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.