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.

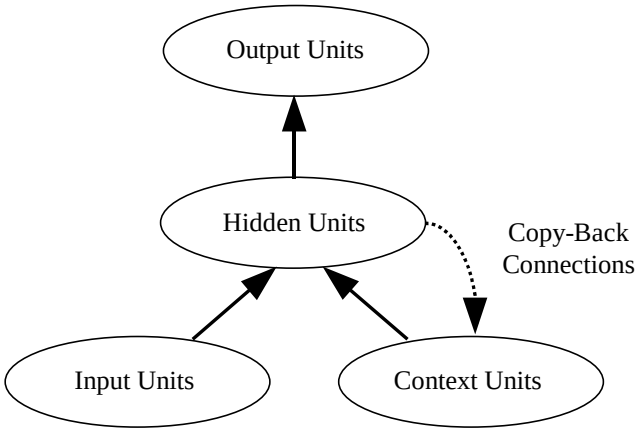

Figure 3.1: A simple recurrent network (Elman, 1990). Solid arrows represent trainable connections, broken arrows show hidden-unit activations from the previous time step, copied back to the context units on a one-to-one basis.

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	/kæptɪn/	bandit	/bændɪt/
	+ offset-embedding	topknot	/topnot/	padlock	/pædlok/
	non-embedding	landed	/lændɪd/	picnic	/pɪknɪk/
Monosyllable	onset-embedded	cap	/kæp/	ban	/bæn/
	offset-embedded	knot	/not/	lock	/lok/
	cohort competitors	dot	/dot/	dock	/dok/
		lick	/lɪk/	lid	/lɪd/
	non-embedded	tap	/tæp/	bat	/bæt/
		knit	/nɪt/	cat	/kæt/
		pot	/pot/	bid	/bɪd/

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.

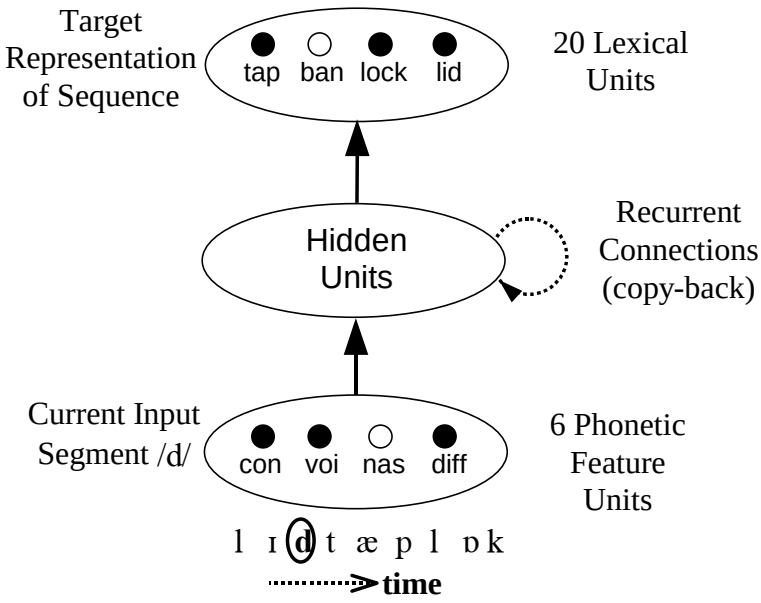

Figure 3.2: A snapshot of the SRN during training on the segment /d/ during *"lid tap lock"*. Throughout each training sequence the target activation for the network is to activate a representation of all the words in that sequence, not just the current word.

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.

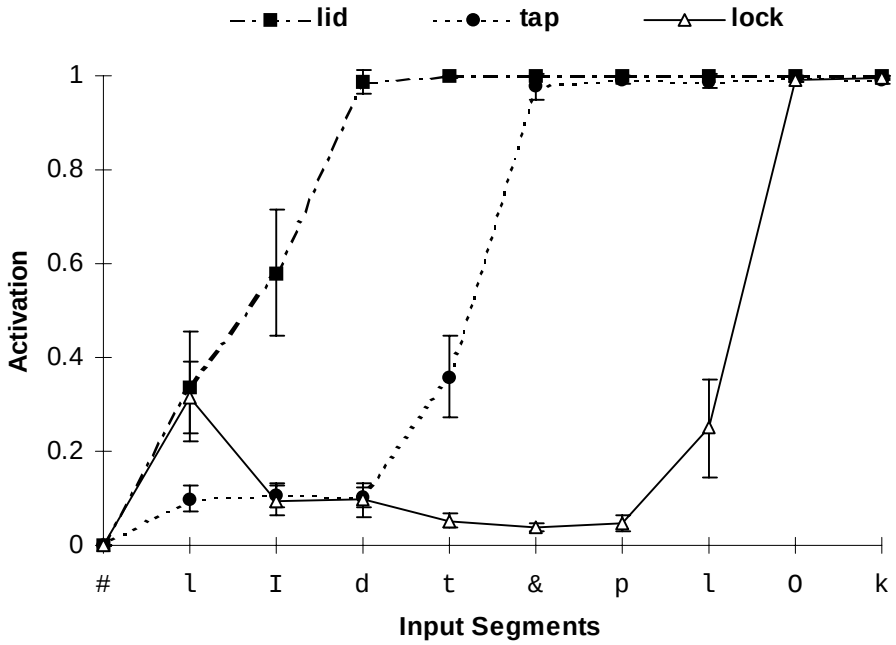

Figure 3.3: Activation of lexical units during the sequence *"lid tap lock"* averaged across ten networks in Simulation 1. Each network activates words as they are presented in the input and preserves their activation until the end of the sequence. Error bars are 1 standard deviation.

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.

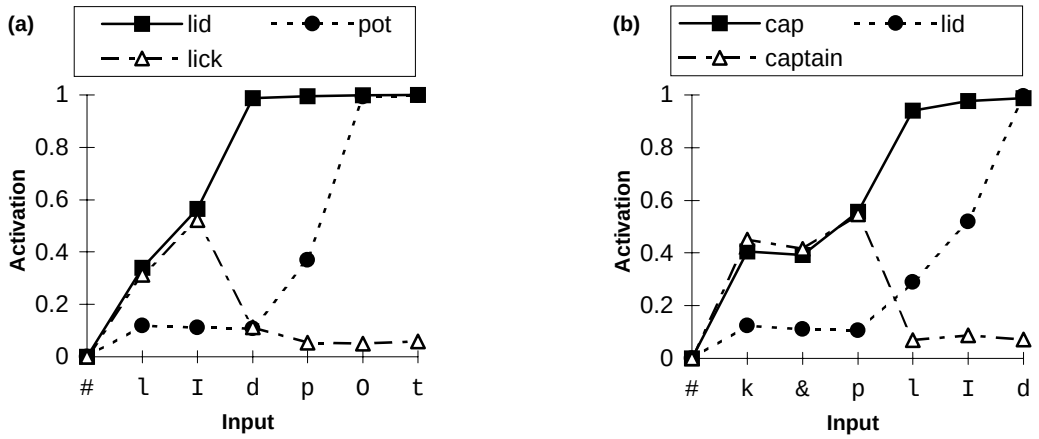

Figure 3.4: Activation of cohort competitors and onset-embedded words in Simulation 1.
(a) Cohort competitors (*lid/lick*) during the sequence *"lid pot"*
(b) Onset-Embedded words (*cap/captain*) during the sequence *"cap lid"*

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.

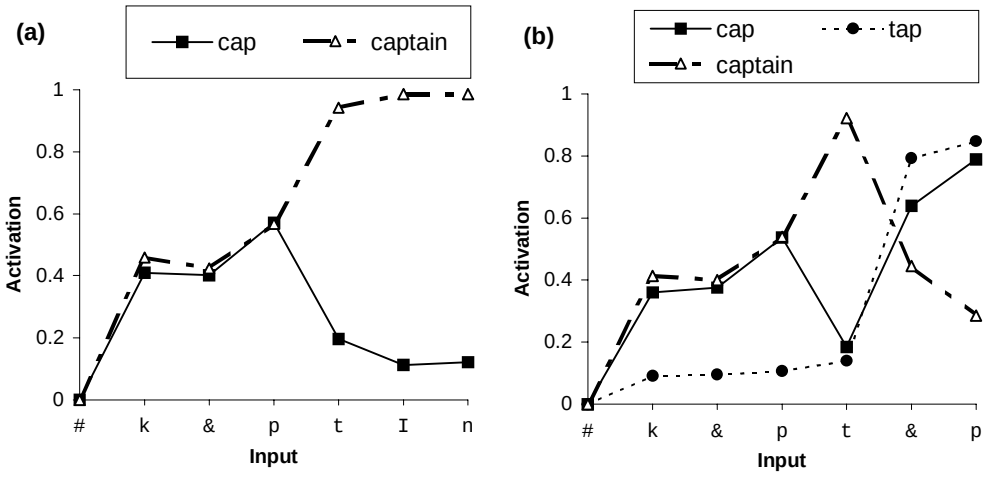

Figure 3.5: Activation of embedded words and competitors for long word sequences and lexical garden-paths in Simulation 1.
(a) Bisyllables with embeddings (*captain/cap*) during *"captain"*
(b) Lexical 'garden paths' (*cap/captain*) during *"cap tap"*

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.

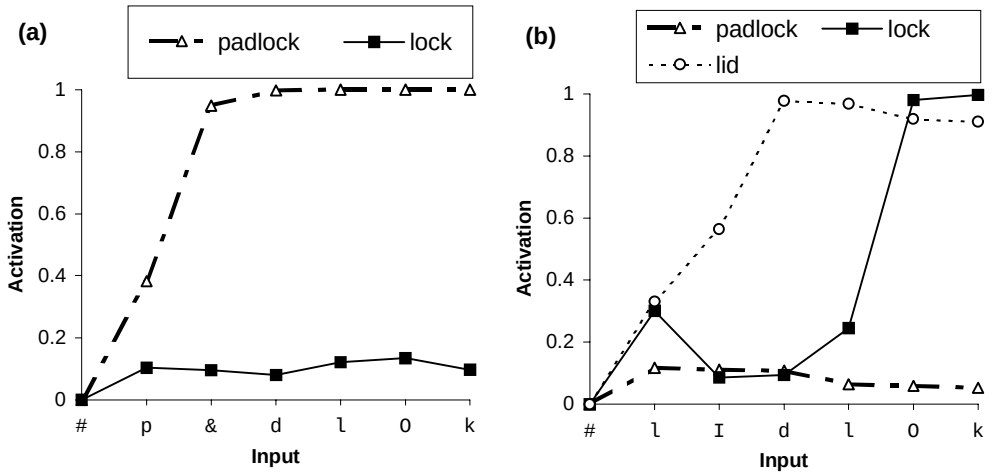

Figure 3.6: Activation of offset-embedded words in Simulation 1:
(a) Bisyllables with offset embeddings (*padlock/lock*) during *"padlock"*
(b) Offset embedded words (*lock/padlock*) during *"lid lock"*

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.

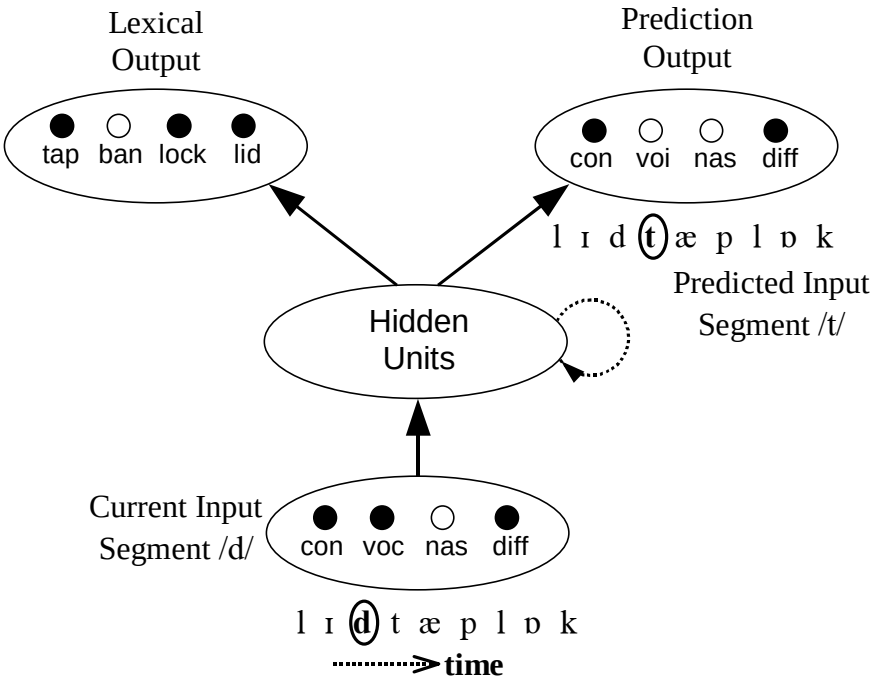

Figure 3.7: A snapshot of the SRN trained in Simulation 2. The architecture and training regime are identical to that shown in Figure 3.2, except for the additional output units trained to predict the input at the following time step.

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.

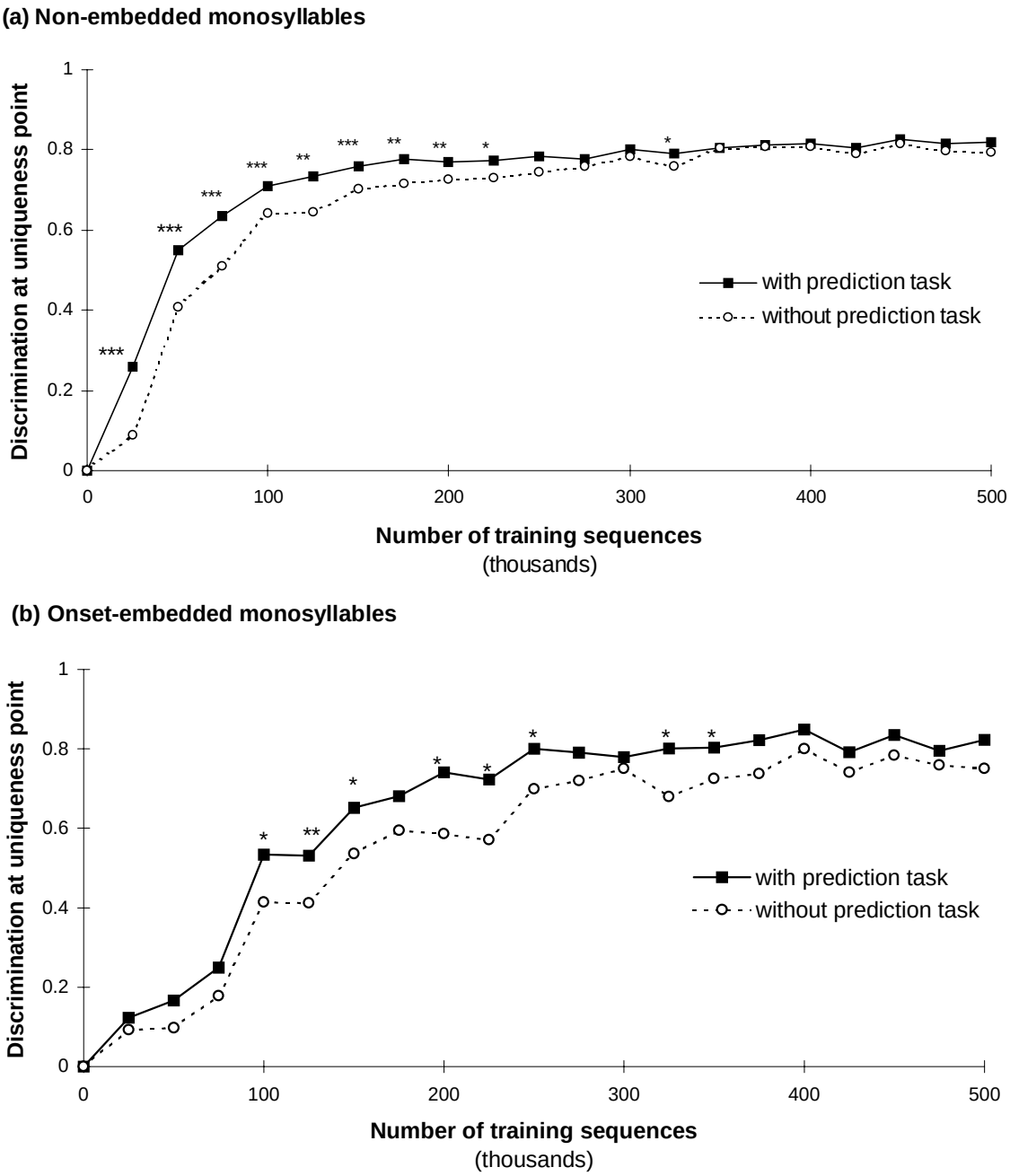

Figure 3.8: Discrimination performance for networks trained with and without the prediction task. Performance measured at uniquness point for (a) Non-embedded monosyllables (b) Onset-embedded monosyllables. Paired t-tests[8] comparing networks with and without the prediction task *** p(1 tail)<0.001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