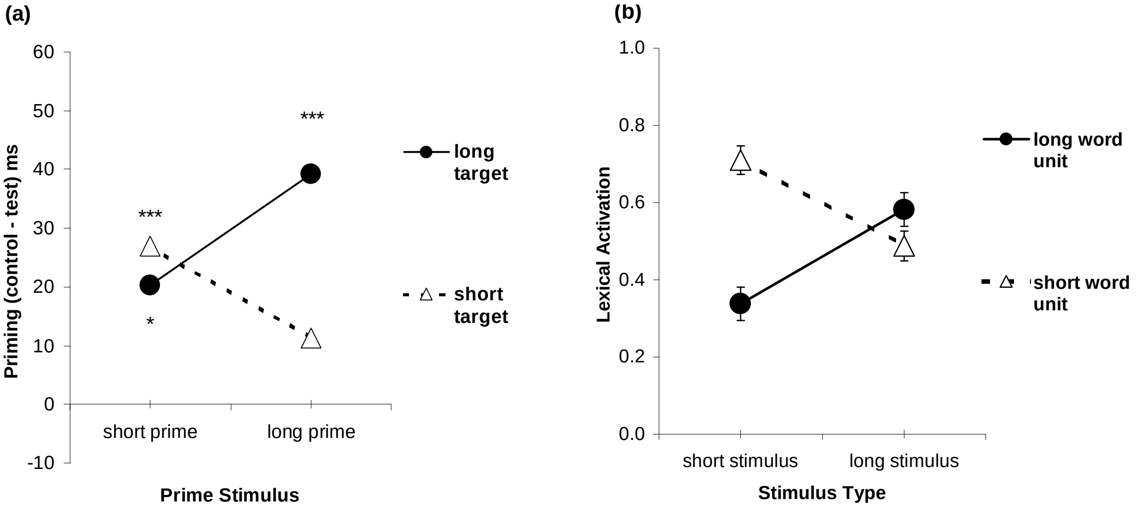


**Figure 6.6: Activation of short and long words at $AP_1$ – following /kæp/ in *cap* or *captain*. Network stimuli presented with ambiguous duration and prior context to allow disambiguation.**

**(a) Magnitude and significance of priming in Experiment 2a (*** p<.001; * p<.05)**

**(b) Mean lexical activation for 10 networks in simulation 4 (error bar = standard error)**

Thus the network simulates the main piece of experimental data suggesting that listeners are able to use acoustic cues to word length in the recognition of onset-embedded words and longer competitors. Increased activation is observed for lexical units that match the stimulus. Since the duration cue associated with these syllables is ambiguous in the absence of prior context, in order to display this result the network must be using prior context to disambiguate the input sequences. Thus in the absence of prior context the network predicts that the two sets of stimuli would be indistinguishable at this point. This prediction could be tested in subsequent experiments.

One difference between the experimental results and the simulation is that the network produces a marginally significant increase in activation for short words over long words that is in the reverse direction to the numerical trend observed in the experimental data. This increased activation may reflect the greater number of monosyllabic words in the networks' vocabulary, since the ambiguous duration code is more frequently paired with a short word than with a long word. However, since neither the effect in the experimental data, nor the reverse effect in the model reaches statistical significance, it can still be concluded that there is good overall agreement between the simulation and experimental data.

### 6.4.3. Experiment 2b – *AP*$_2$

At the second alignment point, information in the continuation of the second syllable becomes available in the speech stream. In the cross-modal priming experiments, this increased the amount of priming observed for long targets – especially for prime stimuli that contained the long word. This is reflected in the analysis of priming effects at this probe position, which found main effects of prime and target type (more priming of long targets and more priming from long primes) in addition to the significant interaction between prime and target type. This pattern of results is shown in Figure 6.7a.

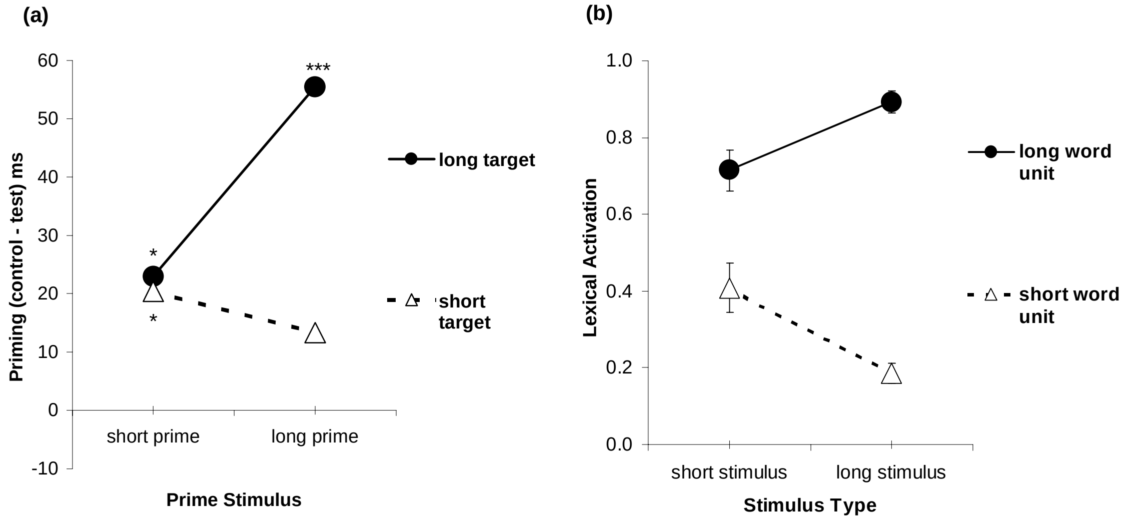


**Figure 6.7: Activation of short and long words at *AP*$_2$ – following /kæpt/ in *cap tucked* or *captain*.**
**(a) Magnitude and significance of priming in Experiment 2b (*** p<.001; * p<.05)**
**(b) Mean lexical activation for 10 networks in simulation 4 (error bar = standard error)**

As can be seen by comparing this graph with Figure 6.7b, the model shows the same interaction between stimulus type and lexical unit as was obtained in the priming data (F[1,9]= 25.55, p<.001). However these networks are less successful in simulating the main effects of prime and target type shown in the experimental data, since they show significantly greater activation for long word units, irrespective of prime condition. This effect is confirmed by the significant main effect of lexical unit in the analysis across all ten networks (F[1,9]= 51.38, p<.001). Conversely the model does not show greater overall activation for long stimuli than short stimuli (F[1,9]= 1.73, p>.1) indicating that while the long word stimuli in the experiment produce greater overall priming (possibly as a result of their reduced ambiguity) the same pattern is not observed in the model.

These discrepancies between simulations and priming data indicate that the model predicts increased activation for long word units for both short word (*cap tap*) and long word (*captain*) stimuli. By comparison, increased priming of long words is only shown where the prime stimulus actually contains a long word. Thus, this comparison of simulations and experiments, suggests that there may be experimental evidence for acoustic cues in word-initial segments that mark word boundaries. Previous experiments (Gow & Gordon, 1995; Nakatani & Dukes, 1977) have suggested that segmental cues in word onsets support the detection of word boundaries. Incorporating these acoustic cues into the model may therefore reduce the effect of garden-path continuations on the activation of long words and improve the networks' simulation of the experimental data. However, without further evidence to support the presence of acoustic cues to word onsets in the stimuli used in Experiment 2, it is premature to alter the input representation of the model.

### 6.4.4. Experiment 2c – $AP_3$

The sentences used in Experiments 1 and 2 were designed such that short and long words only diverged in the vowel of the second syllable of the critical stimuli. Consequently, it was expected that the activation of short word hypotheses would increase at the third alignment point, where information in the vowel becomes available to participants. However, the results of Experiment 2c showed that the short word stimuli remained ambiguous at $AP_3$ as shown in Figure 6.8a. Statistical analysis showed a marginally significant main effect of target type (by participants but not by items) suggesting greater priming of long word targets. There was no main effect of prime type and the interaction between prime and target length was again significant at this probe position.

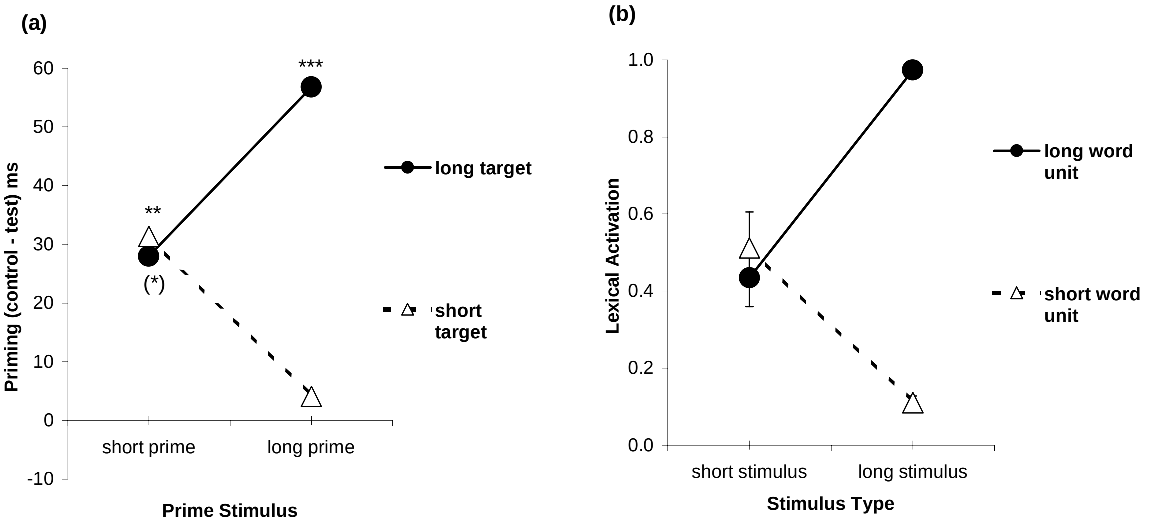


**Figure 6.8: Activation of short and long words at *AP*$_3$ – following /kæptuː/ in *cap tucked* or /kæptɪ/ in *captain***

**(a) Priming results from Experiment 2c (*** p<.001; ** p<.1; (*) p<.1)**

**(b) Activation for 10 networks in simulation 4 (error bar = standard error)**

As can be seen in Figure 6.8 there is a very strong resemblance between the pattern of priming observed in Experiment 2c and the activation of lexical units in simulation 4. This resemblance is supported by statistical analysis of lexical activations in the network which showed a main effect of lexical unit (F[1,9]= 26.83, p<.001) equivalent to the main effect of target type in the experimental data. Effects of stimulus type were non-significant in the network (as for the priming experiment) and the interaction between stimulus type and lexical unit was highly significant (F[1,9]= 54.22, p<.001).

Thus, there is good agreement between the network and the priming data regarding the time course of responses to mismatch in the speech stream. Both the recurrent network and the experimental data suggest that despite the presence of mismatch between short stimuli and long lexical items at *AP*$_3$, effects of mismatch are slow to act in reducing the activation of long words.

## 6.4.5. Experiment 2d – *AP*$_4$

The final probe position tested in the cross-modal priming experiments marked a point at which both short and long word stimuli were expected to be ambiguous. This lack of ambiguity is apparent in the priming effects obtained in Experiment 2d, shown in Figure 6.9a. The only significant effect in the statistical analysis of these data is a cross-over interaction between prime and target length. This reflects the expected pattern – that significant priming is observed only for conditions in which the prime and target match.

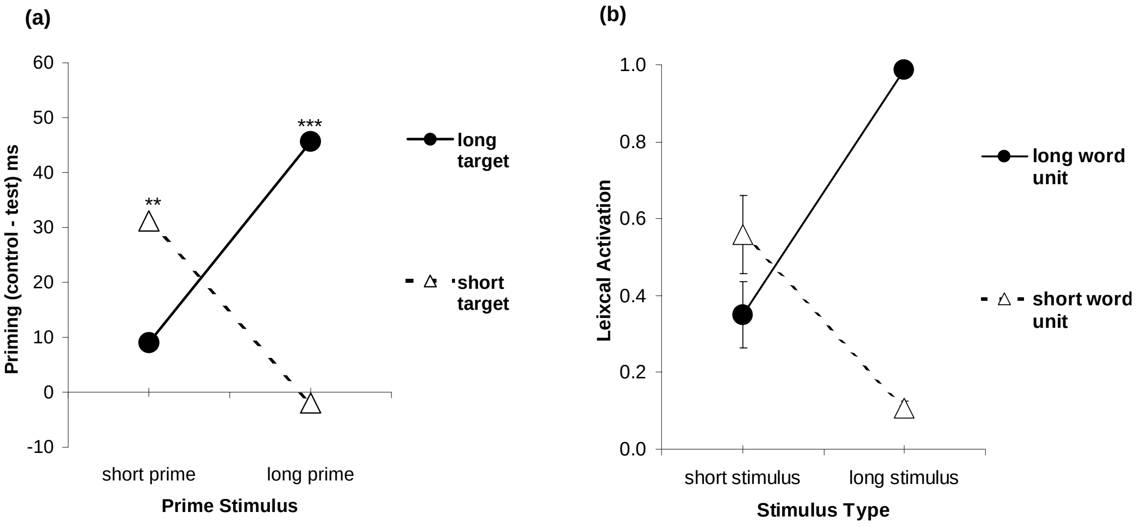


**Figure 6.9: Activation of short and long words at *AP*$_4$ – following /kæptuːk/ in *cap tucked* or /kæptɪn/ in *captain***

**(a) Priming results from Experiment 2d (*** p<.001; ** p<.1)**

**(b) Activation for 10 networks in simulation 4 (error bar = standard error)**

Once more, visual comparison of the priming data and network activations shown in Figure 6.9 are encouraging. Both graphs suggest that short and long stimuli can be identified at this probe position (though long stimuli appear to be less ambiguous than short word stimuli). However, statistical analysis of network activations do not reflect this apparent similarity. ANOVA shows a significant main effect of lexical unit indicating greater overall activation of long word units (F[1,9]= 17.29, p<.01). There is also a marginally significant effect of stimulus type indicating greater activation for long word stimuli (F[1,9]= 3.40, p<.1). More reassuringly, the most significant effect in the analysis of the networks' performance is the interaction between stimulus type and lexical unit (F[1,9]= 55.11, p<.001). This indicates that despite these discrepant main effects, the simulation does capture the lack of ambiguity of the stimuli at this probe position.

Both of the main effects reported in the model appear to result from the very small amount of variance that is observed in output activations for long word stimuli. As suggested by the invisibility of the error bars on the right hand side of Figure 6.9b all ten networks fully activated the long word units and deactivated the short word units at this position in the long word stimuli[1]. In contrast to this unambiguous activation profile for long word stimuli, short word stimuli are rather more ambiguous at this probe position and hence activations are more variable for these stimuli. However, since priming data are highly variable for both short and long target words, main effects of prime and target type are not observed in the analysis of the experimental data.

### 6.4.6. Discussion

The final set of simulations reported here demonstrate that simple recurrent networks are able to account for the integration of segmental and supra-segmental cues to the identification of onset-embedded words in connected speech (post-offset mismatch and syllable duration respectively). Despite the limited vocabulary on which the model was trained, statistical analyses of network activations showed many of the same effects that were reported in the cross-modal priming data presented in Chapter 5. Although the exact details of the results at each probe position are sometimes lacking, the network clearly simulates the initial lack of ambiguity of short and long word stimuli. At subsequent probe positions the network also simulates the bias towards long word interpretations followed by the reduction in ambiguity of the short word stimuli when mismatching input is processed.

The network results illustrate the promise of an account, in which lexical activations are proportional to the conditional probability of individual lexical items in the input. However, caveats remain about the representational assumptions that have been incorporated into these networks. In modelling the adaptive processing of input cues analogous to duration, these simulations have focussed on a highly restricted subset of the variables involved in determining the duration of segments and syllables in connected speech. Although the greater complexity of real speech appears to make the task faced by the network more difficult, the addition of more realistic speech input may reduce the absolute degree of ambiguity that is present in these stimuli. If multiple cues to syllable duration were present, it is unlikely that the length of syllables would be as exactly matched as was assumed in the simulations reported here. This discussion illustrates one

[1] The residual activation of short word units represents the likelihood that they will appear as one of the remaining words in the current sequence

limitation of the recurrent network simulations developed here – the highly unrealistic input representation used in the model. Further work is therefore required to investigate whether these networks can incorporate more realistically structured input and output representations.

### *Representing the speech input*

The networks used in these simulations are provided with inputs that represent the duration of segments and syllables independently of the identity of these input segments. This assumption does not hold for real speech since duration, particularly for vowel segments, is strongly affected by the identity of adjacent segments (Klatt, 1976). Furthermore there is evidence from gating studies (Warren & Marslen-Wilson, 1988) that this duration information can contribute to lexical choice. Finally, in many languages, lexical items are contrasted solely by vowel duration.

Since each of these properties contradicts the assumptions used in developing the networks reported here, further work is required to show how differences in duration can be more appropriately represented in the network. However, simulations representing differences in duration as different numbers of duplicated input segments have not been successful. The SRN that was used here fails to process duration coded as duplicated segments in an adapative manner.

An alternative means of processing duration information is therefore required for a complete account of sensitivity to temporal structure in the speech stream. One account of how a network could adapt to differences in the temporal properties of the speech stream is provided by the dynamic rate adaptation networks investigated by Nguyen and Cottrell, (1997). They investigated recurrent networks in which the time delay on their recurrent connections is adjusted to minimise prediction error. These systems can thereby match their processing properties to the rate of the current input. However this process is implemented as an off-line process and thus may not be appropriate in simulating the on-line processing of spoken input.

An alternative account is provided by the entrainment processes implemented in oscillator based networks (Gasser, Eck, & Port, 1999; McAuley, 1994). These systems use discrepancies between the predicted and actual position of metrical beats in the onsets of stressed syllables to alter the rate of oscillation of processing units. Thus systems of these neurons will gradually adapt to the rate of an ongoing sequence.

Either of these approaches provides an account of how a connectionist system can alter its computational properties to compensate for changes in the rate of presentation of the speech stream. Either of these approaches may therefore be required for a more complete model of temporal processing of speech stimuli in connectionist networks.

## 6.5. General discussion

The networks shown in Simulation 4 are very successful in accounting for the pattern of priming data produced in Experiment 2. However, as is apparent from the small scale of the model, further simulations are required to extend this account to cover more realistically sized vocabularies. This should allow the network to simulate experimental data other than that reported in this thesis. For instance, it may be possible to extend the model to account for a wider range of ambiguous sequences, such as the minimal pairs (e.g. *grey day* and *grade A*) used in the acoustic phonetics literature (see Chapter 2 for further details).

In addition to increasing the model’s coverage of segmental ambiguity, this architecture could also be trained to simulate a wider range of supra-segmental phenomena – such as the difference between metrically stressed and unstressed syllables. In this way it may be possible to simulate results used to motivate the metrical segmentation strategy (Cutler and Norris, 1988). These extensions to the model will further test the probabilistic approach to lexical access and segmentation that has been developed in this thesis.

The recurrent network simulations presented here have shown that a probabilistic approach to the process of spoken word recognition has the potential to simulate many of the results that have previously motivated lexical competition models such as TRACE and Shortlist. These simulations have therefore demonstrated how computational properties of the recognition system such as bottom up activation and inhibition of lexical candidates, competition between lexical items spanning potential word boundaries and adaptive processing of the input can all be learnt by a simple gradient descent algorithm. Further investigations are required to establish if this account is able to simulate a wider range of experimental data.

Even in its current, limited form, the network makes a number of testable predictions for future experiments. The first of these is that the acoustic cues that discriminate syllables of short and long words should be more easily detected with a preceding sentential context than in the absence of such a context. A second prediction is that altering the temporal properties of the critical syllables (or their preceding context) will have an strongly biasing effect on subjects' interpretations of those syllables.

A further prediction made by these simulations is that the segments that follow the offset of an embedded word will affect the ease with which subjects can identify an embedded word. Garden path sequences (such as *"cap tucked"*) in which continuations match a longer lexical item (*captain*) are particularly difficult for the network to identify. In contrast sequences which mismatch with all other lexical items immediately after the offset of the embedded word will be comparatively easy to process. It is this prediction of the network that will be tested in the following chapter.

# 7. Effects of following context in recognising embedded words

In previous chapters it has been shown that acoustic cues to word length are used in the identification of words embedded at the onset of longer words. In none of the repetition priming experiments reported in Chapter 5 was significant priming of onset-embedded words (such as *cap*) observed from stimuli containing longer words that contined these embeddings (*captain*) - even where the prime word was cut off at the offset of a syllable matching an embedded word. However, these experiments also showed that during the identification of sequences containing an embedded word with a garden-path continuation (such as the test sequence *cap tucked*), longer competitors (e.g. *captain*) remain active until after the offset of the embedded word. Significant priming of longer words was observed at probe positions up to the vowel of the following syllable.

In the recurrent network simulations reported in Chapter 6, the lack of ambiguity between short and long words was modelled by incorporating an input cue analogous to the duration difference between syllables in short and long words. The network was able to use this cue even where syllables of short and long words are presented with the same, ambiguous duration. This indicates that the network is able to process duration adaptively, in order to disambiguate onset-embedded words from longer competitors. The model is therefore able to account for priming results suggesting that short and long words are not as ambiguous as previously predicted. The network also simulated the identification of embedded word sequences that contain a lexical garden-path. Where continuations for short word stimuli matched the second syllable of a longer word, activations for the longer lexical item were boosted during identification. Consequently, the garden-path sequences that were used in Experiments 1 and 2 were responsible for the increased activation of longer words like *captain* in the recurrent network simulations.

For this reason, it is expected that short word stimuli would only cause ambiguity for sequences that create a lexical garden-path. Input sequences such as *cap lick* (shown in Figure 6.5a and b) that mismatch with all longer words do not activate longer competitors after the offset of the embedded word. It is therefore a crucial test of the validity of the model to investigate whether listeners also show a different activation profile for sequences containing onset-embedded words in non-garden-path following contexts. The experiments reported in this chapter therefore investigated the identification of short embedded words in contexts that rule out longer interpretations immediately after the offset of the embedded word. The results of these experiments can then be compared with the activation profile predicted by the network.

## 7.1. Experimental materials

The goal of the experiments reported in this chapter was to investigate the time course of recognition of onset-embedded words under conditions where less disruption from longer competitors was predicted. Sequences containing an embedded word were used with a continuation that diverges from all other lexical items immediately after the offset of the embedded word (rather than the delayed mismatch that was used in Experiments 1 and 2). As before, both gating and repetition priming will be used to probe the activation of short and long target words at different points in the speech stream. Since the critical comparisons here involve short word stimuli, it is not necessary to re-run the long word materials a second time. Furthermore since short words with garden-path contexts have already been tested in Experiments 1 and 2 it should not be necessary to re-test these stimuli. The experiments reported here therefore used only a single set of test sentences containing short, embedded words, with continuations that immediately mismatch with all longer competitors.

The items used in this experiment were derived from the same set of onset-embedded monosyllables used in Experiments 1 and 2, placed in the same context sentences as before. Words following the monosyllabic test word were changed so that the onset segments of the following word mismatched with all likely longer competitors. For an example pair of test words *cap* and *captain*, other lexical items that start with the syllable [kæp] include: *caption*, *capsule*, *captive* and *capture*. The onset of the following word was therefore chosen to mismatch with all the longer lexical items in the CELEX database (Baayen, Pipenbrook, & Guilikers, 1995). Since a continuation that begins with the phoneme /l/ mismatches with all these continuations, the sequence *cap looking* was used in the prime sentence for this stimulus item.

Results obtained by McQueen (1998) showed that listeners detected embedded words more easily where they are followed by a phonotactically illegal sequence. Continuations were therefore chosen that such that the segments either side of the word boundary are found word-internally in other English words. These phonotactically legal sequences would not therefore provide a pre-lexical cue to a word boundary. For the example word *cap*, other words in the CELEX database with the segments /æp/ at the end of a syllable include: *clapboard, haphazard, napkin, chaplain, entrapment*. By analogy with the word *chaplain,* continuations starting with the segment /l/, would therefore not provide a phonotactic cue to a word boundary. Hence the test sentence used for the monosyllable *cap* was "*The soldier saluted the flag with his <u>cap</u> <u>looking</u> slightly crumpled*" . A similar process was repeated for each of the onset-embedded words used in Experiments 1 and 2. The complete set of 40 sentences is shown in Appendix A.

These materials were recorded and digitised using the same methods employed previously. Three alignment points, equivalent to those used in Experiments 1 and 2 were marked for these stimuli ($AP_1$ at the offset of the first syllable of the test word, $AP_2$ after the onset segments of the following word and $AP_3$ in the vowel of the second syllable). The durations of each of these sections were compared to those of the short-word stimuli used in Experiments 1 and 2 (as shown in Table 4.1 in Chapter 4). The target monosyllable (e.g. *cap*) averaged 303ms in duration, which does not significantly differ from the mean duration (291ms) of the equivalent syllable in Experiments 1 and 2 (t(39)= 1.32, p>0.1). The onset segments of the second syllable (between $AP_1$ and $AP_2$) were 73ms in duration (compared to 78ms in Experiment 1 and 2, t(39)= 0.42, p>0.1), and the third alignment point was placed an average of 43ms into the vowel of the second syllable (42ms previously, t(39)= 0.66, p>0.1).

To allow comparison between the results of the current experiment and those reported previously, both gating and cross-modal repetition priming experiments were carried out using these stimuli. As in the previous experiments the goal of these investigations was to assess the ongoing competition between short and long word interpretations of these stimuli. With this in mind, methods were used to measure the activation of both the target word and a longer competitor. Since these experiments can be compared to those reported in Chapters 4 and 5, only the new set of short word stimuli are tested here.

## 7.2. Experiment 3 – Gating

The gating task was used initially to investigate whether the identification of embedded monosyllables follows a different pattern from that reported in Experiment 1 for sentences in which mismatch with longer competitors occurs in the onset of the following word.

***Participants***

Eleven subjects from the Birkbeck Speech and Language subject pool took part in the experiment. All were paid £5/hour for their participation. None had taken part in any of the previous experiments.

***Design, materials and procedure***

These were as described in Experiment 1 in Chapter 4, except that instead of stimuli containing either short or long target words, all the test stimuli contained embedded words with following contexts that immediately mismatched with longer lexical items. Gates were set up as previously at the three alignment points described in Figure 4.1 and Table 4.1 with two additional gates 50 and 100ms before $AP_1$, and other gates 50, 100, 200, 300 and 400ms after $AP_3$. Since only a single set of test sentences were investigated, these were presented in a single test version. As in Experiment 1, test stimuli were played out in successive fragments, preceded by 4 practice items and interspersed with 16 filler items. As previously the experiment was divided into 5 blocks – one of practice items and four blocks of test and filler items. The experiment took approximately two hours to complete, including short breaks following each block of test items.

### 7.2.1. Results and discussion

Results were analysed in terms of the proportion of responses at each gate that matched either the target word or a longer competitor. Three items produced a disproportionate number of errors. For *bran* and *ban* over 50% of participants failed to identify these words correctly by the final gate – responses for these items were therefore discarded. The third discarded item (*win*) produced a lower error rate (36%) in the current experiment but was discarded to aid comparisons with Experiment 1 (in which over 60% of participants failed to identify this target word). The proportion of responses matching either short or long target words at different gates is shown in Figure 7.1. Also included in this graph is data from the short word stimuli in Experiment 1 where stimuli included lexical garden-paths after the offset of the short word.

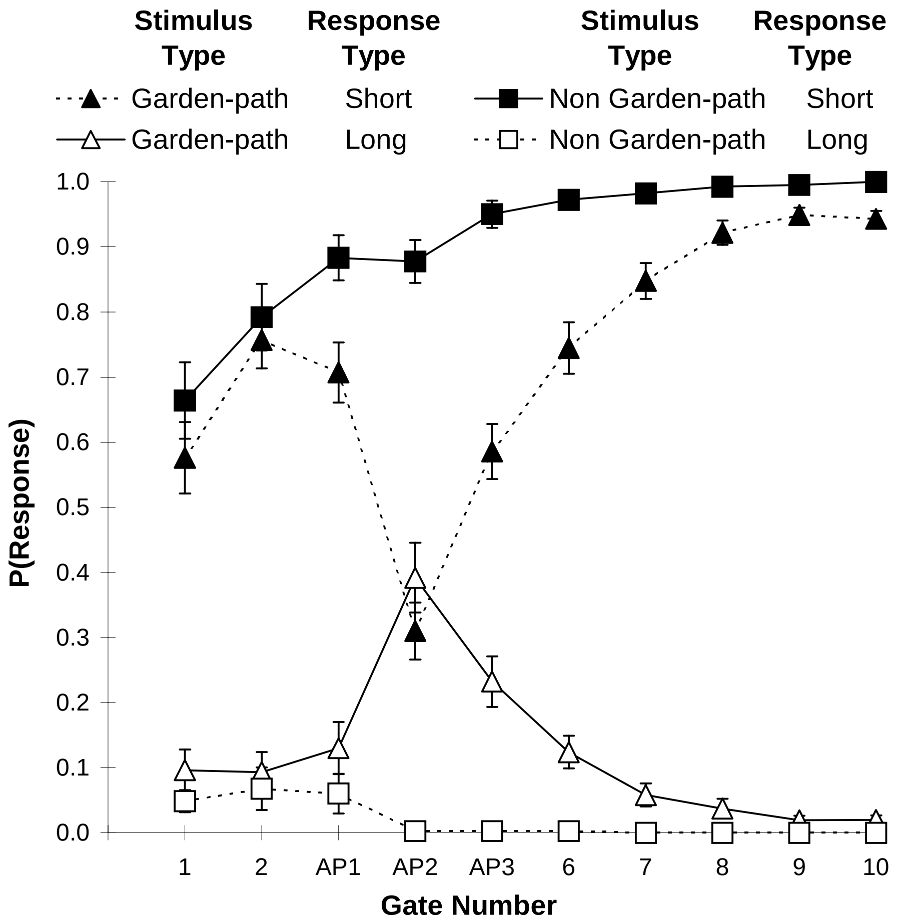


**Figure 7.1: Results of Experiment 3 – Gating. Proportion of responses matching short (CAP) and long words (CAPTAIN) for stimuli containing lexical garden-paths (*cap tucked*) and without garden-paths (*cap looking*). Error bars = 1 standard error. Lexical garden-path data from Experiment 1 (Figure 4.2)**

As can be seen in Figure 7.1, participants in this experiment no longer produced reduced numbers of short word responses at gates where the following word can be heard. This change is especially apparent at $AP_2$ where information about the onset of the following word becomes available. However, there is also a discrepancy at $AP_1$ between short word responses to stimuli with delayed or immediate mismatch ($t_1$(31)= 4.57, p<.001; $t_2$(37)= 4.35, p<.001). This indicates that for non garden-path stimuli additional information supporting the identification of onset-embedded words is available at $AP_1$. This result suggests that the earliest alignment point used in these experiments does include some information from the following context. Given the co-articulation of segments in connected speech it will be difficult to cut-off the end of a word in such a way as to exclude any influence from the onset of the following word.

The effect of garden-path sequences on the identification of embedded words in Experiment 1 can also be seen in the proportion of correct responses made at later gates. The results from the current experiment indicate that (apart from the excluded items) participants correctly identified all the target words at the final gate, whereas previously only 94% made at gate 10 correctly identified the embedded word in garden-path contexts. Significantly fewer correct responses ($t_1(31)= 4.51$, $p<.001$; $t_2(35)= 4.64$, $p<.001$) were produced for stimuli with continuations that match longer lexical items – even where mismatch is only delayed over the onset segments of the following syllable.

Another way of analysing this data is by calculating isolation points. These measure the gate at which participants produce the correct response without subsequently altering their response (Grosjean, 1996). In the analysis of Experiment 1, the isolation point of the short word stimuli was bi-modally distributed with some participants isolating the embedded word before $AP_1$ and some after $AP_3$. In the current experiment a dip in the number of correct short-word responses is no longer observed at $AP_2$. For this reason, measures of isolation point now provide an appropriate summary of the identification of the short word stimuli in non garden-path contexts. The mean isolation point for these stimuli in Experiment 3 (without *ban*, *bran* and *win*) was 268ms; which is shorter than the average duration of the target words (303ms). A paired t-test showed that these isolation points were significantly before the offset of the target word ($t(36)= 3.29$, $p<.01$).

The results of this analysis suggests that these stimuli could be correctly identified at their offset - unlike the lexical garden-path stimuli that were used in Experiment 1. It is possible that other gating experiments that demonstrated delayed (i.e. post-offset) identification of words in connected speech (Bard, Shillcock, & Altmann, 1988; Grosjean, 1985) may have inadvertently contained lexical-garden paths equivalent to those that were deliberately constructed in the short word stimuli for Experiments 1 and 2. Alternatively it may be that the stimuli used in the current experiments provide a stronger contextual constraint than those used previously, producing earlier isolation points in the gating task (Tyler & Wessels, 1983).

## 7.3. Experiment 4 - Cross-modal priming

In the concluding section of Chapter 4, the issue of response biases in gating was discussed. These may limit the accuracy of the gating task as a means of gauging the lexical activation of words in connected speech. Consequently, further studies reported in Chapter 5 used cross-modal repetition-priming to provide an on-line measure of listeners' interpretations of stimuli containing onset-embedded words. The same method will now be applied to short word stimuli with non garden-path continuations. As in Experiment 2, repetition priming will be used to measure the lexical activation of embedded words and longer words that contain the embedded word as their initial syllable. In this way the extent of ambiguities between short and long words can be established for these non-garden-path stimuli, extending the results of Experiment 2.

As in Experiment 3, this study will only use a single set of test stimuli instead of the two sets used in Experiment 2. In order to assess the activation of targets and competitors at a single probe position, four experimental conditions are required (two prime types - test and control - and two target types - short and long words). To test all four probe positions separately as in Experiment 2 would therefore require four experiments each with four versions. To reduce the number of experimental versions (and hence number of subjects) required, control prime conditions were combined so that two probe positions for the test stimuli were tested together with a single control prime. This reduced the size of the experiment to 12 conditions which were tested in two separate six version experiments.

***Participants***

Over the two repetition-priming studies carried out in Experiment 4, 114 participants from the Birkbeck Speech and Language subject panel were tested (56 on Experiment 4a and 58 on Experiment 4b). A shortage of previously untested subjects meant that approximately 20 of the subjects had taken part in Experiment 2(a). However, none had been tested on these experiments within the previous 12 months and none had been tested on the stimuli used in the current experiment (i.e. none had taken part in Experiment 3). All were native English speakers without any hearing or language impairment and were paid for their participation.

***Design***

The 40 stimulus sentences containing short prime words with non-garden-path continuations were paired with the same set of control prime sentences in which the test word is replaced with an unrelated word matched in frequency to either the short or long target (see Appendix A). Since there was only one set of test sentences, two different probe positions were examined in an experiment with three prime types (two test primes and one control prime). To cover the four probe positions assessing the activation of both short and long words, two experiments with six versions were designed. Experiment 4a measured priming at $AP_1$ and $AP_4$ (100ms after $AP_3$) and Experiment 4b probed at $AP_2$ and $AP_3$. Control primes in each experiment were played up to a point equivalent to the earlier of the two probe positions - $AP_1$ in Experiment 4a and $AP_2$ in Experiment 4b.

In all other respects both experiments were identical to each other. Each contained a total of 122 filler sentences, 27 of which were paired with a non-word phonologically related to the word at the cut-off point of the sentence, as well as 54 sentences followed by unrelated non-words and 41 followed by unrelated word targets. This produced experimental versions in which 50% of targets were words and where 33% of targets followed a phonologically similar prime. When the 20 practice items and 10 dummy items (starting each experimental block) are included, the proportion of trials in which a word target was preceded by a related test prime was just over 14%.

***Procedure***

The procedure used in Experiment 4 was identical to that used in Experiment 2 except for the different test stimuli used. These were presented up to $AP_1$ or $AP_4$ in Experiment 4a or up to $AP_2$ and $AP_3$ in Experiment 4b. Details of each test trial are as described for Experiment 2 with visual targets being presented for 200ms at the point where the speech is cut off. Once again, the experiment was divided into four sessions, an initial block of practice items, followed by two blocks of experimental items and finishing with a recognition memory test on some of the filler sentences used in the experiment.

## 7.3.1. Results

Response time data from these two experiments was analysed following the exclusion of slow or error prone participants (mean RT over 750ms or error rates of over 12.5% on the test words). These criteria led to the exclusion of 8 participants from Experiment 4(a) and 11 participants in Experiment 4(b). As in Experiment 2, the target word BRAN elicited a large number of errors and it, along with its matched pair BRANDY was excluded from further analysis. Also excluded were a number of outlying responses slower than 1200ms (4 data-points from Experiment 4a and 1 data-point from Experiment 4b). Response times and error rates for each prime and target type following these exclusions are shown in Table 7.1 below.

| | Experiment 4a | | | |
|---|---|---|---|---|
| | Short Target (CAP) | | Long Target (CAPTAIN) | |
| Prime Type<br>*Probe* | RT (ms) | error % | RT (ms) | error % |
| Test<br>$AP_1$ | 488 | 2.7 | 558 | 5.5 |
| Test<br>$AP_4$ | 480 | 1.6 | 549 | 7.6 |
| Control | 508 | 4.5 | 560 | 7.5 |

| | Experiment 4b | | | |
|---|---|---|---|---|
| | Short Target (CAP) | | Long Target (CAPTAIN) | |
| Prime Type<br>*Probe* | RT (ms) | error % | RT (ms) | error % |
| Test<br>$AP_2$ | 499 | 1.0 | 556 | 5.7 |
| Test<br>$AP_3$ | 500 | 3.5 | 569 | 5.9 |
| Control | 520 | 5.7 | 551 | 6.2 |

**Table 7.1: Results of Experiment 4a and 4b. Mean lexical decision times and error rates by prime and target type.**

Statistical analysis of this experiment used pairwise comparisons of responses following test and control primes to evaluate the magnitude and significance of priming effects. Analyses comparing the priming effects found in different experiments will use differences between normalised control and test prime RTs as the dependent measure. Analysis of response times and errors by prime and target type for each experiment are reported in Appendix B.

### *Experiment 4a ($AP_1$ and $AP_4$)*

Responses to short word targets were significantly facilitated by test primes at $AP_1$ ($\underline{F}_1[1,42]= 5.87$, $p<.05$; $\underline{F}_2[1,33]= 5.00$, $p<.05$) and at $AP_4$ ($\underline{F}_1[1,42]= 14.73$, $p<.001$; $\underline{F}_2[1,33]= 13.62$, $p<.001$). No significant priming effects were observed for long targets (all $\underline{F}_1<1$ and $\underline{F}_2<1$). Differences in error rates following test and control primes failed to show any significant differences (all p>.1 – except differences in error rates to short words following test primes at $AP_4$ which was significant by participants $\underline{F}_1[1,42]= 4.98$, $p<.05$; $\underline{F}_2[1,33]= 2.35$, $p>.1$). The magnitude and significance of the priming effects observed for each prime and target type are shown in Figure 7.2.

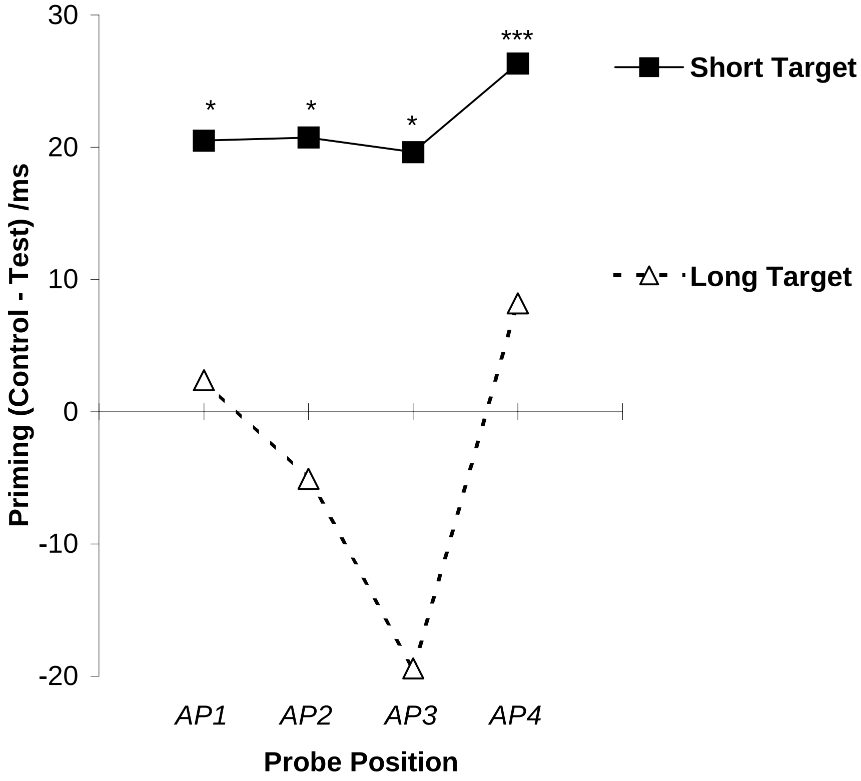


**Figure 7.2: Magnitude and statistical significance of priming in Experiment 4. Primes contain short words with non-lexical continuations (*cap looking*) with short and long word targets (CAP/CAPTAIN) presented at different probe positions (*$AP_1$, $AP_2$, $AP_3$, $AP_4$*). *** p<.001; * p<.05**

### *Experiment 4b ($AP_2$ and $AP_3$)*

Pairwise comparisons of response times following test and control primes illustrated in Figure 7.2 show significant priming for short word targets at $AP_2$ ($\underline{F}_1[1,41]= 7.39$, $p<.01$; $\underline{F}_2[1,33]= 4.72$, $p<.05$) and at $AP_3$ ($\underline{F}_1[1,41]= 4.09$, $p<.05$; $\underline{F}_2[1,33]= 4.77$, $p<.05$). Responses to long word targets in contrast, tended to be slowed following test primes.

This interference effect was not significant at $AP_2$ ($\underline{F}_1$<1, $\underline{F}_2$<1) though it approached significance at $AP_3$ ($\underline{F}_1$[1,41]= 1.89, p>.1; $\underline{F}_2$[1,33]= 4.18, p<.05). Comparisons of error rates showed that there were fewer lexical decision errors for short words following test primes than following control primes at $AP_2$ ($\underline{F}_1$[1,41]= 9.11, p<.01; $\underline{F}_2$[1,33]= 9.20, p<.01). All other comparisons of error rates were non-significant (all p>.1).

Pairwise comparisons of errors made following test and control primes indicate that participants made fewer lexical decision errors for short words following test primes than following control primes. This effect was significant following primes presented up to $AP_2$ ($\underline{F}_1$[1,41]= 9.11, p<0.01; $\underline{F}_2$[1,33]= 9.20, p<0.01) though not at $AP_3$ ($\underline{F}_1$[1,41]= 2.05, p>0.1; $\underline{F}_2$[1,33]= 2.41, p>0.1). There was no significant difference in error rates to long word targets following test primes at either probe position compared to controls (all $\underline{F}_1$<1 and $\underline{F}_2$<1).

### *Combined analysis (Experiment 4a and 4b)*

Combined analysis of priming effects in Experiment 4a and 4b were carried out using data that had been normalised by dividing each response time by the mean for that participant and multiplying by the overall mean response time in both experiments. Test-control difference scores for this normalised data were entered into two-way ANOVAs using the factors target type (short or long words) and probe position. There was a highly significant effect of target type ($\underline{F}_1$[1,186]= 17.70, p<.001; $\underline{F}_2$[1,38]= 7.68, p<.01) reflecting greater priming of short target words across all four probe positions (see Figure 7.2). There was no main effect of probe position ($\underline{F}_1$[3,186]= 1.32, p>.1; $\underline{F}_2$[3,114]= 2.41, p<.1) nor any significant interaction between target type and probe position ($\underline{F}_1$<1; $\underline{F}_2$<1). As in the analysis across different probe positions for Experiment 2, the lack of a significant effect of probe position suggests that interpretations of the stimuli (as indicated by the magnitude of priming) did not significantly change as participants heard more of the prime sentences.

## 7.4. Comparison of Experiment 2 and Experiment 4

Analyses of the two sets of repetition priming experiments were carried out to compare the magnitude of repetition priming for short word stimuli with garden-path and non-garden path following contexts. These ANOVAs used normalised difference scores as the dependent measure, to minimise differences produced by varation between the groups of subjects tested in each part of the two experiments. Analyses with participants as the random variable included the prime type factor (garden-path vs. non garden-path stimuli) as a between participants comparison (since these are the results of separate experiments) while analyses by items were carried out with both prime and target type as repeated-measures factors. The magnitude of priming for short and long target words from embedded words with garden-path continuations (*cap tucked*) and stimuli without lexical garden-paths (*cap looking*) is shown in  Figure 7.3.

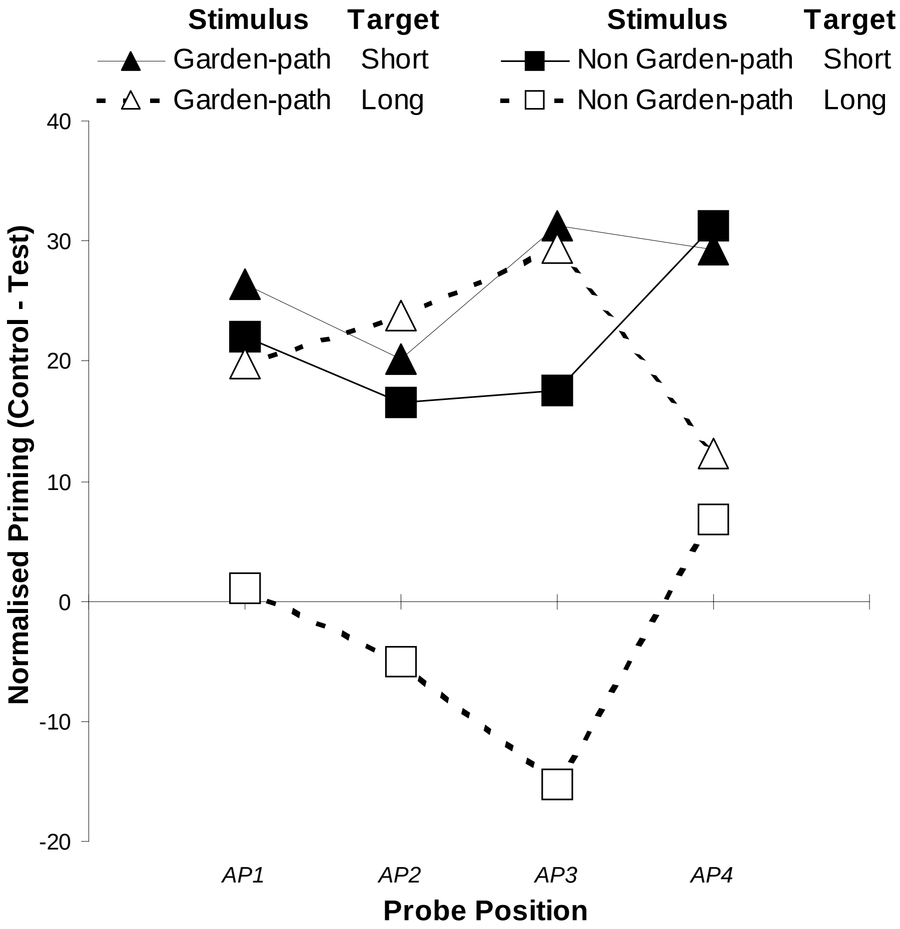


**Figure 7.3: Combined results of Experiments 2 and 4. Normalised priming of short (CAP) and long (CAPTAIN) targets for garden-path stimuli (*cap tucked*) and non garden-path stimuli (*cap looking*) at different probe positions.**

Analyses showed a significant main effect of target type ($\underline{F}_1$[1,400]= 15.09, p<.001; $\underline{F}_2$[1,38]= 4.78, p<.05) indicating greater overall priming for short target words than for long targets. This greater priming of short words is unsurprising given that both sets of prime stimuli involved in this comparison contained short words. A significant main effect of prime type (lexical garden-path vs. non garden-path) was also observed in these analyses ($\underline{F}_1$[1,400]= 9.40, p<.01; $\underline{F}_2$[1, 38]= 11.21, p<.01). The total magnitude of priming was greater for the lexical garden-path sequences than the non garden-path sequences.

An interesting effect here is the interaction between prime type and target type - though this is only marginally significant in the analysis by participants ($\underline{F}_1$[1,400]= 2.86, p<.1; $\underline{F}_2$[1,38]= 4.61, p<.05). As Figure 7.3 shows, the lexical garden-path stimuli used in Experiment 2 produce significant priming of both short and long words at most probe positions, while the non garden-path stimuli in Experiment 4 only facilitate short word targets.

There was no main effect of probe position in this analysis, nor any interaction between probe position and prime or target type. However, pairwise comparisons at individual probe positions suggest that differences in the activation of long lexical items for stimuli with and without lexical garden-paths are observed at specific probe positions. At $AP_1$ and $AP_4$ there are no significant differences between the priming of long words in Experiments 2 and 4 (at $AP_1$, $\underline{F}_1$[1,111]= 1.67, p>.1; $\underline{F}_2$[1,38]= 2.75, p>.1. At$AP_4$, $\underline{F}_1$<1; $\underline{F}_2$<1). However, differences in the priming of long targets do emerge at $AP_2$ and $AP_3$. Long word targets are primed more strongly in Experiment 2 than in Experiment 4 at $AP_2$ ($\underline{F}_1$[1,70]= 3.37, p<.1; $\underline{F}_2$[1,38]= 5.05, p<.05) and at $AP_3$ ($\underline{F}_1$[1,102]= 6.97, p<.01; $\underline{F}_2$[1,38]= 6.93, p<.05)

Thus for stimuli that mismatch with the longer word immediately after the offset of the embedded word (*cap looking*), reduced priming is observed for long targets earlier in the speech stream than for lexical garden-path primes (*cap tucked*). These effects of the presence or absence of garden-path continuations suggest that bottom-up mismatch acts to support or disconfirm alternative lexical hypotheses – even where the competing interpretations do not share word boundaries.

In contrast to the results for long targets, there were no significant differences in the priming of short words in garden-path and non garden-path stimuli at any probe position (all p>.1). Caution is required in interpreting null results in between-experiment comparisons, however the presence or absence of garden-path continuations appears to have no significant effect on the priming of short word stimuli – contrasting with the gating results presented previously. The implications of these results for models of spoken word recognition will be discussed in the concluding section of this chapter.

## 7.5. Comparing experimental data and recurrent network simulations

The recurrent network model developed in this thesis suggests that mismatch between continuations of embedded words and longer lexical items plays an important role in the identification of onset-embedded words. Simulation 1, reported in Chapter 3, did not incorporate any other input cue to distinguish syllables of short and long words. Consequently, lexical units for onset-embedded words were only fully activated when longer competitors were ruled out by mismatching following contexts. Thus, mismatching input provided the only cue that allowed onset-embedded words to be identified. The networks described in Chapter 3 therefore predict marked differences in the activation of short word units depending on whether the following context of an embedded word forms a lexical garden-path with a longer word or not (for an example, compare the activation profile for *cap* in Figure 3.4b and Figure 3.5b).

Since the simulations reported in Chapter 6, however, included an input cue analogous to the acoustic difference between syllables in short and long words. These networks no longer rely solely on post-offset mismatch to recognise embedded words. Therefore, it is unclear the extent to which networks in Simulation 4 still predict differences in activation for short word stimuli depending on their following context. Further tests of these networks were therefore carried out to compare the activation of short and long word units in response to stimuli containing short words followed by garden-path and non garden-path contexts. Results from these simulations can then be compared to that obtained from the combined priming data from Experiments 2 and 4.

Data for the garden-path sequences will be identical to that used in simulating the results of Experiment 2. The non-garden-path stimuli were newly generated sequences in which the initial segment following the offset of the embedded word mismatched with the longer lexical item that contained the embedded word. Since there was only one longer word containing each of the embedded monosyllables in the networks vocabulary these sequences were easier to generate than the equivalent experimental stimuli. All the stimuli included the embedded word as the second word in a sequence so that, although presented with an ambiguous duration, cues to the length of the target word could be detected by the network. The word before the test item was either monosyllabic or bisyllabic and matched neither the embedded word nor its longer competitor. The test set contained 32 combinations of prior contexts, embedded words and following contexts. As in the previous chapter, the performance of the ten networks trained for Simulation 4 was investigated by averaging across all the items on which each network was tested. Analyses of activations produced for different test stimuli at different lexical units were carried out as repeated-measures comparisons across these ten computational subjects.

In simulating the results of experiments using fragments of prime sentences it is necessary to decide what position in the sequence of segments presented to the network is equivalent to each probe position tested in the experiments. As in the simulations of priming data reported in Chapter 6, it was assumed that the descriptions of the alignment points given in Chapter 4 provide a reasonable estimate of the phonemic information that is available at each probe position. Thus, $AP_1$ is placed at the offset of the embedded word with no information about following context available at this point. Although there is evidence from short word responses in gating to suggest that some information from the following word is available at $AP_1$, since the network's input is coded as discrete segments, it is not possible to simulate effects of coarticulated information. $AP_2$ is placed following the onset of the initial segment of the following word – a point at which there will be information that mismatches with the longer competitor for the non garden-path stimuli, but where the garden-path stimuli still match a longer lexical item. $AP_3$ and $AP_4$ were placed after the vowel and offset segments of the following syllable.

### 7.5.1. Recognising embedded words

Figure 7.4a shows the activation of short lexical units in the recurrent network for embedded word stimuli presented in garden-path and non garden-path contexts. These lexical activations are shown alongside equivalent results obtained at each probe position in the priming experiments reported earlier in this chapter (Figure 7.4b). As can be seen by comparing the two graphs, the network overestimates the effect of following context on the activation of short word units. For the non garden-path stimuli, input that mismatches with longer lexical items (at $AP_2$) leads the network to maximally activate the embedded word. Thus at all probe positions after $\mathrm{AP}_1$ the network produces significantly greater activation for short word units in non garden-path sequences ($AP_1$ - t(9)= 8.96, p<.001; $AP_3$ - t(9)= 4.70, p<.001; $AP_4$ - t(9)= 4.03, p<.01) while the equivalent comparisons for the priming data are non-significant (and show a numerical trend in the reverse direction).

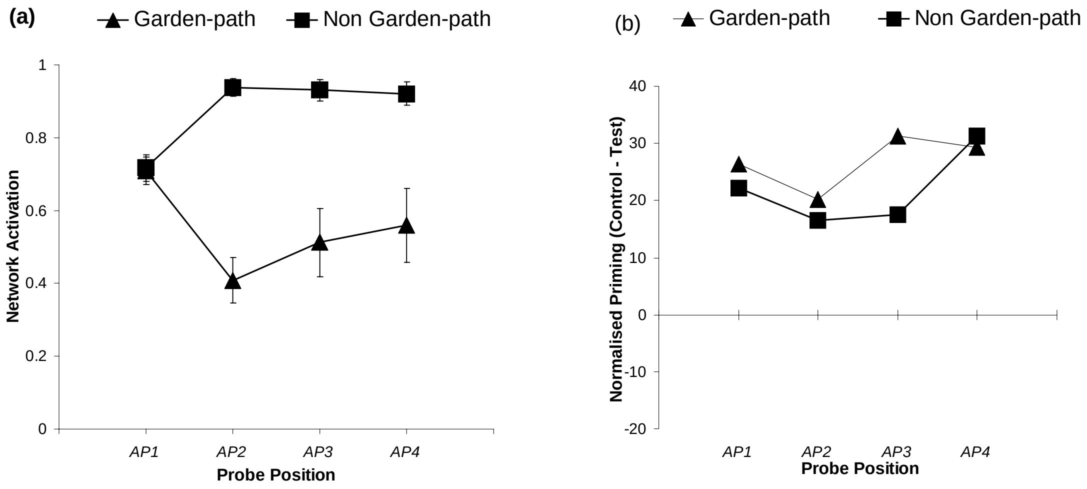


**Figure 7.4: Network activation and priming results for short lexical items (CAP) in response to short word stimuli in garden-path (*cap tucked)* and non garden-path contexts (*cap looking*): (a) short word activations over ten networks in Simulation 4. Error bars are one standard error (b) priming results for short targets in Experiments 2 and 4.**

This discrepancy between lexical activations observed in the networks and priming data suggests that the model would not predict the null result obtained in comparing the priming of short words in garden-path and non garden-path sequences. Despite the presence of acoustic cues to word length that bias the network towards short word hypotheses at the offset of an embedded word, following context still affects the activation of lexical units for short words. Where following context rules out longer competitors it boosts the activation of short lexical items in the model. No equivalent increase in the priming of short words in non garden-path contexts is observed in the experimental data.

Although the recurrent network simulations reported here lack direct inhibitory connections between lexical units, effects of competition between short and long lexical items are still observed. The comparison between network activations and priming data in Figure 7.4 suggests that even without direct lexical competition, these networks still predict an effect of following context that is not shown in the cross-modal priming data. Explanations of this discrepancy are explored in the concluding section of this chapter.

### 7.5.2. Ruling out longer competitors

Despite the lack of any significant effect of following context on the activation of short words in the priming data, it is clear that mismatching input is being perceived by listeners. Effects of mismatch were shown by significant differences between the magnitude of priming for long word targets in different following contexts. Greater priming is observed for long word targets from short word stimuli in garden-path contexts, than from stimuli in which there is immediate mismatch between short word stimuli and long lexical items.

The networks from Simulation 4 succeed in simulating this difference between the activation of long words produced by embedded word stimuli with garden-path and non garden-path following contexts. As illustrated in Figure 7.5a, the model predicts significantly greater activation for long words where the following context of an embedded matches a longer competitor compared to sequences where long words are ruled out earlier. This increased activation was significant by pairwise comparisons at all three probe positions where information in the following context was available ($\mathrm{AP}_2$ – t(9)= 15.35, p<.001;$AP_3$ – t(9)= 6.09, p<.001;$AP_4$ – t(9)= 4.03, p<.01).

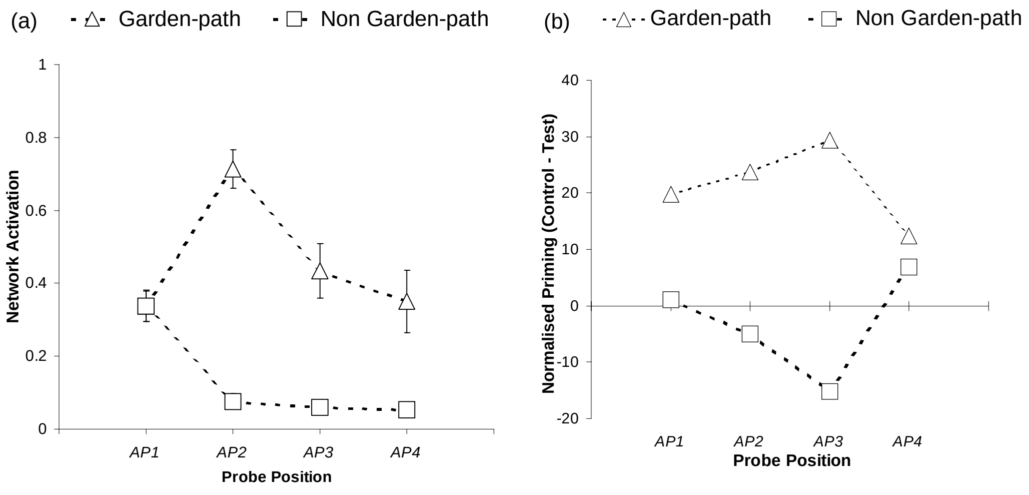


**Figure 7.5: Network activation and priming results for long lexical items (CAPTAIN) in response to short word stimuli in garden-path (*cap tucked)* and non garden-path contexts (*cap looking*): (a) long word activations over ten networks in Simulation 4. Error bars are one standard error. (b) priming results for long targets in Experiments 2 and 4.**

However, in the priming data shown in Figure 7.5b, differences in priming are only reliable at $AP_2$ and $AP_3$. The numerically greater priming observed for long words from garden-path primes was not significant at $AP_1$ or at $AP_4$. Compared to the priming data, it seems that the network has greater difficulty in ruling out longer lexical hypotheses for the short word stimuli at the last probe position.

## 7.6. General discussion

The networks in Simulation 4 are therefore able to account for both the results of Experiment 2 and also the reduced priming of long word targets in Experiment 4. However, some significant discrepancies remain in simulating the priming data for short word targets in Experiment 4. The recurrent networks in Simulation 4 predict that increased priming will be observed for onset-embedded words in contexts where longer words are ruled out immediately after the offset of the embedded word. Since increased priming was not observed in the experiments, it appears that the model has problems accounting for the integration of lexical and acoustic cues to word segmentation. However, before interpreting these results as ruling out the recurrent network account, it is worth considering whether any confounding factors in the experiments or assumptions made in interpreting network activations are responsible for the discrepancies between the networks' predictions and the experimental results.

### *Alternative explanations of experimental data*

Perhaps the simplest argument against interpreting the experimental data as falsifying the model is to suggest that differences between the design of the two sets of experiments are responsible for the inconsistent results. For instance, in Experiment 2 investigating garden-path continuations, prime stimuli came from both short and long words whereas in Experiment 4 prime sentences all contained short words. It is therefore possible that strategic effects produced by different forms of prime-target overlap in the two sets of repetition priming experiments may introduce discrepancies between the priming effects obtained for garden-path and non garden-path stimuli.

In Experiment 4, strategic effects might result from participants noticing that only short word primes are present. These might lead to decreased overall priming of long word targets – explaining the inhibition observed for long word targets at $AP_3$. However, this strategic difference would not account for the priming results for short word targets. A strategic explanation would predict increased priming for short words, while this result (predicted by the network through effects of following context) was absent from the experimental data. Since the lack of significant differences between priming effects for short words in garden-path and non garden-path stimuli is the most difficult result to explain, it seems that strategic differences between the two sets of priming experiments do not help the recurrent network account.

Another concern in interpreting the experimental data is that the comparison of garden-path and non garden-path stimuli does not take place within a single experiment. The conclusion that following context does not affect the activation of short word stimuli is a null result based on a between-experiment comparison. It is therefore possible that the current set of experiments have insufficient statistical power to detect a difference between garden-path and non garden-path stimuli. A within-experiment replication would therefore be valuable in supporting the conclusions drawn from these experiments. In the absence of this data, however, it is necessary to consider whether assumptions made in relating priming data to network activations could account for the failure of the model to simulate the experimental data.

### *Relating priming data to network activations*

In considering how best to interpret this discrepancy between results in the second set of priming experiments and the predictions of models of spoken word recognition, one interesting finding is that in the gating experiment onset-embedded words are more easily identified in non garden-path contexts. Thus, ignoring the overall bias towards short word responses, the pattern of results shown in gating (Figure 7.1) are rather closer to the network's predictions than the priming data. Although not intended to suggest that gating data is to be preferred over results in cross-modal priming, this indicates that behavioural data obtained in psycholinguistic experiments cannot be relied upon to provide a transparent measure of the internal processes of the language processing system.

In comparing network simulations to experimental data, the priming task has not been modelled directly. Instead the simplifying assumption has been made that the activation of lexical units in the network predicts the magnitude of priming observed in more complete simulations that incorporated orthographic inputs and a lexical decision mechanism. However, this assumption may be difficult to justify in the absence of a rather more sophisticated model of language processing. For instance, the recurrent network model investigated by Gaskell & Marslen-Wilson (1997) includes separate output representations for the phonological form and meaning of the speech input. In simulating the results of semantic and repetition priming experiments, Gaskell and Marslen-Wilson assume that semantic priming is predicted by overlap in the semantic representation, while repetition priming results from similarity in representations of both phonological form and meaning. This account therefore predicts a different relationship between the conditional probability of a word given the current input and the amount of facilitation that is observed of semantically related or repeated targets (Gaskell and Marslen-Wilson, in press). These predictions are confirmed by experiments investigating the relationship between the likelihood of different cohort competitors being present in fragments of speech, and the magnitude of semantic or repetition priming observed.

For semantically related targets, the model predicts that the magnitude of priming is directly proportion to the likelihood of the prime word being present in the input. Where this probability is near zero – for instance, where a much more frequent word also matches the speech input – no reliable priming is predicted. For example, a fragment of speech like /strɪː/ would not be predicted to prime a word semantically related to *streak* since this word has a low probability for this fragment (given the presence of a much more frequent cohort competitor *street*). Results of a series of cross-modal priming experiments investigating how the magnitude of semantic priming varies with competitor environment show that the magnitude of semantic priming is directly proportional to the probability of a word given the current input. No priming was observed in these experiments as the conditional probability of the prime word in its cohort environment tends towards zero.

Conversely, in repetition priming, even a word with a low probability of occurrence in its cohort set can be significantly primed through form overlap with its phonological neighbours. So, for the example given above, significant facilitation of responses to the target word *streak* from the spoken fragment /strɪː/ would be observed in cross-modal repetition priming. In the model, graphs plotting the magnitude of repetition priming against conditional probability have a significant offset at the origin, as a consequence of the priming effect observed through form-based overlap. These results confirm the predictions of the Gaskell and Marslen-Wilson model since the magnitude of repetition priming, although varying with conditional probability, is not directly related. Form-based priming can arise even where the probability of a word in its cohort set is near zero.

On the basis of these findings, measures of repetition priming may not relate as closely to the activation of lexical/semantic representations as has been assumed in this thesis. Significant repetition priming can be obtained where there is minimal lexical activation of the target word but there is form overlap between the prime and target. Consequently the magnitude of priming effects measured in Experiment 2 and Experiment 4 may not reflect the conditional probability of words in the speech input as directly as was assumed in relating repetition priming data to network simulations. In particular, since form based priming can arise where lexical activations are low, priming effects for short words may be relatively unaffected by the activation of longer competitors.

Since the magnitude of semantic priming has been shown to be more closely correlated with the activations predicted by a probabilistic model (Gaskell & Marslen-Wilson, in press) it may be of interest to use semantic priming rather than repetition priming in follow up experiments. Such data would provide a more stringent test of the recurrent network account developed here in which lexical activations are suggested to be directly proportional to conditional probability.

### *Lexical activation and probabilistic behaviour*

Given these difficulties in relating the predictions of the recurrent network model to behavioural data obtained in priming experiments the status of the probabilistic account of lexical activation presented here remains unclear. The lack of any significant change in priming effects for embedded words depending on whether longer competitors have been ruled out presents a challenge to a probabilistic account. However, it is unclear that these discrepancies would falsify the model rather than simply suggesting that the statistical structure of the network's training environment does not match the properties of natural language.

For instance, the statistical properties of the duration cue used in the network simulations is highly simplified compared to real speech. In order to ensure that the network did not focus on duration at the expense of the segmental input, the duration cue was made deliberately unreliable in the network. Thus, longer words were still robustly activated at the offset of an embedded word, increasing the effect of following context in the network. If the duration cue were made more reliable for the network, then longer competitors would be more weakly activated initially and following context would play a reduced role in altering the activation of the embedded word.

Secondly, the recurrent network simulations presented in this thesis use a highly impoverished vocabulary by comparison with experimental participants' lexicons. Each embedded word had only one longer competitor in the network, and very few neighbouring lexical items. As a result, only two words were activated during the identification of embedded word stimuli. The network is therefore likely to treat evidence ruling out a longer competitor as ruling in the embedded word. In simulations with multiple long words that all contain the embedded word, this symmetrical competition between short words and a single long competitor would no longer be apparent. Thus networks with a more accurate competitor enviroment may not predict such large differences between garden-path and non garden-path stimuli.

The failure of the recurrent network account to simulate the experimental data therefore need not indicate that the architecture or computational properties of the system are inappropriate – merely that the training set does not adequately capture the statistical structure of the language. This conclusion is hardly surprising; the training set used for the simulations in this thesis contained only 20 lexical items. However, this finding suggests that caution be exercised in interpreting the failure of the network to exactly simulate the behavioural profile observed in a particular experiment. Small network simulations may be best interpreted as demonstrations of the processing possibilities offered by a particular computational architecture rather than specific predictions about the results of behavioural experiments. In order to simulate behavioural data at an item specific level more realistically sized training sets must be used.

# 8. Concluding Remarks

The research reported in this thesis has investigated the segmentation and recognition of words in connected speech. An account has been developed that uses a simple recurrent network trained to map from a sequence of input segments to an output representation of the lexical/semantic content of that sequence. This final chapter discusses ways in which this distributed recognition system contrasts with other computational models that use localist representations and that include direct, inhibitory connections between units representing competing lexical items. This chapter also describes conclusions drawn from experimental investigations that were designed to test the predictions of these different accounts and proposes future work to develop and test this recurrent network model further.

A critical test case used throughout this thesis to evaluate these different computational accounts has been the recognition of words that are embedded at the onset of longer words. In the introductory chapters of this thesis, arguments were reviewed suggesting that direct, inhibitory competition between lexical items is necessary to account for the identification of embedded words. Lexical competition provides a mechanism by which a following context that rules out longer competitors can boost the activation of embedded words, allowing their identification. However, simulations described in Chapter 3 demonstrate that the crucial property of systems like TRACE and Shortlist is not the inclusion of direct intra-lexical competition, but that the recognition system is not confined to using sections of the speech stream to identify only a single word at a time. Recurrent networks in which the goal of the recognition process is to activate a representation of an entire sequence of words are also capable of using post-offset information to identify onset-embedded words. The approach taken in this thesis has interesting implications for theories of language comprehension and acquisition.

## 8.1 Lexical representation in a distributed system

Training a recurrent network to activate a representation of an entire sequence of words was suggested in Chapter 3 to have an interesting developmental interpretation. The assumption made by these networks is that the task involved in learning to understand spoken language involves a mapping from whole utterances of connected speech to the intended meaning of that entire sequence. Thus the network learns the mapping from speech to meaning in the absence of explicitly segmented input in either the spoken or the conceptual domain and without being supplied with one-to-one correspondences between units in the speech stream and units of meaning. Although systems that learn the statistical properties of the speech stream have been described that provide a preliminary segmentation of the speech stream into lexical units, the account proposed here suggests that the structure of adult lexical representations is primarily determined by the properties of the mapping from speech to meaning.

In this view of language comprehension, lexical representations emerge as the fundamental unit of regularity in this mapping between speech and meaning. As has been proposed in other domains, such as reading aloud (Plaut, McClelland, Seidenberg and Patterson, 1996) and derivational morphology (Gonnerman, Devlin, Anderson and Seidenberg, submitted) the advantage of the distributed connectionist approach is that these systems are not committed to extracting structure at a single level of representation. Thus the modeller need only specify the input and output representation; during training the network will develop appropriately structured internal representations to capture the statistical regularities that exist in the mapping.

For instance, experimental evidence described in Chapter 1, supported by the results of dictionary searches in Chapter 2, suggest that many morphologically complex words in English are decomposed into their constituent morphemes at a lexical level. In a connectionist account of the form-meaning mapping (Gonnerman, et al., submitted) this decomposed representation arises as an emergent property of the regularities that exist between the form of a particular morpheme (such as *happy*) and the meaning of semantically transparent derived forms (*happily*, *happiness*, *unhappy*, etc.). Importantly, this distributed system will also acquire the correct form-meaning mapping for semantically-opaque derived forms (such as *department*, which is unrelated to the embedded morpheme *depart*), in which lexical representations are suggested not to be decomposed (Marslen-Wilson, et al, 1994). Conversely, localist connectionist accounts require separate systems to account for items that are morphologically decomposed and items that are processed at a whole word level (see for instance Schreuder and Baayen, 1995).

The strength of the distributed account is thus that a multiplicity of different ‘grain-sizes’ of lexical representation can co-exist within a single system. Recent experimental evidence suggesting that common word combinations (such as *first lady* or *greasy spoon*) are also lexically represented (Harris, 1994; 1996) would therefore not require additional computational mechanisms to be accomodated within this account. In a distributed form-meaning mapping, these combinations would be lexically represented where they capture regularities in the form and meaning mapping that could not be accounted for by combining the representations of single words.

Importantly, the modelling work presented here is not intended to suggest that form-meaning mappings are the only means by which the segmentation of the speech stream can arise. Simulations reported Chapter 3 demonstrate the role of distributional information (as simulated through the inclusion of input-prediction tasks in these networks) in assisting lexical acquisition. The recurrent networks trained in Simulation 2 demonstrate that the inclusion of prediction tasks significantly speeds the acquisition of the form-meaning mapping. Thus, the systems investigated here provide a concrete illustration of the role of statistical learning in boot-strapping lexical acquisition. However, since the goal of lexical segmentation is to extract meaningful units in the speech stream, the model proposed here, in which lexical representations capture form-meaning regularities, will account for an important aspect of the language comprehension system.

## 8.2 Lexical segmentation and identification

The model developed in this thesis is not only intended to illustrate the role of different sources of information in lexical segmentation and vocabulary acquisition: it is primarily proposed as an account of the time-course of identification of words in connected speech. In this context, it is of interest that recurrent network accounts predict a different activation profile in identifying onset-embedded words than do localist systems that incorporate direct, inter-lexical competition. Where multiple lexical items match the speech stream, accounts incorporating direct competition predict increased activation for short word hypotheses. Conversely recurrent networks display probabilistic behaviour, in which multiple candidates are each activated in proportion to the conditional probability of that item being present in the current input, irrespective of length. Thus the recurrent network simulations reported in Chapter 3 predict that embedded words and longer competitors will be equally activated where both words match the speech stream.

Experiments reported in Chapters 4 and 5 were carried out to investigate the time course of identification of onset-embedded words and longer competitors in order to test these alternative accounts. Sentences were created in which these two competing interpretations were equally plausible. For the short, embedded word stimuli, following contexts were generated that created a ‘lexical garden-path’ with the longer word. The presence of segments at the onset of the subsequent word that matched the second syllable of the longer lexical item would be expected to maximise the ambiguity between short and long words. These stimuli will therefore provide the most stringent test of predictions following from the two computational accounts that have been described.

However these experiments in fact produced the novel result that early on in the processing of the test sequences, stimuli containing embedded words and longer competitors are not as ambiguous as would be predicted by both of these models of spoken word recognition. Gating results reported in Chapter 4 showed that responses to matched stimuli containing onset-embedded words and longer competitors differed from the earliest point tested. Since this test position occurs before the stimuli diverge phonemically these results suggest that short and long words can be distinguished before the point predicted by computational models that use a phonemically coded input.

Results obtained in the repetition priming experiments reported in Chapter 5 provide an even clearer demonstration that non-phonemic cues can be used by the perceptual system to distinguish long words from shorter lexical items that are embedded at their onset. In Experiment 2 no significant priming was observed from a sentence containing the word *captain* to an embedded word like *cap* - even where the prime sentence was cut off at the offset of the syllable /kæp/. The lack of significant priming of embedded words from a matching syllable of a longer word presents a considerable challenge to models that predict that onset-embedded words create substantial ambiguities during the processing of connected speech. Some additional cues must therefore be present in the speech stream to enable the perceptual system to distinguish embedded words from the start of longer competitors.

## 8.3 Acoustic cues to word boundaries

As described in the review of the acoustic-phonetics literature in Chapter 2, various acoustic cues have been proposed that may assist the recognition system in detecting word boundaries. The most reliable of these cues that could be measured in these experimental stimuli were differences in the duration of segments and syllables in short and long words. For this reason, results indicating the early differentiation of syllables from short and long words (at positions where duration cues, but not segmental cues to word boundaries are likely to be available) were taken as evidence that syllable duration provides an important cue to the detection of word boundaries. Input cues analogous to syllable duration were therefore incorporated into the recurrent network account in order to simulate the time-course of identification of onset-embedded words.

An important computational property of the syllable duration cue is that it is likely to require adaptive processing of spoken sequences to be used as a cue to word boundaries. Since multiple factors that can alter the duration of a spoken syllable (such as speech rate, metrical stress and the location of prosodic boundaries) compensation for some or all of these factors will be required to detect the small, but reliable differences in the duration of syllables in short and long words. Therefore, in order to use syllable duration to discriminate between short and long words additional processes may be necessary to compensate for changes in syllable duration caused by these alternative factors.

Simulations described in Chapter 6 incorporated a duration code that depended not only on the length of the word from which the syllable was taken but also on the overall rate at which the sequence was presented. This code was used to create sequences which included an identical duration for embedded syllables in short and long words. These short and long words could therefore only be disambiguated where prior context was used in processing. These network simulations were able to increase the activation of appropriate lexical units depending on whether an ambiguous syllable came from a short or a long word. This work therefore shows that recurrent networks are capable of the adaptive processing of an input analogous to syllable duration.

Networks trained in these simulations were also able to simulate the time course of activation of short and long words as was inferred from the cross-modal priming experiments reported in Chapter 5. Thus the recurrent network model developed here provides an appropriate account of the integration of phonemic and non-phonemic cues in the identification of onset-embedded words. However, despite this agreement between the experimental results reported here, and simulations that include a duration cue, further experiments are required to demonstrate that it is syllable duration that provides the acoustic cue to word length that is required to account for the experimental data. Only by directly manipulating the duration of syllables in short and long words can experiments establish that differences in syllable duration are responsible for the differential activation of onset-embedded words and longer competitors in these cross-modal priming experiments.

One further prediction of these recurrent networks is that where preceding speech is not presented, or presented at an inappropriate rate (preventing the adaptive processing of duration cues to word length) increased ambiguity of embedded words and longer competitors will result. Since speech rate in prior contexts have been shown to affect VOT boundaries in the discrimination of voiced and unvoiced stop consonants (Wayland, Miller and Volatis, 1994), effects of speech rate on the perception of syllable duration in short and long words would not be unexpected. However, in order to conclude that syllable duration is processed relative to preceding context requires experiments in which altering the preceding context of embedded words affects the perception of syllables taken from short and long words.

## 8.4 Sequential recognition, lexical competition and embedded words

The presence of acoustic cues that distinguish short words from the onset of words in which they are embedded might be used to rehabilitate accounts in which segmentation occurs through the early identification of words in connected speech, such as in the original sequential-recognition form of the Cohort model (Marslen-Wilson and Welsh, 1978). Acoustic cues to word length could allow the pre-offset identification of embedded words, allowing the recognition system to use lexical identification to determine the location of word boundaries even for onset-embedded words.

However, the results of Experiment 2 also demonstrated that long words continue to be activated after the offset of an embedded word in garden-path sequences like *cap tucked* (as indicated by significant priming of targets like *captain*). Consequently, not all ambiguity can be resolved by the acoustic offset of an embedded word. This result is evidence that post-offset information plays a role in the recognition of words in connected speech. If, as described in sequential recognition accounts, embedded words are identified at, or before, their acoustic offset, longer candidates would not need to be ruled out during the following contexts of these embedded words.

Further evidence of the role of following context was obtained in the cross-modal priming experiments reported in Chapter 7. These experiments indicate that where information mismatching with longer words appears earlier in the speech input, longer words can be ruled out more rapidly and more effectively than in the lexical garden-path sequences used in Experiment 2. These results provide evidence supporting the use of bottom-up mismatch to rule out inappropriate lexical hypotheses. Models such as TRACE (McClelland & Elman, 1986) that do not allow mismatching input to directly reduce the activation of inappropriate lexical hypotheses would therefore be challenged by this data. Either Shortlist (Norris, 1994) or the recurrent network account developed in this thesis would be able to simulate this data since both of these models allow mismatching information to decrease the activation of lexical candidates through bottom-up inhibition.

However, while there is evidence that information after the offset of a word is used to rule out mismatching competitors, comparisons of Experiments 2 and 4 suggest that post-offset mismatch does not increase the activation of embedded words. Equal priming of short targets is observed from sequences where longer competitors are ruled out immediately after the offset of the embedded word, and from sequences where the following context creates a garden-path matching a longer word. This result appears to go against the predictions of both recurrent network and lexical competition accounts of spoken word recognition. As suggested in the previous chapter, this null-result in a between-experiment comparison may only indicate a lack of statistical power. Follow-up investigations using a within-experiment comparison are worthwhile.

## 8.5 Summary and future directions

In this concluding chapter, several further experiments have been proposed, to help establish the correct interpretation of some of the experimental results presented in this thesis. However, as has been argued throughout this thesis, interpretations of experimental data may not be suitably constrained where there is not an implemented computational model available for comparison. The computational modelling work presented in this thesis, has demonstrated that recurrent neural networks provide a powerful and flexible processing system for spoken word recognition. These networks have been shown to have interesting computational properties that are appropriate for the segmentation and identification of words in connected speech.

In order to go beyond these ‘demonstrations’, however, and produce a model capable of a detailed account of experimental data requires rather more from a network simulation. To produce a good quantitative match to empirical data the system must do more than just display the appropriate computational properties (such as sensitivity to following context, or adaptive processing of duration). A complete network account should also be able to simulate experimental data on an item-by-item basis. This, however, would require representations that adequately capture the properties of the input and output domains. This may be rather harder to achieve in modelling spoken word recognition than in modelling the processes involved in reading aloud (Plaut, et al, 1996) where input and output representations can be more easily specified. A complete model of spoken word recognition would also require a realistically sized and appropriately structured vocabulary (including full competitor environments and detailed morphological structure). Finally, a selection of input and output routes and attendant control process would also be required to simulate the different patterns of behavioural data obtained in different tasks.

Computational psycholinguistics may in future be capable of addressing these various challenges. It is, however, possible that the wealth of empirical data provided by neuro-imaging techniques will reduce the importance of simulating behaviour as a goal for computational modelling. It may be that psycholinguistics in the new millennium will see the constraints provided by neuro-imaging data as being of greater importance than behavioural data in explaining spoken language comprehension.

# Appendix A – Stimulus sentences

Target words (short/long) appear in capitals following by stimulus sentences in the following order: (a) short word test stimuli (Experiments 1 and 2), (b) long word test stimuli (Experiments 1 and 2), (c) short word control prime stimuli (Experiment 2), (d) long word control prime stimuli (Experiment 2), (e) short word test prime with mismatching continuation (Experiments 3 and 4). Critical words are emphasised.

1. ANT/ANTLER

   a. It is because the **ant** lived under the rocks that it survived the explosion.
   b. It is because the **antler** is fully grown that you can tell the deer is male.
   c. It is because the **horn** was so loud that we all jumped.
   d. It is because the **trumpet** was so loud that we all jumped.
   e. It is because the **ant** found its way into the kitchen that we had to fumigate.

2. BAN/BANDAGE

   a. Mike explained that the **ban** dates from the late 1930s.
   b. Mike explained that the **bandage** was very tight in order to stop the bleeding.
   c. Mike explained that the **arch** had been built by the Romans.
   d. Mike explained that the **cabbage** always tasted horrible.
   e. Mike explained that the **ban** solved the drinking problem.

3. BILL/BUILDING

   a. It was agreed that the **bill** doesn't have to be paid immediately.
   b. It was agreed that the **building** doesn't have to be pulled down immediately.
   c. It was agreed that the **name** of the ship would be the Titanic.
   d. It was agreed that the **program** me was hardly worth watching.
   e. It was agreed that the **bill** for food should be paid immediately.

4. BOWL/BOULDER

   a. We were lucky that the **bowl** didn't break when it hit the floor.
   b. We were lucky that the **boulder** didn't crush us to death when it rolled down the hillside.
   c. We were lucky that the **rope** didn't break with our combined weight.
   d. We were lucky that the **hammer** was kept in the toolbox.
   e. We were lucky that the **bowl** matched the one that we'd broken earlier.

5. BRAN/BRANDY

a. Susan claimed that the **bran** didn't taste nearly so bad.

b. Susan claimed that the **brandy** tasted much nicer.

c. Susan claimed that the **chrome** would never tarnish.

d. Susan claimed that the **cupboard** was much cheaper in the sale.

e. Susan claimed that the **bran** tasted much nicer.

6. CAN/CANTEEN

a. Opening the **can** takes a long time with a rusty penknife.

b. Opening the **canteen** was the cooks first job in the morning.

c. Opening the **barn** let the sheep out into the field.

d. Opening the **hostel** on a Sunday was a good idea.

e. Opening the **can** shouldn't take long with the right tool.

7. CAP/CAPTAIN

a. The soldier saluted the flag with his **cap** tucked under his arm.

b. The solider saluted the flag with his **captain** looking on.

c. The soldier saluted the flag with his **palm** facing forwards.

d. The soldier saluted the flag with his **rifle** by his side.

e. The soldier saluted the flag with his **cap** looking slightly crumpled

8. CHAP/CHAPLAIN

a. During the speech, the **chap** laughed at all the jokes.

b. During the speech, the **chaplain** started snoring really loudly.

c. During the speech, the **hum** died down.

d. During the speech, the **platform** started creaking alarmingly.

e. During the speech, the **chap** shut his eyes and went to sleep.

9. CREW/CRUSADE

a. It was unfortunate that the **crew** celebrated their victory so loudly.

b. It was unfortunate that the **crusade** was so violent.

c. It was unfortunate that the **fog** was so thick.

d. It was unfortunate that the **garage** was closed at weekends.

e. It was unfortunate that the **crew** veered into the bank at the start of the race.

10. CROW/CROQUET

a. After the lawn was mowed the **crow** could continue looking for food.

b. After the lawn was mowed the **croquet** match could begin.

c. After the lawn was mowed the **weeds** could be seen more clearly than ever.

d. After the lawn was mowed the **picnic** could take place.

e. After the lawn was mowed the **crow** gave up looking for worms.

11. CRY/CRISIS

a. Everyone was worried as the **cry** seemed to come from the attic.

b. Everyone was worried as the **crisis** was getting worse by the minute.

c. Everyone was worried as the **exam** was much harder than expected.

d. Everyone was worried as the **engine** had started making loud noises.

e. Everyone was worried as the **cry** didn't sound like it came from the TV.

12. DEN/DENTIST

a. At the end of a hard day, the **den** tends to be the place I choose to relax.

b. At the end of a hard day, the **dentist** needed somewhere to relax.

c. At the end of a hard day, the **chores** are the last thing I want to do.

d. At the end of a hard day, the **washing** up is the last thing I want to do.

e. At the end of a hard day, the **den** should be an ideal place to relax.

13. DOCK/DOCTOR

a. On Saturdays the **dock** teemed with people.

b. On Saturdays the **doctor** was always very busy.

c. On Saturdays the **ducks** are usually very well fed.

d. On Saturdays the **circus** is fully booked.

e. On Saturdays the **dock** should be fairly quiet.

14. DOLL/DOLPHIN

a. The children thought the **doll** felt softer than usual.

b. The children thought the **dolphin** was beautiful.

c. The children thought the **clown** was very funny.

d. The children thought the **museum** was very boring.

e. The children thought the **doll** could be fun to play with.

15. FAN/FANCY

a. Everyone agreed that the **fan** suited Catherine's new outfit.

b. Everyone agreed that the **fancy** clothes suited Catherine.

c. Everyone agreed that the **bait** should be suitable for catching rats.

d. Everyone agreed that the **rations** were inadequate for adults.

e. Everyone agreed that the **fan** should be left on during the afternoon.

16. GIN/GINGER

a. A splash of **gin** just about makes the drink perfect.

b. A splash of **ginger** makes whiskey taste really good.

c. A splash of **soup** ruined my outfit.

d. A splash of **curry** ruined my outfit.

e. A splash of **gin** tastes really good with ice and lemon.

17. GREY/GRAVY

a. Some time later, the **grey** van was all that people talked about.

b. Some time later, the **gravy** was all that people talked about.

c. Some time later, the **feast** began to get livelier.

d. Some time later, the **dagger** was found.

e. Some time later, the **grey** car was all that people talked about.

18. HAM/HAMSTER

a. During the summer it is best if the **ham** stays in the fridge.

b. During the summer it is best if the **hamster** stays in the shade.

c. During the summer it is best if the **shrubs** are watered regularly.

d. During the summer it is best if the **moped** is kept in the garage.

e. During the summer it is best if the **ham** never gets left out of the fridge.

19. HELL/HELMET

a. The soldiers thought that **hell** might be more comfortable than their barracks.

b. The soldiers thought that **helmets** would save their lives.

c. The soldiers thought that **tents** wouldn't stay dry if it rained.

d. The soldiers thought that **aeroplanes** were the best way to travel.

e. The soldiers thought that **hell** tormented the souls of their enemies.

20. JUNK/JUNCTION

a. It was obvious that the **junk** should be moved somewhere else.

b. It was obvious that the **junction** was dangerous to drive around.

c. It was obvious that the **gems** weren't worth very much money.

d. It was obvious that the **cider** was much stronger than usual.

e. It was obvious that the **junk** made the house look less tidy.

21. KID/KIDNEY

a. We were concerned when the **kid** knocked over the priceless vase.

b. We were concerned when the **kidney** infection hadn't got any better.

c. We were concerned when the **flight** was delayed by a couple of hours.

d. We were concerned when the **bouquet** of flowers didn't arrive.

e. We were concerned when the **kid** laughed at violent movies.

22. LAWN/LAUNDRY

a. On sunny days, the **lawn** dried out leaving large brown patches.

b. On sunny days, the **laundry** was hung out in the garden to dry.

c. On sunny days, the **bay** was crowded with holidaymakers.

d. On sunny days, the **canyon** was filled with haze.

e. On sunny days, the **lawn** tends to be covered with people sunbathing.

23. NAP/NAPKIN

a. Taking a **nap** can help you to stay up later

b. Taking a **napkin** from the restaurant was a good idea.

c. Taking a **dip** in the sea is very nice during the summer.

d. Taking a **hostage** allowed the robbers to make their escape.

e. Taking a **nap** tends to help me stay up later.

24. PAIN/PAINTING

a. John replied that the **pain** tempted him to abort the climb.

b. John replied that the **painting** was very colourful.

c. John replied that the **songs** were quite good.

d. John replied that the **record** was quite good.

e. John replied that the **pain** wouldn't stop him climbing.

25. PAN/PANTRY

a. Although he was an experienced cook, the **pan** transformed Bruce's cooking.

b. Although he was an experienced cook, the **pantry** contained ingredients Bruce had never seen before.

c. Although he was an experienced cook, the **sauce** was a real challenge to get right.

d. Although he was an experienced cook, the **onions** still made him cry when he chopped them.

e. Although he was an experienced cook, the **pan** saved Bruce a lot of trouble.

26. PEN/PENSION

a. We all noticed that the **pen** shook when the young man signed the form.

b. We all noticed that the **pension** payments were worth less and less each month,

c. We all noticed that the **skirt** didn't match Annes blouse.

d. We all noticed that the **trousers** didn't match Peters jacket.

e. We all noticed that the **pen** changed Phillip's handwriting for the better.

27. PIG/PIGMENT

a. Because of its odd appearance, the **pig** made everyone gasp with astonishment.

b. Because of its odd appearance, the **pigment** was rejected by Dulux.

c. Because of its odd appearance, the **tie** attracted attention.

d. Because of its odd appearance, the **bicycle** was never stolen.

e. Because of its odd appearance, the **pig** never got sold at market.

28. PILL/PILGRIM

a. They hoped that the **pill** granted them immunity from the disease.

b. They hoped that the **pilgrim** would save them.

c. They hoped that the **hint** would be understood.

d. They hoped that the **basement** would not get flooded by the storm.

e. They hoped that the **pill** didn't have any unpleasant side effects.

29. POLE/POULTRY

a. As we climbed over the farm gate, the **pole** tripped us up.

b. As we climbed over the farm gate, the **poultry** ran away from us.

c. As we climbed over the farm gate, the **heel** on my shoe came loose.

d. As we climbed over the farm gate, the **orchard** could be seen.

e. As we climbed over the farm gate, the **pole** didn't support our weight.

30. SHELL/SHELTER

a. Although badly battered, the **shell** tempted the collector.

b. Although badly battered, the **shelter** was warm and dry.

c. Although badly battered, the **yacht** was still watertight.

d. Although badly battered, the **vessel** was still watertight.

e. Although badly battered, the **shell** might still be valuable.

31. SPY/SPIDER

a. We had to be careful that the **spy** didn't overhear our conversations.

b. We had to be careful that the **spider** didn't crawl into our sleeping bags.

c. We had to be careful that the **jeans** were washed inside out.

d. We had to be careful that the **ferry** was on time.

e. We had to be careful that the **spy** listened to the fake recording.

32. STAY/STATION

a. They thought that the **stay** became boring after a while.
b. They thought that the **stable** would cost more than the house to heat.
c. They thought that the **kiln** was hot enough to fire the pots.
d. They thought that the **pistol** belonged to the criminal.
e. They thought that the **stay** ceased being interesting after the first week.

33. TRACK/TRACTOR

a. When it reached the house, the **track** turned north towards the forest.
b. When it reached the house, the **tractor** came to a halt.
c. When it reached the house, the **cat** was offered a saucer of milk.
d. When it reached the house, the **parcel** remained unopened for several days.
e. When it reached the house, the **track** got more difficult to follow.

34. TRAY/TRAITOR

a. After a while, the **tray** tempted him too much and he started to eat.
b. After a while, the **traitor** became careless and he was caught.
c. After a while, the **flag** was raised to the top of the flagpole.
d. After a while, the **kettle** came to the boil.
e. After a while, the **tray** should have been returned to the kitchen.

35. TREE/TREATY

a. For the last fifty years there has been a **tree** towering above this house.
b. For the last fifty years there has been a **treaty** between England and Germany.
c. For the last fifty years there has been a **race** to see who could climb the hill fastest.
d. For the last fifty years there has been a **butchers** in the high street.
e. For the last fifty years there has been a **tree** standing on this spot.

36. TRY/TRIFLE

a. We were disappointed that the **try** failed to win the match.
b. We were disappointed that the **trifle** hadn't been touched.
c. We were disappointed that the **queen** didn't come to visit the school.
d. We were disappointed that the **princess** didn't come to visit the school.
e. We were disappointed that the **try** very nearly lost us the match.

37. WALL/WALNUT

a. A severe storm left the **wall** nearest the house badly damaged.

b. A severe storm left the **walnut** tree badly damaged.

c. A severe storm left the **town** with a large bill for the clear-up operation.

d. A severe storm left the **locals** with a large bill for the clear-up operation.

e. A severe storm left the **wall** teetering on the brink of collapse.

38. WELL/WELCOME

a. In the village, the **well** can't cope with this summers drought.

b. In the village, the **welcome** given to tourists is very friendly.

c. In the village, the **fumes** from the factory are unbearable.

d. In the village, the **parson** is very friendly.

e. In the village, the **well** might not cope with this summers drought.

39. WIN/WINTER

a. After a bad start to the season, the **win** turned the teams fortunes around.

b. After a bad start to the season, the **winter** became much milder than usual.

c. After a bad start to the season, the **drought** was eased by the arrival of the monsoon.

d. After a bad start to the season, the **public** stopped attending the matches.

e. After a bad start to the season, the **win** helped our team to avoid relegation.

40. WIT/WITNESS

a. Everyone thought Tom's **wit** nearly deserved a prize.

b. Everyone thought Tom's **witness** was the least convincing.

c. Everyone thought Tom's **socks** were a horrible colour.

d. Everyone thought Tom's **jacket**  made him look very smart.

e. Everyone thought Tom's **wit** made him an ideal companion for the trip.

# Appendix B – Supplementary analyses

## B.1. Experiment 2b

Two three-way analyses of variance on participants and items were carried out on response times with the factors of prime type (short test word, long test word, control word) and target length (short word, long word) as well as the between groups factor of version or item group. There was a main effect of target length ($F_1[1,43]= 13.98$, $p<.001$; $F_2[1,33]= 5.68$, $p<.05$) with faster responses to shorter visual targets. Participants also responded faster to word targets following related primes as reflected in a significant main effect of prime type both ($F_1[2,86]= 11.64$, $p<.001$; $F_2[2,66]= 13.86$, $p<.001$). There was also a significant interaction between these two factors ($F_1[2,86]= 5.91$, $p<.005$; $F_2[2,66]= 5.38$, $p<.01$) suggesting that the magnitude of the priming effect differed depending on the identity of both the prime and target word.

Analyses on arcsine transformed error rates showed a similar pattern of results to those found in the response time data. There was a main effect of prime type ($F_1[2,86]= 4.45$, $p<.05$; $F_2[2,66]= 5.04$, $p<.1$) as well as a marginal effect of target length ($F_1[1,43]= 3.54$, $p<.1$; $F_2[1,33]= 3.37$, $p<.1$). There was also a significant interaction between prime type and target length ($F_1[2,86]= 3.62$, $p<.05$, $F_2[2,66]= 4.06$, $p<.05$).

## B.2. Experiment 2c

Analysis of variance showed significant main effects of both target type ($F_1[1,51]= 17.79$, $p<.001$; $F_2[1,33]= 5.66$, $p<.05$) and prime type ($F_1[2,102]= 13.96$, $p<.001$; $F_2[2,66]= 10.55$, $p<.001$) as well as a significant interaction between these factors ($F_1[2,102]= 11.70$, $p<.001$; $F_2[2,66]= 10.14$, $p<.001$). A similar pattern is found in the analysis of arcsine transformed error rates with a main effect of target type ($F_1[1,51]= 7.49$, $p<.01$; $F_2[1,33]= 4.61$, $p<.05$) and of prime type ($F_1[2,102]= 3.98$,

$p<.05$; $F_2[2,66]= 5.50$, $p<.01$). The interaction between prime and target type in error rates was only marginally significant by participants and items ($F_1[2,102]= 2.92$, $p<.1$; $F_2[2,66]= 3.13$, $p<.1$).

## B.3. Experiment 2d

Response times were analysed with two three-way ANOVAs using the factors prime type, target type and an additional factor of version or item group. This showed significant main effects of prime type ($F_1[2,82]= 7.01$, $p<.01$; $F_2[2,66]= 5.43$, $p<.01$) and of target type ($F_1[1,41]= 26.41$, $p<.001$; $F_2[1,33]= 11.58$, $p<.01$) as well as an interaction between these factors ($F_1[2,82]= 13.68$, $p<.001$; $F_2[2,66]= 15.47$, $p<.001$). This interaction between prime type and target type was also shown in anova on the arcsine transformed error rates in each condition ($F_1[2,82]= 3.47$, $p<.05$; $F_2[2,66]= 3.29$, $p<.05$) though neither of the main effects of prime type ($F_1<1$; $F_2<1$) and target type ($F_1[1,41]= 1.33$, $p>.1$; $F_2[1,33]= 1.17$, $p>.1$) were significant in this analysis.

## B.4. Experiment 4a

Analysis of variance carried out on the response time data showed the expected effect of target type ($F_1[1,42]= 111.66$, $p<.001$; $F_2[1, 33]= 37.28$, $p<.001$) and of prime type ($F_1[2,84]= 4.13$, $p<.05$; $F_2[2,66]= 4.45$, $p<.05$) indicating that responses were faster to short word targets and to targets preceded by related test primes. The interaction between prime and target type was significant by participants and not by items ($F_1[2,84]= 3.91$, $p<.05$; $F_2[2,66]= 1.35$, $p>.1$). Analyses of arc-sine transformed error proportions showed an effect of target type ($F_1[1,42]= 14.97$, $p<.001$; $F_2[1,33]= 9.04$, $p<.01$) indicating that error rates were also lower for shorter targets. However, there were no effects of prime type on error rate, either as a main effect ($F_1[2,84]= 2.15$, $p>.1$; $F_2[2,66]<1$) or by interaction with target type ($F_1[2,84]= 1.11$, $p>.1$, $F_2[2,66]<1$).

## B.5. Experiment 4b

Analysis of variance on response time data showed a highly significant main effect of target type. This indicates that lexical decision responses were again faster to short words than to long words ($F_1$[1,41]= 102.29, p<.001; $F_2$[1,33]= 25.91, p<.001). Unlike in previous experiments there was no main effect of prime type in these analyses ($F_1$[2,82]<1, $F_2$[2,66]<1) although the interaction between prime and target type was significant ($F_1$[2,82]= 3.56, p<.05; $F_2$[2,66]= 3.96, p<.05). Analysis of error rates showed a marginally significant effect of target type by participants but not by items ($F_1$[1,41]= 3.78, p<.1; $F_2$[1,33]= 2.27, p>.1) and a marginally significant effect of prime type, again by participants but not by items ($F_1$[2,82]= 2.60, p<.1; $F_1$[2,66]= 2.04, p>.1). There was no significant interaction between these factors, in analyses with either participants or items as the random factor ($F_1$[2,82]= 1.28, p>.1; $F_2$[2,66]= 1.82, p>.1).

# References

Abu-Bakar, M., & Chater, N. (1995). Time-warping tasks and recurrent neural networks. In J. P. Levy, D. Bairaktaris, J. A. Bullinaria, & P. Cairns (Eds), *Connectionist models of memory and language* (pp. 289-310). London: UCL Press.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419-439.

Anderson, S., & Port, R. F. (1994). Evidence for syllable structure, stress and juncture from segmental durations. *Journal of Phonetics, 22*, 283-315.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effects of subphonetic differences on lexical access. *Cognition, 52*, 163-187.

Aslin, R. N., Woodward, J. Z., La Mendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. L. Morgan & K. Demuth (Eds), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117-134). Mahwah, NJ: Erlbaum.

Baayen, R. H., Pipenbrook, R., & Guilikers, L. (1995). The Celex Lexical Database Version 2.5 (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Bard, E. G., Shillcock, R. C., & Altmann, G. T. M. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception and Psychophysics, 44*, 395-408.

Barry, W. J. (1981). Internal juncture and speech communication. In W. J. Barry & K. J. Kohler (Eds), *Beitrage zur experimentalen und angewandten phonetik*, Kiehl, Germany: AIPUK.

Bertinetto, P. M. (1981). *Strutture prosodiche*. Firenze, Italy: Accademia della crusca.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Bloom, P., & Markson, L. (1998). Capacities underlying word learning. *Trends in Cognitive Sciences, 2*, 67-73.

Bradley, D. C., & Forster, K. I. (1987). A reader's view of listening. *Cognition, 25*, 103-134.

Brent, M. R. (1997). Towards a unified model of lexical acquisition and lexical access. *Journal of Psycholinguistic Research*, *26*, 363-375.

Brent, M. R. (1999a). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences, 3*, 294-301.

Brent, M. R. (1999b). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning, 34*, 71-105.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61,* 93-125.

Briscoe, E. J. (1989). Lexical access in connected speech recognition, *Proceedings of the 27th Congress, Association for Computational Linguistics*. (pp. 84-90). Vancouver.

Bullinaria, J. A., & Chater, N. (1995). Connectionist modelling: Implications for cognitive neuropsychology. *Language and Cognitive Processes, 10*, 227-264.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1995). Bottom-up connectionist modelling of speech. In J. P. Levy, D. Bairaktaris, J. A. Bullinaria, & P. Cairns (Eds), *Connectionist models of memory and language* (pp. 289-310). London: UCL Press.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus based approach to speech segmentation. *Cognitive Psychology*, *33*, 111-153.

Carey, S., & Bartlett, E. (1978) Acquiring a single new word. *Papers and Reports on Child Language Development, 15*, 17-29.

Chomsky, N., & Halle, M. (1968). *The sound patterns of English*. New York: Harper & Row.

Christiansen, M. H., & Allen, J. (1997). Coping with variation in speech segmentation. In A. Sorace, C. Heycock & R. Shillcock (Eds), *Proceedings of the 1997 GALA conference: Language acquisition: Knowledge representation and processing.* University of Edinburgh.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13,* 221-268.

Christie, W. M., Jr. (1974). Some cues for syllable juncture perception in English. *Journal of the Acoustical Society of America, 55*, 819-821.

Christophe, A., Dupoux, E., Bertoncini, J. & Mehler, J. (1994). Do infants perceive word boundaries: An empirical study of the bootstrapping of lexical boundaries. *Journal of the Acoustic Society of America, 73*, 1570-1580.

Christophe, A., Guasti, T., Nespor, M., Dupoux, E., & Ooyen, B. V. (1997). Reflections on prosodic bootstrapping: Its role for lexical and syntactic acquisition. *Language and Cognitive Processes, 12,* 585-612.

Clark, A. (1993). *Associative engines: Connectionism, context and representational change*. Cambridge, MA: MIT Press.

Clark, A., & Thornton, C. (1997). Trading spaces: Computation, representation and the limits of uninformed learning. *Behavioural and Brain Sciences, 20,* 57-90.

Cole, R. A., & Jakimik, J. (1980). A model of speech perception. In R. A. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, NJ: Erlbaum.

Content, A., & Sternon, P. (1994). Modelling retroactive context effects in spoken word recognition with a simple recurrent network. In A. Ram & K. Eiselt (Eds), *Proceedings of the 16th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

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

Crowder, R. G., & Morton, J. (1969). Pre-categorical acoustic storage (PAS). *Perception and Psychophysics, 5,* 365-373.

Crystal, T. H., & House, A. S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America, 88*, 101-112.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language, 31*, 218-236.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language, 2*, 133-142.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 113-121.

Dahan, D., & Brent, M. R. (1999). On the discovery of novel word-like units from utterances: An artificial language study with implications for native-language acquisition. *Journal of Experimental Psychology: General, 128*, 165-185.

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 914-927.

Dupoux, E., & Hammond, M. (submitted). Are stress units used in pre-lexical processing in English? Submitted to *Journal of Experimental Psychology: Learning Memory and Cognition*.

Echols, C. H., Crowhurst, M. J., & Childers, J. B. (1997). The perception of rhythmic units in speech by children and adults. *Journal of Memory and Language, 36*, 202-225.

Elman, J. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211.

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition, 48*, 71-99.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for co-articulation of lexically restored phonemes. *Journal of Memory and Language, 27*, 143-165.

Elman, J., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness*. Cambridge, MA: MIT Press.

Fear, B. D., Cutler, A., & Butterfield, S. (1995). The strong-weak syllable distinction in English. *Journal of the Acoustical Society of America, 97,* 1893-1904.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3-73.

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

Forster, K. I. (1989). Basic issues in lexical processing. In W. D. Marslen-Wilson (Ed.), *Lexical representation and process*. Cambridge, MA: MIT Press.

Foulke, E., & Sticht, T. (1969). Review of research on the intelligibility and comprehension of accelerated speech. *Psychological Bulletin, 72*, 50-62.

Frauenfelder, U. H., & Peeters, G. (1990). Lexical segmentation in TRACE: An exercise in simulation. In G. T. M. Altmann (Ed.), *Cognitive Models of Speech Processing*. Cambridge, MA: MIT Press.

Frauenfelder, U. H., Segui, J., & Dijkstra, T. (1990). Lexical effects in phonemic processing: Facilitory or inhibitory. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 77-91.

Frazier, L., & Clifton, C. (1996). *Construal*. Cambridge, MA: MIT Press.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*, 110-125.

Gaskell, M. G. (1994). *Spoken Word Recognition: A Combined Computational and Experimental Approach.* Unpublished PhD thesis, Birkbeck College, University of London.

Gaskell, M. G., & Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 144-158.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12,* 613-656.

Gaskell, M. G., & Marslen-Wilson, W. D. (in press). Ambiguity, competition and blending in spoken word recognition. *Cognitive Science*.

Gaskell, M. G., & Marslen-Wilson, W. D. (submitted). Discriminating local and distributed models of competition in spoken word recognition. Manuscript submitted to *Journal of Memory and Language.*

Gaskell, M. G., Hare, M., & Marslen-Wilson, W. D. (1995). A connectionist model of phonological representation in speech perception. *Cognitive Science, 19*, 407-439.

Gasser, M. (1992). Learning distributed representations for syllables. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Gasser, M., Eck, D., & Port, R. (1999). Meter as mechanism: A neural network that learns metrical patterns. *Connection Science, 11*, 187-205.

Gleitman, L. R. (1994) Words, words, words. *Philosophical Transactions of the Royal Society of London. Series B - Biological Sciences. 346*(1315), 71-77.

Goldinger, S. D. (1996a). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22,* 1166-1183.

Goldinger, S. D. (1996b). Auditory lexical decision. *Language and Cognitive Processes, 11*, 559-567.

Goldowsky, B. N., & Newport, E. L. (1993). Modeling the effects of processing limitations on the acquisition of morphology: The less is more hypothesis. In E. V. Clark (Ed.), *Proceedings of the 24th Annual Child Language Forum.* Stanford, CA: CSLI.

Gonnerman, L. G., Devlin, J. T., Anderson, E., Seidenberg, M. S. (submitted) Derivational morphology as an emergent inter-level representation. Manuscript submitted to *Journal of Memory and Language*.

Gow, D. W., Melvold, J., & Manuel, S. (1996). How word onsets drive lexical access and segmentation: Evidence from acoustics, phonology and processing. In

*Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, PA.

Gow, D., J., Jr., & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 344-359.

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

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

Grosjean, F. (1996). Gating. *Language and Cognitive Processes, 11*, 597-604.

Grosjean, F., & Gee, J. P. (1987). Prosodic structure and spoken word recognition. *Cognition, 25*, 135-156.

Gupta, P., & Mozer, M. C. (1993). Exploring the nature and development of phonological representations. in Kintsch, W. (Ed.) *Proceedings of the 15th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading and dyslexia: Insights from connectionist models. *Psychological Review, 106*, 491-528.

Harris, C. L. (1994). Coarse coding and the lexicon. In C. Fuchs & B. Victorri (Eds), *Continuity in linguistic semantics*. Amsterdam: Benjamins.

Harris, C. L. (1996). *Recognition of common word combinations: Towards a lexicon of variable sized units*. Research report, Psychology Department, Boston University, MA.

Harris, Z. S. (1955). From phoneme to morpheme. *Language, 31*, 190-222.

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence, 40*, 185-234.

Huggins, A. W. F. (1975). Temporally segmented speech and "echoic" storage. In A. Cohen & S. G. Nooteboom (Eds), *Structure and process in speech perception* (pp. 209-225). New York: Springer-Verlag.

Jakimik, J., Cole, R. A., Rudnicky, A. I. (1985). Sound and spelling in spoken word recognition. *Journal of Memory and Language, 24*(2), 165-178.

Jakobson, R., Fant, G., & Halle, M. (1952). *Preliminaries to speech analysis*. Cambridge, MA: MIT Press.

Joanisse, M. F., & Seidenberg, M. S. (1999). Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences of the United States of America, 96*(13), 7592-7597.

Jones, D. (1931). The "word" as a phonetic entity. *Le maitre phonetique, 3rd series, 36*, 60-65.

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences, 3,* 323-327.

Jusczyk, P. W., & Aslin, R. N. (1995) Infants detection of sound patterns of words in fluent speech. *Cognitive Psychology, 29*, 1-23.

Jusczyk, P. W., Cutler, A., & Redanz, N. (1993). Preference for the predominant stress pattern in English. *Child Development, 5*, 265-286.

Kessinger, R. H., & Blumenstein, S. E. (1998). Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies. *Journal of Phonetics, 26*, 117-128.

Klatt, D. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, *59*, 1208-1221.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. In R. A. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, NJ: Erlbaum.

Kohsom, C., & Gobet, F. (1997). Adding spaces to Thai and English: Effects on reading. in M. G. Shafto & P. Langley (Eds). *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, Stanford, CA.

Lehiste, I. (1960). An acoustic-phonetic study of internal open juncture. *Phonetica, 5*(supplement), 5-54.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America, 51*, 2018-2024.

Liberman, M. Y., & Streeter, L. A. (1978). Use of non-sense syllable mimicry in the study of prosodic phenomena. *Journal of the Acoustical Society of America, 63*, 231-233.

Luce, P. A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception and Psychophysics, 39*, 155-158.

Luce, P. A., & Cluff, M. S. (1998). Delayed commitment in spoken word recognition. *Perception and Psychophysics, 60*, 484-490.

Luce, P. A., & Lyons, E. A. (1999). Processing lexically embedded spoken words. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 174-183.

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1990). Similarity neighbourhoods of spoken words. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing*. Cambridge, MA: MIT Press.

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

Manaster-Ramer, A. (1996). A letter from an incompletely neutral phonologist. *Journal of Phonetics, 24,* 477-489.

Marslen-Wilson, W. (1984). Function and processing in spoken word recognition: A tutorial review. In H. Bouma & D. G. Bouwhuis (Eds), *Attention and Performance X: Control of Language Processing*. Hillsdale NJ: Erlbaum.

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

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. *Cognition, 25*, 71-102.

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

Marslen-Wilson, W. D. (1999). Abstractness and combination: The morphemic lexicon. In S. Garrod & M. Pickering (Eds), *Lexical Processing*. London: UCL Press.

Marslen-Wilson, W. D., Ford, M., Older, L., & Zhou, X. (1996). The combinatorial lexicon: Affixes as processing structures. In G. W. Cottrell (Ed). *Proceedings of the 18th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Marslen-Wilson, W. D., & Gaskell, G. (1992). Match and mismatch in lexical access. [Abstract]. *International Journal of Psychology*, 27, 61.

Marslen-Wilson, W. D., Moss, H. E., & van Halen, S. (1996). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 1376-1392.

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

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

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

Marslen-Wilson, W. D., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes and features. *Psychological Review, 101*, 653-675.

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

Maskara, A., & Noetzel, A. (1993). Sequence recognition with recurrent neural networks. *Connection Science, 5*, 139-152.

Mattys, S. L. (1997). The use of time during lexical processing and segmentation: A review. *Psychonomic Bulletin and Review, 4,* 310-329.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Word segmentation in infants: How phonotactics and prosody combine. *Cognitive Psychology, 38*, 465-494.

McAuley, J. D. (1994). Time as Phase: A dynamic model of time perception. In A. Ram & K. Eisert (Eds) *Proceedings of the 16th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

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

McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel distributed processing: Explorations in the Microstructure of Cognition. (Vol. 2: Psychological and Biological Models).* Cambridge, MA: MIT Press.

McQueen, J. M. (1996). Word spotting. *Language and Cognitive Processes, 11*, 695-699.

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

McQueen, J. M., Cutler, A., Briscoe, T., & Norris, D. (1995). Models of continuous speech recognition and the contents of the vocabulary. *Language and Cognitive Processes, 10,* 309-331.

McQueen, J. M., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20,* 621-638.

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General, 126,* 99-130.

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

Mertus, J. (1989). *BLISS users manual*. Providence: Brown University.

Miller, J. L., & Lieberman, A. M. (1979). Some effects of late-occurring information on the perception of stop consonant and semi-vowel. *Perception and Psychophysics, 25,* 457-465.

Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and perception for the voicing contrast. *Phonetica, 43,* 106-115.

Monsell, S., & Hirsh, K. W. (1998). Competitor priming in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 24*, 1495-1520.

Morgan, J. L. (1996). A rhythmic bias in preverbal speech segmentation. *Journal of Memory and Language, 35*, 666-688.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review*, *76*, 165-178.

Moss, H. E., & Marslen-Wilson, W. (1993). Access to word meanings during spoken language comprehension: Effects of sentential semantic context. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*, 1254-1276.

Moss, H. E., Hare, M. L., Day, P., & Tyler, L. K. (1994). A distributed memory model of the associated boost in semantic priming. *Connection Science, 6*, 413-426.

Moss, H. E., Ostrin, R. K., Tyler, L. K., & Marslen-Wilson, W. D. (1995). Accessing different types of semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21*, 863-883.

Moss, H., & Older, L. (1996). *Birkbeck word association norms*. Hove: Psychology Press.

Nakatani, L. H., & Dukes, K. D. (1977). Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America, 62*, 715-719.

Nakatani, L. H., & Schaffer, J. A. (1978). Hearing words without words: Prosodic cues for word perception. *Journal of the Acoustical Society of America, 63*, 234-245.

Nakatani, L. H., O'Connor, K. D., & Aston, C. H. (1981). Prosodic aspects of American English speech rhythm. *Phonetica, 38*, 84-106.

Nguyen, M. H., & Cottrell, G. W. (1997). Tau-Net: A neural network for modeling temporal variability. *Neurocomputing, 15,* 249-271.

Norris, D. (1990). A dynamic-net model of human speech recognition. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing*. Cambridge, MA: MIT Press.

Norris, D. (1993). Bottom up connectionist models of 'interaction'. In G. Altmann & R. Shillcock (Eds), *Cognitive models of language processes: Proceedings of the second Sperlonga meeting*. Hove, UK: Erlbaum.

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

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21*, 1209-1228.

Norris, D., McQueen, J. M., & Cutler, A. (in press). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences.*

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

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language, 32*, 258-278.

Page, M. (in press). Connectionist modelling in Psychology: A localist manifesto. *Behavioural and Brain Sciences*.

Pallier, C., Christophe, A., & Mehler, J. (1998). Language-specific listening. *Trends in Cognitive Sciences, 1,* 129-132.

Pallier, C., Sebastian-Galles, N., Dupoux, E., Christophe, A., & Mehler, J. (1998). Perceptual adjustment to time-compressed speech: A cross-linguistic study. *Memory and Cognition, 26*, 844-851.

Perruchet, P., & Vinter, A. (1998). PARSER: A model of word segmentation. *Journal of Memory and Language, 39,* 246-263.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28*, 73-193.

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

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for co-articulation mediated by the lexicon? *Journal of Memory and Language, 39*, 347-370.

Plaut, D. C. (1995). Semantic and associative priming in a distributed attractor network. In J. D. Moore & J. F. Lehman (Eds), *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 37-42). Mahwah, NJ: Erlbaum.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. E. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*, 56-115.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition, 38*, 43-102.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition, 48*, 21-69.

Plunkett, K., & Sinha, C. (1992). Connectionism and developmental theory. *British Journal of Developmental Psychology, 10*, 209-254.

Plunkett, K., Sinha, C., Møller, M. F., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. *Connection Science, 4*(3/4), 293-312.

Port, R. F. (1990). Representation and recognition of temporal patterns. *Connection Science, 2*(1/2), 151-176.

Port, R. F. (1996). The discreteness of phonetic elements and formal linguistics: A response to A. Manaster-Ramer. *Journal of Phonetics, 24*, 491-511.

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes, 8,* 1-56.

Quartz, S. R., & Sejnowski, T. J. (1997). The neural basis of cognitive development: A constructivist manifesto. *Behavioural and Brain Sciences, 20*, 537-596.

Radeau, M., & Morais, J. (1990). The uniqueness point effect in the shadowing of spoken words. *Speech Communication, 9*, 155-164.

Robinson, A. J. (1994) An application of recurrent networks to phone probability estimation. *IEE Transactions on Neural Networks, 5,* 298-305.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior, 9*, 487-494.

Rumelhart, D. E., & McClelland, J. L. (1986a). On learning the past tenses of English verbs. In J. L. McClelland, & D. E. Rumelhart (Eds), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition (Vol. 2: Biological and Psychological Models).* Cambridge: MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986b). *Parallel distributed processing: Explorations in the microstructure of cognition. (Vol. 1: Foundations).* Cambridge, Mass: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, & J. L. McClelland (Eds), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition (Vol. 1: Foundations).* Cambridge: MA: MIT Press.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical language learning by 8 month olds. *Science, 274*(5294), 1926-1928.

Schreuder, R., & Baayen, H. (1995). Modelling morphological processing. In L.B. Feldman (Ed.) *Morphological aspects of language processing* (pp. 131-154). Hillsdale, NJ: Erlbaum.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*, 523-568.

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems, 1*, 145-168.

Selfridge, O. G. (1959). Pandemonium: A paradigm for learning, *Symposium on the mechanisation of thought processes*. London: HMSO.

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

Servan-Schreiber, D., Cleeremans, A., & McClelland, J. L. (1991). Graded state machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning, 7*, 161-193.

Shelton, J. R., & Martin, R. C. (1992). How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning, Memory and Cognition, 18*, 1191-1210.

Shillcock, R. C. (1990). Lexical hypotheses in continuous speech. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing*. Cambridge, MA: MIT Press.

Shillcock, R., Levy, J., & Chater, N. (1991). A connectionist model of auditory word recognition in continuous speech, In *Proceedings of the 13th Annual Conference of the Cognitive Science Society* (pp. 340-345). Hillsdale, NJ: Erlbaum

Simpson, G. B. (1984). Lexical ambiguity resolution and its role in models of word recognition. *Psychological Bulletin, 96*, 316-340.

Siskind, J. M. (1996). A computational study of  cross-situational techniques for learning word-to-meaning mappings. *Cognition, 61*, 39-91.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart, & J. L. McClelland (Eds), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition (Vol. 1: Foundations)*. Cambridge: MA: MIT Press.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioural and Brain Sciences, 11*, 1-74.

Sougné, J. (1998). "Connectionism and the problem of multiple instantiation." *Trends in Cognitive Sciences, **2***, 183-189.

St. John, M., F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46,* 217-257.

Stevens, K. N., & Blumstein, S. E. (1981). The search for invariant acoustic-phonetic correlates of phonetic features. In P. D. Eimas & J. L. Miller (Eds), *Perspectives on the study of speech*. Hillsdale, NJ: Erlbaum.

Swinney, D., Onifer, W., Prather, P., & Hirshkowitz, M. (1979). Semantic facilitation across modalities in the processing of individual words and sentences. *Memory and Cognition, 7*, 159-165.

Tabossi, P. (1996). Cross-modal semantic priming. *Language and Cognitive Processes, 11*, 569-576.

Tabossi, P., Burani, C., & Scott, D. (1995). Word identification in fluent speech. *Journal of Memory and Language, 34*, 440-467.

Taft, M., & Hambly, G. (1986). Exploring the cohort model of spoken word recognition. *Cognition, 22*, 259-282.

Tanenhaus, M. K., Burgess, C., & Seidenberg, M. (1988). Is multiple access on artifact of backward priming?  In Small, Cottrell, & Tanenhaus (Eds),*Lexical ambiguity resolution*. San Mateo, CA: Morgan Kaufmann.

Toothaker, L. E. (1991). *Multiple comparisons for researchers*. Newbury Park, CA: Sage.

Tyler, L. K. (1984). The structure of the word initial cohort: Evidence from gating. *Perception and Psychophysics, 36,* 417-427.

Tyler, L. K., & Frauenfelder, U. H. (1987). The process of spoken word recognition: An introduction. *Cognition, 25*, 1-20.

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

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

Ullman, M. T., Corkin, S., Coppola, M., Hickok, G., Crowden, J. H., Koroshetz, W. J., & Pinker, S. (1997). A neural dissociation within language: Evidence that the mental dictionary is part of declarative memory and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience, 9*, 266-276.

Umeda, N. (1975). Vowel duration in American English. *Journal of the Acoustical Society of America, 58*, 434-445.

Volatis, L. E., & Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the acoustic society of America, 92*, 723-735.

Vroomen, J., & de Gelder, B. (1997). Activation of embedded words in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 710-720.

Vroomen, J., van Zon, M., & de Gelder, B. (1996). Cues to speech segmentation: Evidence from juncture misperceptions and word-spotting. *Memory and Cognition, 24*, 744-755.

Walley, A., Michela, V., & Wood, D. (1995). The gating paradigm: Effects of presentation format on spoken word recognition by children and adults. *Perception and Psychophysics, 57*, 343-351.

Warren, P., & Marslen-Wilson, W. D. (1987). Continuous uptake of acoustic cues in spoken word-recognition. *Perception and Psychophysics, 41*, 262-275.

Warren, P., & Marslen-Wilson, W. D. (1988). Cues to lexical choice: Discriminating place and voice. *Perception and Psychophysics, 43,* 21-30.

Wayland, S. C., Miller, J. L. & Volaitis, L. E. (1994). The effect of sentential speaking rate on the internal structure of phonetic categories. *Journal of the Acoustical Society of America, 95,* 2694-2701.

Wolff, J. G. (1977). The discovery of segmentation in natural language. *British Journal of Psychology, 68,* 97-106.

Zhou, X., & Marslen-Wilson, W. D. (in press). Lexical representation of compound words: Cross-linguistic evidence. *Language and Cognitive Processes*.

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

Zwitserlood, P. (1996). Form priming. *Language and Cognitive Processes, 11*, 589-596.

Zwitserlood, P., & Schriefers, H. (1995). Effects of sensory information and processing time in spoken-word recognition. *Language and Cognitive Processes*, *10*, 121-136.