Connectionist modelling of lexical segmentation and vocabulary acquisition

Matt H. Davis

MRC Cognition and Brain Sciences Unit,

15 Chaucer Road, Cambridge, UK

Address for correspondence:

Matt Davis, Tel: 01223 355 294 #266

MRC Cognition & Brain Sciences Unit, Fax: 01223 359 062

15 Chaucer Road Email: matt.davis@mrc-cbu.cam.ac.uk

Cambridge

CB2 2EF

Connectionist modelling of lexical segmentation and vocabulary acquisition

Words are the building blocks of language. Native speakers typically know tens of thousands of words which they combine into sentences to communicate an indefinite number of possible messages. A significant question in understanding how infants learn language is therefore to understand how they acquire words. This chapter focuses on two of the obstacles facing children learning words – firstly, how they discover which sequences of speech sounds cohere to form words (lexical segmentation) and secondly, how they learn to associate sound sequences with meanings (vocabulary acquisition). The connectionist simulations presented in this chapter provide a modelling framework for these two aspects of language acquisition. Although the simulations fall short of the scale and complexity of the learning task faced by infants, they provide an explicit account of some of the sources of information that are available to infants and how this information might be deployed in learning.

Adults typically hear sentences in their native language as a sequence of separate words. We might assume, that words in speech are physically separated in the way that they are perceived. However, when listening to an unfamiliar language we no longer experience sequences of discrete words, but rather hear a continuous stream of speech with boundaries separating individual sentences or utterances. Examination of the physical form of speech confirms the impression given by listening to foreign languages. Speech does not contain gaps or other unambiguous markers of word boundaries – there is no auditory analog of the spaces between words in printed text (Lehiste, 1960). Thus the perceptual experience of native speakers reflects language-

specific knowledge of ways in which to divide speech into words. An important set of questions, therefore, concern the sources of information that are used for segmentation and how the need to segment the speech stream affects infants learning their first words.

The continuous nature of speech might not be a problem for infants learning language if they were 'spoon-fed' with single-word utterances. However, while infant-directed speech contains exaggerations of many aspects of adult speech (e.g. distinctive intonation patterns – Fernald et al., 1989), child-directed speech does not primarily contain single word utterances. For instance in the Korman corpus (1984) less than a quarter of all utterances are single words. Furthermore, parents that are requested to teach their children new words, do not simply repeat single words to them (Aslin, Woodward, La Mendola & Bever, 1996). Even if infants were taught isolated words, this would be poor preparation for perceiving connected speech where words are produced more quickly and with greater acoustic variation than in isolation (Barry, 1981). For these reasons, infants learning their first words must first learn to segment a stream of connected speech into smaller units that communicate meaning.

Theories of how adult listeners segment the speech stream into words emphasise the role that knowledge of individual words plays in the segmentation of speech. Most current models of spoken word recognition in adults propose that segmentation arises through the identification of words in connected speech. Either by using the recognition of words to predict the location of word boundaries (Marslen-Wilson & Welsh, 1978; Cole & Jakimik, 1980) or through processes of lexical competition which ensure that only words that make up a consistent segmentation of the speech stream are activated (McClelland & Elman, 1986; Norris, 1994).

However, since words can not be learnt until the speech stream can be segmented, it seems unlikely that infants will be able to use word recognition to segment connected speech. For this reason, researchers have proposed a variety of strategies and cues that infants could use to identify word boundaries without being able to recognise the words that these boundaries delimit. This chapter, describes some computational simulations proposing ways in which these cues and strategies for the acquisition of lexical segmentation can be integrated with the infants' acquisition of the meanings of words. The simulations reported here describe simple computational mechanisms and knowledge sources that may support these different aspects of language acquisition.

Modelling language acquisition

In creating computational models of language acquisition, a variety of approaches have been taken. As with other papers in the current volume, this chapter focuses on theories which have been implemented using artificial neural networks (connectionist models). The ability of neural networks to extract structure from noisy and probabilistic input suggest that these models provide a plausible account of learning processes that are at the heart of cognitive development. Although there has been some debate on whether learning algorithms such as back-propagation can be neurally instantiated in the brain (see for example, Crick, 1989 and O'Reilly, 1996) it is clear that 'gradient-descent' learning algorithms provide more neurally-plausible accounts of learning than accounts that propose symbolic systems (see Elman et al, 1996 for further discussion and other chapters in this volume).

An important aspect of the computational modelling of psychological processes is to provide an account of the behavioural profile of language learners. Recent years

have seen rapid advances in experimental methods for investigating the abilities of pre-linguistic infants. These empirical data provide informative constraints on computational models of lexical segmentation and vocabulary acquisition and will be reviewed in the chapter. We begin by describing the pre-existing cognitive abilities that infants bring to these domains.

Pre-requisites for language acquisition

By the age of 6 months, infants have begun to acquire knowledge that is specific to their native language. Although it has been shown that new-born infants are able to discriminate between utterances that differ by a single phoneme (Eimas, Siqueland, Jusczyk & Vigorito, 1971), it is only at around 6 months of age that infants organise their phonetic categories in an adult-like manner. For instance, infants start to lose sensitivity to phonemic contrasts that are not used in their native language (Kuhl et al., 1992; Werker & Tees, 1984; see Jusczyk, 1997 for a review). The ability to detect phonemic differences while ignoring other, non-contrastive differences provides an important first step towards language acquisition. Infants will be able to focus on those aspects of the speech signal that have the potential to convey meaning in their (soon-to-be) native language.

Similarly, children's knowledge of objects in the world around them shows rapid development during the first six months of life. By this age, infants have acquired knowledge of the physical properties of objects and their interactions, as shown by their performance on tests of object permanence and their correct predictions concerning the fate of colliding and occluded objects (Baillargeon, 1994; Mareschal, 2000, this volume).

A key problem in language acquisition can then be framed by asking how infants pair the sounds of their native language with objects and actions in the outside world[1]. This problem can be divided into two distinct aspects – (1) lexical segmentation, i.e. how infants chunk speech into words, and, (2) vocabulary acquisition, i.e. how infants map those words onto objects and meanings. We will review experimental evidence and computational models relating to these aspects of the language acquisition problem.

The acquisition of lexical segmentation

In investigating the acquisition of lexical segmentations, researchers have focused on what knowledge infants have acquired about the structure of words in their native language. Two aspects of word structure that have been of primary interest are knowledge of frequent and infrequently occurring sequences of phonemes (phonotactics) and knowledge of the rhythmic alteration of stressed and unstressed syllables in words (metrical information). Both phonotactics and metrical stress have been invoked as cues that infants may use in beginning to segment the speech stream[2].

Experimental investigations

Experimenters have used a head-turn preference procedure (Fernald, 1985) to evaluate infants knowledge of higher-level structure in spoken language. This procedure allows experimenters to compare the amount of time that infants remain interested in two sets of speech stimuli, as indicated by the duration of head-turns towards either of a pair of loudspeakers that present the stimuli. This measure of infants' preferences for a given speech stimulus can be used to infer that the infants tested are sensitive to differences that exist between the two sets of stimuli.

For example, Jusczyk, Cutler and Redanz (1993a) showed that 9-month-old infants prefer listening to lists of bisyllabic words in which the two syllables of the word were stressed then unstressed (strong/weak words such as “butter” and “ardour”) rather than words that followed the reverse pattern (weak/strong words like “between” or “arouse”). This preference is of interest, since in ‘stress-timed’ languages such as English, the majority of nouns start with a strong syllable (Cutler and Carter, 1987). Thus 9-month-old infants show a consistent preference for words with the more frequent strong/weak pattern. Since this pattern was not observed in 6-month-olds, it suggests that between 6 and 9-months, infants learn something of the metrical structure of words in their native language.

A similar preference for more commonly occurring patterns has also been observed for phonotactic regularities – sequences of phonemes that are permitted or not permitted in a language. For instance in English, the sequence of phonemes /br/ can occur at the start of a word like “bread” or “brief” but is not permitted at the end of word. Conversely the sequence /nt/ can occur at the end of a word (“want”, “tent”), though not at the beginning of a word. These constraints are in many cases unique to a particular language, for instance, English does not permit the sequence /vl/ at the start of a word, whereas this sequence is commonly found at the start of words in Dutch or Russian.

Preferential listening experiments suggested that infants may use these phonotactic constraints to distinguish between languages. For instance, Jusczyk, Friederici, Wessels, Svenkerud and Jusczyk (1993b) demonstrated that 9-month old infants prefer to listen to words in their native language (though this pattern was not shown at 6 months). Although both the Dutch and English nouns used in this

experiment typically have a strong/weak stress pattern, they have different constraints on legal and illegal phoneme sequences. These results therefore suggest that by the age of 9 months infants may be aware of the phonotactic properties of words in their native language. Further evidence for this proposal comes for Jusczyk, Luce and Charles-Luce (1994) who observed that 9 month-old infants prefer to listen to lists of monosyllabic non-words that contain high-frequency phoneme sequences (e.g. “chun”) than to lists containing low-probability sequences (e.g. “yush”).

Thus by the age of 9 months, infants have acquired knowledge of the typical sound patterns (both metrical and phonotactic) of words in their native language. These findings indicate that infants have acquired some knowledge of words in their native language. These findings are significant for our understanding of lexical segmentation; both metrical stress and phonotactic information has been proposed as cues that could be used to break the speech stream into words. Research has therefore focussed on whether infants can use this knowledge in segmenting words from longer utterances.

An extension to the headturn preference procedure has allowed investigations of infants abilities to segment words from connected speech (Jusczyk & Aslin, 1995; see Jusczyk, 1999 for a review). Infants are first familiarised with multiple repetitions of a word (either in a list of isolated words or as a word that is found in several unrelated sentences). In a subsequent test phase, infants are then presented with lists or sentences (whichever was not presented previously) containing the same or different words as the familiarisation phase. The duration of head-turns towards the loudspeakers used to present each test stimulus provides a measure of familiarity. Any significant difference in listening times between the two stimuli provides evidence that infants retain knowledge of the familiarised word forms.

For instance, Jusczyk and Aslin (1995) showed that 7.5 month-old infants familiarised with repetitions of the word “cup” listen longer to sentences that contain “cup” than to sentences which did not contain that word. Similar results were also obtained when infants are tested with words when familiarised with sentences. In a follow up experiment, it was shown that infants at this age did not show an equivalent preference when familiarised with near neighbours of the test word (e.g. training on “tup” did not produce a listening preference for “cup”). Thus, infants of 7.5 months (though not 6 month-olds) are able to retain a detailed representation of the sound patterns of words in order to detect those same words subsequently. Further investigations have shown that infants retain some memory of these familiarised words in testing sessions two weeks after the initial familiarisation (Jusczyk & Hohne, 1998).

These findings demonstrate that infants are able to segment word-forms from connected speech. An experiment carried out by Saffran and colleagues (Saffran, Aslin & Newport, 1996) suggests one cue that appears to be used by infants to divide the speech stream into words. In this study, infants were presented with two-minute long sequences of synthetic speech composed of continuous repetitions of four different tri-syllabic words (e.g. “tibudo” or “pabiku”). Since each syllable occurred in more than one word, infants would have to learn the order of syllables (as well as their identity) if they were to segment words from this continuous stream. Nonetheless, after only a few minutes of training, 8 month-old infants prefered to listen to words from the training set than words generated by combining the last syllable of one word with the first two syllables of another (e.g. “dopabi” or “kutibu”). Since the only information available to infants during training concerned which syllables followed others, Saffran and

colleagues conclude that this information alone (transitional probabilities) was sufficient for infants to segment words from continuous sequences of speech.

Further evidence of the cues used by infants in detecting words in connected speech come from experiments investigating infants familiarised with bisyllabic words with different stress patterns (Jusczyk, Houston and Newsome, 1999). A series of experiments demonstrated that 7.5 month-old infants were able to segment strong/weak bisyllables ("kingdom" or "hamlet") from sentences (showing a familiarity preference for these words but not related items like "king" or "ham"). However, infants at this age were still liable to mis-segment words with a weak/strong stress pattern; for example detecting the words "tar" and "vice" following familiarisation with words like "guitar" and "device". Furthermore, when these weak/strong items were followed by a consistent syllable (for example, if the word "guitar" always followed by "is" to make the weak/strong/weak sequence "guitaris") then the infants would tend to treat the strong/weak unit ("taris") as familiar rather than the word "guitar". These results indicate that sequential constraints on syllable sequences are combined with a strong bias towards assuming that metrically-stressed syllables mark the start of a word.

Computational simulations

These experimental studies illustrate two forms of knowledge acquired during the first year of life that contribute to infants' ability to segment words from connected speech. Interpretations of these experimental findings have been enhanced by neural network models designed to simulate the means by which knowledge of phonotactics and metrical stress contribute to lexical segmentation.

For both metrical and phonotactic cues, simple models can be proposed in which the occurrence of a particular pattern of input can inform the placement of word boundaries. For instance, a sequence of phonemes like /mgl/ is unlikely to occur within a word in English, but can occur between words (such as in the sequence "same glove"). Knowledge of sequences of sounds that are unlikely to occur within a word therefore provide a cue that can be used to propose word boundaries in an otherwise unsegmented speech stream (Harrington, Watson and Cooper, 1989). Similarly, since content words typically begin with a strong syllable placing a word boundary before fully-stressed syllables would correctly segment many words in connected speech (Cutler and Carter, 1987; Cutler and Norris, 1988)

However, although these models can detect word boundaries in a stream of connected speech, neither will suffice as an account of how infants learn to use metrical or phonotactic cues. Since infants do not hear substantial numbers of single words in parental speech, computational accounts of the acquisition of lexical segmentation are faced with a boot-strapping problem. How could a system learn these or other cues to segmentation without prior knowledge of the location of word boundaries? Connectionist models of the acquisition of lexical segmentation have described two strategies that could be used to learn cues to the location of word boundaries without explicitly marked boundaries being present in the input.

Learning from utterance boundaries

One account of how infants' learn to segment connected speech is that they learn the metrical and phonotactic properties of word boundaries by generalising from the properties of boundaries between utterances (Aslin et al, 1996). Since there are consistent acoustic cues (e.g. pauses and changes in pitch) to indicate boundaries

between utterances, infants can use these cues to identify word boundaries that fall at utterance boundaries. Infants can then use the metrical and phonotactic properties of utterance boundaries for the segmentation of words within utterances.

Aslin et al. (1996) presented a connectionist model that implemented this segmentation strategy. They trained a three-layer, feed-forward neural network to map from phoneme trigrams to an output unit that was activated at boundaries between utterances. Input to the network was provided by a 3 segment window that stepped through a corpus of child-directed speech one segment at a time. When exposed to a test corpus, the network activated the output unit not only at utterance boundaries but also at many word boundaries. With an appropriate threshold on the output, the network identified over half of the word boundaries in the test corpus. This network therefore learns to lexically segment connected speech by detecting trigrams which typically straddle boundaries between utterances.

The task of identifying utterance boundaries provides a psychologically plausible means by which infants could learn a trigram-based segmentation strategy similar to that proposed by Harrington et al. (1989). A system such as this could therefore account for results suggesting that infants are sensitive to phonotactic constraints on sequences that occur at the start and end of words (c.f. Jusczyk et al, 1993b) by learning sequences that occur before or after an utterance boundary. However, other results in the experimental literature might prove more problematic for this model. For instance, Saffran et al. (1996) report that infants are able to detect words in a stream of speech that is presented without pauses or utterance boundaries.

Distributional accounts of segmentation

An alternative computational mechanism for acquiring segmentation operates by dividing the speech stream into frequently occurring sequences. It is assumed that these high frequency sequences will form meaningful units or words, while infrequent phoneme sequences are likely to straddle a word boundary. Thus, the frequency and distribution of sound sequences can be used for lexical segmentation without requiring that word boundaries are explicitly marked in the input. This technique for dividing utterances into words was originally proposed as a technique for linguistic analysis by Harris (1955). More recently this idea has been proposed as an account of how infants divide speech into words – under the catch-all term 'distributional regularity' (Wolff, 1977; Brent & Cartwright, 1996; Brent, 1999).

As suggested above, these distributional approaches encompass two distinct, but related strategies – grouping frequently occurring sequences to form words (a 'synthetic' approach), and placing boundaries at infrequent transitions in order to break longer utterances into words (an 'analytic' approach). In symbolic computational systems, these distinct approaches could be implemented separately using different computational mechanisms (see Brent, 1999 for further discussion). However, recent connectionist simulations suggest that both approaches may be served by a single computational mechanism – the prediction of upcoming input.

One interesting implementation of this distributional approach involves training a neural network to predict subsequent phonemes based on the current and previous input. This prediction task is made easier if the next phoneme occurs within the same word as previous phonemes. For instance, following the sequence of phonemes /trɛs/,

there are only two likely segments that could continue this sequence as a single word (/p/ in “trespass” or /l/ in “trestle”). However, if this same sequence of sounds occurred before a word boundary (for instance, at the end of the word “actress”) then the following segments will be much less constrained. Thus segment prediction will be much more accurate during a word than immediately before a word boundary.

One influential demonstration of this approach was reported by Elman (1990). A simple recurrent network was trained on this segment prediction task using a small artificial language, presented a segment at a time without word boundaries. Elman observed two properties of the network’s output error on test sequences. First, the network’s prediction error decreased later on in a word – as the current input matched fewer and fewer words in the network’s vocabulary, future segments in the word could be predicted with greater accuracy. The network therefore acquired some knowledge of words in its training vocabulary. Elman’s second observation was that prediction error increased sharply at the end of a word, as a result of the greater variety of phonemes that can occur after a word boundary. The predictability of segments therefore provides a metric not only for grouping segments into words but also for detecting word boundaries.

Models that include the prediction task have been suggested as an account of the experimental results of Saffran, Newport and Aslin (1996). In these experiments, infants became familiar with words from training sequences that are presented without boundaries between words or utterances. Infants learn that the sequences includes groups of predictable syllables. Test sequences which contain whole words from the training set will therefore be more predictable (and hence more familiar) than sequences which combine parts of different words. Recurrent network simulations using similar

materials have shown that models using the prediction task are able to simulate these experimental results (Allen and Christiansen, 1996).

Simulations reported by Cairns, Shillcock, Chater and Levy (1997) extend these recurrent network prediction systems to a large corpus of phonologically transcribed conversations; scaling up the Elman (1990) simulations to a realistic input. Consistent with these earlier simulations they showed that error peaks in segment-prediction can be used to detect word boundaries in a test corpus. However, even when carefully optimised this system only detects 21% of word boundaries. Although this is superior to chance performance, it still falls short of the level of performance that would be required for word identification especially since the network placed boundaries within many words (a hit:false-alarm ratio of 1.5:1). In describing the performance of their network, Cairns and colleagues suggested that boundaries were placed between phonotactically well-formed syllables rather than between words. Thus the lexical knowledge that was acquired in Elman's small scale simulations may not scale up to larger training sets.

Combining multiple cues for segmentation and identification

One way to improve the performance of segmentation systems is to incorporate more than one strategy for the detection of word boundaries (Christiansen, Allen & Seidenberg, 1998). Christiansen et al. described simulations in which a simple recurrent network was trained with three cues that have been proposed for segmentation – utterance boundaries (cf. Aslin et al., 1996), phoneme prediction (cf. Elman, 1990) and metrical stress (cf. Cutler and Carter, 1987). These cues were presented to different networks either singly or in combination. Interestingly, the performance of networks trained on all three cues exceeded the performance of networks trained on single cues or

pairs of cues. This combined system also out-performed the systems reported by Aslin et al (1996), or Cairns et al (1997) detecting 74% of word boundaries, with a hit:false-alarm ratio of 2.3:1. Christiansen and colleagues propose that combining multiple cues is a particular productive strategy for language acquisition in general and lexical segmentation in particular. Segmentations that are predicted by a single cue may be unreliable, whereas segmentations supported by multiple cues are more likely to correspond to true word boundaries in the input.

Nonetheless, this combined approach still falls short of segmenting all words in the speech stream. As might be expected, a system that detects less than three-quarters of word boundaries still fails to segment half of all words from connected speech. Furthermore, the lexical effects that observed for small vocabulary recurrent network models (Elman, 1990) were still not observed. Christiansen, Allen and Seidenberg (1998) reported that the prediction error reflected knowledge of phonological clusters that occurred in many words in the training set, and did not capture specific phoneme sequences that occurred in any single lexical item.

Thus, despite the potential for systems trained on prediction-tasks to learn sequences corresponding to individual lexical items, it is clear that for realistically-sized vocabularies these systems do not segment the speech stream by storing chunks of speech as familiar words. It is possible that the prediction task does not place sufficient demands on these networks to retain information from early time-points (since local information may be sufficient to predict subsequent segments). Alternatively, some more fundamental limitations on the memory capacity of recurrent neural networks for learning long distance dependencies provide the limiting factor on the performance of

these systems (see, for instance, Servan-Schrieber, Cleeremans & McClelland, 1991; Maskara & Noetzel, 1993; see Rhode & Plaut, this volume for further discussion).

In view of these limitations, computational models of how infants begin to acquire and store word forms have mostly proposed symbolic systems that determine the most likely (i.e. maximum probability) set of words in an utterance (Brent and Cartwright, 1996; Brent, 1999). Sections of speech that have been hypothesised to be words are stored and reused to segment subsequent utterances. Although this approach successfully simulates how infants discover words in connected speech, unrealistic assumptions are made regarding the computational resources available to infants. In particular these algorithms require (1) an unbounded and undecaying memory for storing potential vocabulary items (2) pre-existing mechanisms to compensate for the noise and variability that exists in all speech and, in some cases (3) built-in knowledge of phonotactic constraints on viable and non-viable segmentations. The increased effectiveness of these models therefore comes at some cost to their psychological plausibility.

One goal of the computational account developed here is to explore the potential for recurrent neural networks to not only simulate the development of lexical segmentation, but also to account for the identification of lexical items and the acquisition of the mapping from spoken word-forms to meaning. These simulations allow us to explore whether the failure to observe lexical effects in recurrent networks reflects an intrinsic limitation of the processing capacities of these neural network models or, simply, that alternative cues or learning strategies are required. Given this interest in vocabulary acquisition we will review the developmental literature concerning how infants learn to map from word forms to meaning.

Vocabulary acquisition

Infants face a number of difficult problems in learning how words are paired with meaning. From a philosophical perspective there are a potential infinity of referents that a single new word could denote (Quine, 1960). For instance, a child learning the pairing between the word 'rabbit' and a furry stuffed toy may be unclear whether the word refers to the whole animal, a part of the animal, or indeed some entity that is not present in the current scene. A variety of strategies have been proposed to account for infants' ability to learn language in spite of this seemingly intractable problem – for instance, cues that help infants determine the likely referents of words that they hear.

For example, experiments on how infants categorise objects that are accompanied by novel words suggest a bias towards assuming that new words refer to whole, bounded objects rather than to other possible referents such as parts of these objects, the materials from which they are made, their colour etc. (Waxman & Markow, 1996). It is unclear whether this bias reflects the operation of a constraint that is specifically tuned to detecting those aspect of the environment that have been linguistically-labelled, or whether infants share with adults a more general bias towards treating objects in the world as salient (Bloom, 1994). Nonetheless, since many of the earliest words that are learnt by infants refer to concrete nouns (Fenson, et al., 1994) this bias is apparent in early language acquisition even if the precise cause remains unclear.

Another sources of constraint that infants could use to help determine the possible referents of words in connected speech is to pay attention to non-verbal cues. For instance, by observing the direction of gaze of the speaker or other forms of

'pointing' behaviour infants are provided with cues to the referents of an utterance even in the absence of any comprehension of the linguistic label. Infants appear to use this cue from an early age in determining the objects to which words in speech refer (Baldwin, 1991; 1993).

Mapping from speech to meaning

These and other strategies assist the infant in determining the meanings of unknown words and will therefore reduce the number of possible target meanings for the words that they hear. However, learning the speech-to-meaning mapping cannot be reduced to a simple one-to-one association. On any occasion on which more than one word is spoken (i.e. for the majority of the utterances heard by infants) there will be more than one word that can be learnt and therefore more than one target referent for the words in that utterance. For instance, infants' experience of the word "cat" may arise from multiple-word utterances like "look at the fat cat", "that cat is sitting on the fence again", "does the cat need feeding?". A one-to-one mapping between the sounds of words and their meanings is not specified in the learning environment, but must be discovered by infants. An important question therefore remains: how it is that children discover the one-to-one correspondence that exist between sequence of sounds and their referents in order to associate the word /kæt/ with the furry animal to which this word commonly refers?

One proposal concerning how infants discover these one-to-one correspondences is that they will analyse multiple utterances and make use of the word-meaning pairings that these different utterances have in common. This very general idea of 'cross-situational learning' has been proposed by many authors, including Pinker

(1989) and Gleitman (1994). Symbolic models of vocabulary acquisition have included a more formal description of algorithms that permit these cross-situational inferences (Siskind, 1996). However, connectionist models of early vocabulary acquisition have so far not considered the problems that are involved in discovering word-form to meaning mappings. Existing connectionist models of early vocabulary acquisition (such as Plunkett, Sinha, Møller and Strandsby, 1992, or Plaut and Kello, 1998) have focussed on other issues, and have therefore used training sets in which word-forms and word-meanings are paired on a one-to-one basis. The simulations conducted here investigate the acquisition of one-to-one mappings between words and meanings without requiring that these correspondences are explicitly specified in the training set.

Experimental investigations of early vocabulary acquisition

In specifying mechanisms for learning the mapping from words to meaning, it might be expected that infants build on their pre-existing linguistic knowledge. As reviewed previously, a variety of sources of evidence have demonstrated infants' acquisition of language specific knowledge during the second half of their first year of life. In particular, evidence has suggested that, at the age of 7.5 months, infants are first able to isolate and recognise single words from connected speech (Jusczyk & Aslin, 1995). The age at which infants' develop the ability to relate words to meanings is largely consistent with the assumption that vocabulary acquisition begins after infants are able to segment words from connected speech.

Investigators have used preferential looking procedures to derive evidence of infants' earliest comprehension of the meanings of words. Typically, infants are presented with an array of two or more objects, and their tendency to fixate a particular

object if an appropriate name is produced is compared with fixations following a novel or unfamiliar word (Oviatt, 1980; Thomas, Campos, Shucard, Ramsay & Shucard, 1981). This method, has shown comprehension of words for concrete objects in infants as young as 13 months. However, since this method is susceptible to biases arising from infants' visual preferences (e.g. a preference for fixating objects for which a name is known – Shafer, Plunkett & Harris, 1999) only cautious conclusions should be drawn from these comparisons. More robust evidence is obtained from experiments comparing preferences for looking at appropriate versus inappropriate objects where names for both objects are known. With this more careful procedure, the age of earliest comprehension is raised to approximately 15 months (Golinkoff, Hirsh-Paskek, Cauley & Gordon, 1987). These results are consistent with the earliest estimates of when infants can be shown to learn novel pairings between words and objects (Shafer and Plunkett, 1998). Such demonstrations require that infants are taught two novel names for two novel objects (avoiding confounding effects of pre-existing familiarity with either words or concepts and hence confounding biases). Under these tightly controlled conditions, 15 month old infants show learning of new names such as "bard" and "sarl" for photos of two novel objects after only twelve pairings of the word and the concept.

One important issue for investigations of the representation and processing of spoken words concerns how word forms are represented and how those representations are activated during word comprehension. It has been suggested that infants' representations of their first words are not structured in an adult-like, segmental fashion, but may comprise holistic, whole-word representations (Walley, 1993). Infants may only structure their word representations using individual phonemes (e.g. storing "cat"

as /kæt/) when their vocabulary has reached a sufficient size for other neighbouring words such as “rat”, “kit” and “cap” to be known (Charles-Luce & Luce, 1995).

Experimental evidence concerning the time-course of word identification in children, however, has cast doubt on the holistic representations proposed by Walley (1993). Investigations of the timing of fixations towards the referents of heard words (Fernald, Pintos, Swingley, Weinberg & McRoberts, 1998) have demonstrated that during the second year of life, infants become increasingly skilled at mapping the sounds of speech onto their meanings. Between 15 and 24 months infants fixations towards pictures referred to by spoken words become increasingly rapid – despite little evidence of developmental changes in the speed with which saccadic eye-movements can be programmed and executed. Fernald and colleagues (1998) have shown that by 24 months in age, infants can initiate a saccade towards the appropriate picture before the acoustic offset of a word. This finding suggests that infants, like adults, can identify words at the earliest point at which sufficient information becomes available in the speech stream (see Marslen-Wilson, 1984 for discussion).

This theory is further supported by experiments reported by Swingley, Pinto and Fernald (1999) showing that this rapid and efficient word processing is accompanied by an adult-like time-course of identification. In 24-month-old infants, fixations to target pictures are delayed for stimuli in which two competing items share their initial sound (e.g. “tree” and “truck”), exactly as predicted by accounts of spoken recognition in which speech processing keeps track with the speech stream. Even with the small receptive vocabularies typical of infants at this age, these results suggest that word representations during early stages of acquisition are organised as sequences of

phonemes and processed incrementally – consistent with the sequential processes observed in adult listeners (Marslen-Wilson & Welsh, 1978).

The goal of the simulations reported here is to investigate connectionist networks that can account for the developmental profile that has been observed in the literature on lexical segmentation and vocabulary acquisition. Any psychologically plausible account must fit two primary requirements:- (1) the system must simulate the behavioural profile that has been observed in infants, and, (2) the model must make realistic assumptions concerning the processing mechanisms and sources of information that are available to infants.

Computational models of spoken word identification

The task of recognising words in connected speech can be described as a mapping from sequences of input representing the sounds of speech to a lexical/semantic representation of the word or words contained in the speech stream. Recurrent network models of this mapping (Norris, 1990; Gaskell & Marslen-Wilson, 1997) simulate the time-course of word identification for adult listeners as a continuous process in which the activation of lexical representations responds immediately to incoming information in the speech signal. In training these recurrent networks, the target representation is specified throughout the word, irrespective of whether sufficient information is available in the input at that time. Thus the network is frequently given an impossible task – to identify words from only their initial segments. The effect of this time pressure, however, is to ensure that in testing the network, the appropriate lexical representation is activated as soon as sufficient information becomes available. Input sequences which are consistent with more than one word lead the network to

activate multiple lexical representations in proportion to the probability that they represent the current word in the speech stream. For example, in response to the sequence of sounds /kæptɪ/, lexical representations for “captain” and “captive” will be partially activated. At the offset of /kæptɪn/, when the input matches only a single word, the appropriate lexical representation is fully activated.

An important limitation of these simulations was revealed by Norris (1990; 1994). Specifically, these networks have problems in recognising short words embedded at the start of longer words (such as the word “cap” that is embedded in “captain”). At the offset of the syllable /kæp/, the network will activate short words and longer competitors equally. For a sequence like “cap fits”, in which a longer word can be ruled out, the network uses the following context to identify the subsequent word (activating lexical items that begin with /f/ such as “fits”, “feels”, etc.), but is unable to use this input to revise its interpretation of the syllable /kæp/. Thus onset-embedded words like “cap” remain ambiguous and can not be identified by the network.

One solution to this problem is to allow speech input arriving after the offset of a word to play a role in the identification of previous words in the speech stream. In models such as TRACE (McClelland & Elman, 1986) or Shortlist (Norris, 1994) this is achieved by incorporating inhibitory connections between lexical candidates, such that information arriving later on in the speech stream can affect the lexical activation of earlier words. However, these inhibitory connections are hard-wired in TRACE or dynamically re-wired in Shortlist. It is at present unclear how this additional competition mechanism can be incorporated into a network trained by back-propagation or some other gradient descent algorithm.

A further limitation of these recurrent network models as an account of development is that the training procedure assumes that a one-to-one correspondence between speech stream and lexical/semantic representations is available in the learning environment. The network is provided with a target representation that, at every time-step in the input, specifies the appropriate lexical/semantic representation for the current word. This training regime requires not only that word-boundaries are specified beforehand but also, and more importantly, that a target lexical representation can be assigned to each word in the speech stream. The assumption that is implicit in this training procedure is that infants are supplied with the one-to-one relationship between words and meanings prior to vocabulary acquisition. This is exactly analogous to the one-to-one pairings that we described as unrealistic in some connectionist models of vocabulary acquisition (Plunkett et al., 1992; Plaut & Kello, 1998). These models all assume that words are learnt by a process of ostensive definition by which infants hear a single word utterance and are directed to the meaning of that word. As we described previously, this situation does not capture crucial aspects of the learning problem faced by infants.

The recurrent network simulations that are explored here demonstrate that a very simple change to the training assumptions of the model provides a solution to both of these limitations of previous recurrent network simulations. Providing a recurrent network with a more developmentally-plausible training set (i.e. not including one-to-one correspondences between speech and meaning) results in a system that is able to identify all the words in the training set, including onset-embedded words. To compensate for the increased complexity produced by removing these one-to-one correspondences we make an extreme, simplifying assumption that the meaning of each

word in a sequence can be represented by a single, lexical node. Although this assumption is clearly false – the meanings of individual words vary considerable depending on context and therefore cannot be represented by a single, fixed representation – this assumption can in part be justified by suggesting that for the names of concrete objects (which form the heart of infants' early vocabularies), a categorical representation of certain classes of concrete concepts may be available to infants prior to vocabulary acquisition (see, for example, Quinn, Eimas & Rosenkrantz, 1993).

Simulation 1: Learning to identify words in connected speech

This simulation explores the effect of altering the training task for a recurrent network model of spoken word identification. Whereas previous models (e.g. Norris, 1990; Gaskell & Marslen-Wilson, 1997) were trained to activate a representation of the current word in the input, the networks presented here are trained to activate a representation of an entire sequence of words. The network must maintain an active representation of all the words that have been heard until the end of an utterance. By extending the networks' task so that it must continue to activate an identified word after its acoustic offset, we can ensure that the system can resolve the temporary ambiguities created by onset-embedded words.

Since the network is trained using a fixed target representation for an entire sequence of words this training regime no longer includes a one-to-one pairing of words and their meanings. Instead, the network is exposed to a variety of word sequences paired with a target representation in which all the words in each sequence are activated. The task of the network is to uncover the set of one-to-relationships that best capture the contribution of each input word to the target representation (cf. Goldowsky & Newport,

1993; Roy & Pentland, 2002; Siskind, 1996). This training assumption is analogous to the cross-situational learning proposed by Pinker (1989) and Gleitman (1994).

A similar approach to language comprehension is described by St. John and McClelland (1990) for a model in which the goal of the comprehension system is to activate a 'sentence gestalt' capturing the meaning and thematic relationships between words in sentences. The output representation used in the current simulations is simpler than that used by St. John and McClelland (1990), since it employs localist units, each representing the meaning of a single word in the networks' vocabulary. Although structured as discrete lexical units, this aspect of the model is intended as a computational convenience rather than as an integral part of the account. Distributed output representations would provide a more realistic account since the network would then be forced to extract a consistent lexical/semantic representation from the noisy and contextually variable meanings of words in different sequences. However, this change to the model would greatly increase the amount of computer time required to train the networks without substantially altering the behavioural profile of the simulation except for circumstances in which multiple items were very weakly activated (see Gaskell & Marslen-Wilson, 1999 for illustrative simulations and further discussion).

A further advantage of the localist output representations used here is that they avoid the binding problem that is incurred in combining existing representations of single words to produce a distributed representations of a sequence (see Page, 2000; Sougné, 1998 for further discussion). More complex representation schemes such as temporal binding (Shastri & Ajjanagadde, 1993) or tensor-product binding (Smolensky, 1990) have been proposed to allow the use of distributed representations that can represent multiple words simultaneously without interference. However, these more

complex output representation would further increase the size and complexity of the simulations. The networks presented here provide a simple demonstrations of the computational properties of recurrent networks without requiring a solution to this contention issue. However these simulations must therefore come with the caveat that scaling up to more realistic semantic representations may present additional problems.

Method

A simple recurrent network (Elman, 1990) was used for these simulations. The network was trained with back-propagation to map from sequences of distributed representations of phonemes (as sets of binary phonetic features) to an unordered localist representation of all the words in each sequence. This training target remains static throughout each sequence of words so that the network is not provided with any information about the location of word boundaries, nor which segments in the input map onto individual lexical items. The network must extract the one-to-one correspondences between speech and lexical items from this many-to-many mapping.

<INSERT FIGURE 1 APPROXIMATELY HERE>

The training sequences for the network were generated from an artificial language with 7 consonants and 3 vowels placed in CVC syllables. This language contained 20 lexical items (14 monosyllables and 6 bisyllables) which varied in the segment at which they became unique from other words. This word set included 'cohort' pairs (such as "lick" and "lid" that shared onsets), onset-embedded words ("cap" and "captain") and offset-embedded words ("lock" and "padlock"). Words were selected at random (without replacement) to create sequences between 2 and 4 words in length. Individual sequences were separated by a boundary marker (an input and output vector of zeros). Ten networks were trained from different sets of random initial weights

and with different random sequences using back-propagation of error (r=0.02, no momentum, cross-entropy output error – see Davis, Gaskell & Marslen-Wilson, 1997, for more details) until output error stabilised (500 000 training sequences). The architecture of the network and a snapshot of activations during training is shown in Figure 1.

Results

Figure 2 shows the activation of lexical units for an illustrative test sequence averaged over ten fully trained networks. The network activates words as their constituent segments are presented at the input. Lexical units are partially activated in cases of ambiguity (for example "lick" is partially activated by the onset of "lid"), with the output activation approximating the conditional probability of each word being present in the input sequence. Full activation is consequently only observed when words are uniquely specified in the input. In contrast to previous recurrent network simulations the network is also able to identify onset-embedded words, by using segments in the onset of the following word to rule out longer competitors. For example, in Figure 2, the word "cap" is only identified when information in the onset of the following syllable (/l/ from "lock") rules out the longer word "captain". Thus the network can resolve the temporary ambiguity created by onset-embedded words (see Davis, Gaskell & Marslen-Wilson, 1997; 2002 for further discussion).

<INSERT FIGURE 2 ABOUT HERE>

Since this model is intended to account for the development of spoken word identification in infants, it is important that the developmental profile is assessed. The performance of each network was therefore tested at after every 5000 training sequences; measuring the networks' ability to recognise individual words in test

sequences. A word was considered to be "recognised" if at some point during the input sequence, the network activated the appropriate output unit to a value 0.5 higher than all other competitors[3]. To simplify the test procedure, only responses to the first word in a sequence were considered in this analysis. Results were averaged over every possible two word sequence (19 sequences for each word) in each of the ten networks.

The networks' recognition performance throughout training is shown in Figure 3. As can be seen, the network shows a rapid growth in recognition performance. This vocabulary spurt (as in other acquisition models – e.g. Plunket et al, 1992) may be analogous to the rapid increase in the number of words that infants comprehend that is typically observed during the second year of life (Fenson et al., 1994).

<INSERT FIGURE 3 ABOUT HERE>

A more interesting question is whether the network shows the same gains in the speed of word identification shown in eye-movement data by Fernald et al. (1998). Although these gains in performance may appear unsurprising – after all, infants show improvements on a range of tasks as they get older – however, this increased speed of identification is accompanied by an increase in the number of similar sounding words that infants know. Word identification will therefore require more fine-grained discriminations for older infants with larger vocabularies. The presence of lexical competitors has been shown to delay word identification in adults (see Monsell & Hirsch, 1998, for relevant experimental data and further discussion); it is therefore of interest to observe whether the network also shows improvements in the speed of word identification when rapid increases in vocabulary size are seen.

A measure was therefore taken of the time at which individual words are identified by the network. These recognition points were calculated as the exact number of phonemes (starting from word onset) at which the identification threshold was achieved (i.e. output activation should be 0.5 greater for the target word than for competitors). As in Gaskell & Marslen-Willson (1997) we used linear interpolation between activations at successive segments to improve the accuracy of identification points; though near-identical results were obtained without interpolation. The mean recognition point throughout the training of ten networks is shown in Figure 3. As can be seen, the networks show marked changes in the speed with which words can be identified throughout vocabulary acquisition. Recognition points become substantially earlier with increased training, consistent with the experimental data reported by Fernald et al. (1998) and despite increases in the size of the networks' receptive vocabulary (i.e. improved recognition performance).

Recognition points were also computed for two subset of the items in the network that had different lexical environments. As shown in Figure 4, cohort-competitors (pairs that share their initial CV like "lick" and "lid") are identified at a later phoneme than words that have an equivalent phonemic overlap, but do not share an initial CV (e.g. "bat" and "cat"). This result is consistent with the experimental data presented by Swingley, Pinto & Fernald (1999) in which identification of pairs like "truck" and "tree" that share an onset are delayed by comparison with rhyming pairs like "duck" and "truck".

Interestingly, the advantage for non-cohort pairs is not observed at the earliest time points at which these items are recognised (i.e. before approximately 40 000 training sequences – a point at which only around 30% of words are correctly

identified). It may be that early on in training, these networks do not use sequential information to rule out mismatching items in the same way as at later stages of development when words can be identified with greater accuracy. This prediction that cohort effects only emerge once word recognition is reasonably reliable could be tested experimentally were it possible to repeat the experiments of Swingley et al. (1999) with younger infants.

<INSERT FIGURE 4 ABOUT HERE>

Discussion

This model provides an effective simulation of the time course of identification of words in sequences – progressively updating lexical activations as more input is presented. Unlike previous recurrent network accounts (Gaskell & Marslen-Wilson, 1997; Norris, 1990) the model is able to resolve temporarily ambiguous input for sequences in which post-offset information is required – as is the case for onset-embedded words like “cap” in “captain”. The developmental profile shown by this model is suggestively similar to the profile shown by infants during vocabulary acquisition. The network shows gains in the speed of word processing during a period of rapid learning, consistent with the advances shown by infants during the second year of life.

One important difference between these simulations and previous recurrent network accounts of spoken word recognition is the use of a training set in which the input is not segmented into words and in which output representation does not have pre-specified correspondences with the speech input. In these simulations the target representation provides only the identity of the words contained in an utterance; the networks are not provided with information on the order in which words occur or the

location of boundaries between words. By generalising from experience of different input sequences and activated output units, the network learns the set of one-to-one correspondences between the speech stream and lexical representations. At least for the artificial language investigated here, input-output correspondences (analogous to regularities in the mapping from word-form to word-meaning) provide a cue that the network can use to segment and identify words in connected speech.

Although the networks in Simulation 1 learn to segment the speech input into lexical items by detecting input-output correspondences, this is clearly not the only means by which the speech stream can be divided into words. These networks come to the task of vocabulary acquisition without any knowledge of words and word-boundaries in connected speech – an assumption that is unlikely to be true for the developing infant. A range of evidence has already been reviewed suggesting that, by the end of the first year of life, infants have considerable knowledge of their native language at their disposal (e.g. phonotactics, metrical-stress); that they can use as cues to identify the boundaries between words. This evidence questions one of the assumptions made in Simulation 1 – namely that the speech input is unsegmented prior to lexical acquisition. Further simulations were therefore carried out to explore how infants' abilities to segment the speech stream may contribute to vocabulary acquisition in this model.

Simulation 2: Combining phonological learning and vocabulary acquisition

The success of the distributional accounts of lexical segmentation that were reviewed previously suggest that simple, statistical mechanisms play an important role

in discovering the boundaries between words. Two developmentally plausible mechanisms were described which allow connectionist networks to discover the location of word boundaries – generalising from the properties of utterance boundaries (Aslin, et al., 1996) and using prediction error to determine sections of the speech stream that cohere as words or are likely to contain a word boundary (Elman, 1990; Cairns et al., 1997). Simulations have also shown that the combination of these two strategies produces more accurate segmentation than either approach alone (Christiansen, Allen & Seidenberg, 1998).

Given the success of these statistical mechanisms in detecting the boundaries between words, and the segmentation abilities of infants in their first year of life (before vocabulary acquisition), it is of interest to investigate whether similar computational mechanisms might benefit the connectionist model of vocabulary acquisition presented here. It is likely that a system provided with information concerning the location of word boundaries would be more successful at learning associate speech and meaning than a system without this information. However, the second set of simulations asked a different question – namely whether providing the network with mechanisms previously shown to support the learning of cues to word boundaries will assist the acquisition of word-meaning correspondences.

Method

The approach taken here was to retrain the networks from Simulation 1, adding an additional set of output units that predict future input segments and utterance boundaries. By comparing the developmental profile of networks trained with this prediction task with the results of Simulation 1 the role of distributional analysis in vocabulary acquisition can be explored. Ten networks were therefore trained using the

same network architecture, learning parameters and randomly generated training sets as the ten previous simulations. Those weights common to the two sets of networks (i.e. connections from input to hidden units, recurrent hidden unit connections and connections linking the hidden units to the output) were initialised to the same random values. The one change to the network was to add a set of output units which were trained to activate a representation of the input segment or utterance boundary (an output vector of zeros) that would be presented at the next time step. The only difference between the two sets of simulations was the presence of the additional prediction output in Simulation 2. A depiction of the network during training is shown in Figure 5.

<INSERT FIGURE 5 APPROXIMATELY HERE>

Results and discussion

Inspection of the lexical output of ten fully-trained networks that used the prediction task showed an identical behavioural profile to that reported for Simulation 1 and illustrated in Figure 2. The network identifies words in speech as they are presented in the input and those lexical items remain active until the end of the current utterance. More interesting results, however, are obtained in the comparison of the networks' profile during training. The analysis of the networks' recognition performance was again conducted after every 5 000 training sequences, and performance was compared with results from Simulation 1. Figure 6 shows the average performance of ten networks trained with and without the prediction task, comparing percent correct recognition (Figure 6a) and recognition points (Figure 6b).

<INSERT FIGURE 6 APPROXIMATELY HERE>

As can be seen in Figure 6a, the addition of the prediction task significantly speeds lexical acquisition. Comparing the results of Simulations 1 and 2 shows that networks trained with the prediction task recognise more words than networks that receive the same amount of training without the prediction task. Thus vocabulary acquisition in the network is significantly speeded by the addition of the prediction task. Furthermore, comparison of the recognition points depicted in Figure 6b indicates that networks trained with the prediction task not only recognise more words, but also recognise them more rapidly than the equivalent network trained in Simulation 1.

Both sets of output units in Simulation 2 (prediction task and lexical identification) are trained concurrently, using the same set of hidden units, learning algorithm and parameters. However, the network may not be learning the two tasks at the same rate. To compare the networks' learning profile on each of these tasks, RMS output error (normalised for the number of units in each portion of the output) was measured at each set of output units for a test set presented every 5 000 sequences during training (Figure 7).

<INSERT FIGURE 7 APPROXIMATELY HERE>

Root mean square error measures for networks in Simulation 2 suggest that the prediction task is learnt more rapidly than the lexical task. Learning in the prediction task is mostly complete after 10 000 sequences, and reaches asymptote at around 25 000 sequences, whereas lexical learning continues until later in training. The networks' performance on the prediction task provides evidence that the network has learnt something of the structure of the artificial speech input before the network is able to map the speech input onto the correct lexical output.

This time course of acquisition is similar to the pattern observed in the developmental literature. As described previously, infants become sensitive to statistical aspects of the speech input during the first year of life. By the age of 9 months, infants prefer listening to stimuli that follow the typical sound pattern of words in their native language (e.g. words containing high frequency phonotactic sequences or a strong/weak stress pattern). It is exactly these forms of knowledge that are encoded by networks performing the prediction task (cf. Cairns et al., 1997). Therefore, learning the prediction task may be sufficient for the network to account for infants' knowledge of phonotactics and metrical stress. However, despite early acquisition of the form of words in their native language, it is only during the second year that infants readily associate words with the objects to which they refer. In these simulations it is proposed that vocabulary acquisition is modelled by the lexical output task. Thus the developmental profile observed in these combined simulations is broadly consistent with that observed in infants; lexical learning continues on from earlier phonological learning.

An important question concerns how it is that the addition of the prediction task assists the network in learning to recognise words. To pursue this issue, hidden unit representations developed by networks trained with and without the prediction task were compared. One informative measure is the amount that hidden unit activations change between successive segments in the input – i.e. the distance that the hidden-unit representations change at each time step. The Euclidean distance between hidden unit representations for successive input segments was calculated for the set of ten networks trained with and without the prediction task. These distance measures were averaged for two types of segment position in the input: (1) between segments that occur within the

same word and (2) between segments that cross a word boundary. These results, averaged over the ten networks trained in Simulations 1 and 2 are shown in Figure 8.

<INSERT FIGURE 8 APPROXIMATELY HERE>

Results indicate that, throughout training, networks with the additional prediction task made larger steps through hidden unit space in processing the input. Furthermore, for those networks trained with the prediction task, hidden unit representations changed more within words than across word boundaries. This effect of segment position on movement through the networks' hidden-unit space is particular apparent at the beginning of training. Even at the earliest test phase, networks trained with the prediction task process sections of the input that occur within words differently from sections that cross word boundaries – i.e. they show signs of having lexically segmented the input sequences. Networks from Simulation 1 that were only trained on the lexical task did not show an equivalent difference in processing input within and between words.

Thus the inclusion of the prediction task enables the networks to develop a more structured internal representations of the speech input. Phonological learning provided by the prediction task serves to 'bootstrap' lexical acquisition by chunking the speech input into units that potentially correspond to units in the output representation.

General discussion

The computational simulations that have been presented here illustrate two convergent aspects of the modelling of spoken word identification in recurrent network models. Firstly, these simulations show that training a network to preserve the activation of previous words produces an appropriate activation profile for the

identification of words in connected speech – in particular for words that are embedded at the start of longer competitors. Secondly these networks provide a plausible simulation of the developmental profile of word recognition in infants. The networks are trained without the one-to-one correspondences between speech and meaning that have been provided previously. In discovering the appropriate mapping between the speech input and lexical/conceptual representations, these networks show a realistic developmental profile since the speed and accuracy of word identification increases throughout training, consistent with experimental data from infants.

Interestingly, a single novel assumption appears to be responsible for both of these successes. The developmental plausibility of these simulations is enhanced by being trained to map entire sequences of words to a representation of an entire utterance. Similarly the networks' success at identifying onset-embedded words arises as a direct consequence of being trained to preserve the activation of lexical representations over an entire utterance.

As was discussed, before, one aspect of Simulation 1 is unrealistic by comparison with the developmental literature. These networks were presented with an unsegmented sequence of words in the input. As reviewed in the introductory section there is substantial evidence to suggest that infants are able to segment the speech stream into word-sized chunks before beginning to acquire the mapping from speech to meaning.

The ability of infants to use distributional information to segment connected speech was explicitly incorporated in Simulation 2. However, rather than supplying these networks with pre-segmented sequences, these networks were provided with an additional mechanism to assist in segmenting the speech input. The networks in

Simulation 2 were required to activate an additional prediction output trained in parallel with the lexical output. Prior simulations have shown that this prediction task allows a network to identify a substantial proportion of word boundaries (Cairns et al., 1997; Christiansen et al., 1998). An important finding from these simulations was that the addition of the prediction task significantly improved the speed with which the network learnt the task of recognising words in connected speech. Interpretations of the effect of this additional task in assisting lexical acquisition will be discussed in more detail.

Bootstrapping vocabulary acquisition

The simulations reported here demonstrated that an additional, input-prediction task assists learning in a network model of vocabulary acquisition. Such a result may appear counter-intuitive – it might have been expected that giving a network an additional task to perform would reduce the processing resources available for lexical acquisition. Connectionist simulations of other domains have shown that training a network to perform multiple tasks with a single set of hidden units can impair performance by comparison with networks which are provided with separate hidden-unit modules for each task (e.g. for recognising stems and affixes of morphologically complex words, Gasser, 1994; for the what and where vision task see Rueckl, Cave & Kosslyn, 1989; though see also Bullinaria, 2001). However, in the simulations reported here, forcing two tasks to share the same internal representation assists acquisition.

In the current simulation, both the prediction and lexical acquisition tasks were imposed on the network from the beginning of training. Nonetheless, the network showed more rapid learning of segment prediction than lexical acquisition (Figure 7). Several properties of the two mappings may be relevant in this respect. For instance, the prediction task, may be more easily solved using the input available at the current time-

step, while the lexical task depends on representations of preceding segments. Alternatively, this result may reflect greater input-output similarity in the prediction task, since the same distributed representations are used at the input and at the prediction output. In either case, since the prediction task is learnt first, it seems that it is this task (and not lexical identification) that provides an early influence on the structure of the networks' representations of the input sequences. This is evident in the marked differences between the networks' hidden representations when trained with and without the prediction task (Figure 8).

This finding helps explain the benefit that is provided by simultaneously training a network with both the lexical identification and prediction task. By re-using the hidden-unit representations that develop to perform the prediction task, the network gains a substantial head-start on lexical acquisition. As indicated in Figure 8, the hidden unit representations resulting from the prediction task provide an initial segmentation of the speech stream into lexical units which benefits lexical acquisition. This simulation therefore provides an explicit demonstration of how learning the statistical and distributional structure of the speech stream may serve to bootstrap vocabulary acquisition.

This simulation demonstrates the benefits of statistical learning as a starting point for the acquisition of higher-level mappings. This finding has obvious similarities with physiologically-inspired simulations in which the computational properties of system force a similar development from learning first-order statistics to developing more abstract representations of the input either through change to the number of hidden units in the network (Fahlman & Lebiere, 1990), through changes to the timing of weight changes in different regions of the network (Shrager & Johnson, 1996) or

through changes to the memory capacity of the network (Elman, 1993) – see papers by Quartz (this volume) and Rhode and Plaut (this volume) for further discussion. What is particularly striking in the simulations reported here is that no changes to the processing properties of the network are required to ensure that the network learns simple statistical properties of the input first. Whether this is a consequence of the particular tasks and training sets used or is a more general property of recurrent networks is unclear. Other authors have demonstrated that additional tasks assist in training recurrent networks (Maskara & Noetzel, 1993) it is therefore at least possible that this demonstration reflects a general property of recurrent neural networks.

Puzzles and contradictions in vocabulary acquisition

The simulations that have been presented here are largely consistent with developmental evidence concerning the acquisition of lexical segmentation and vocabulary acquisition. The simple assumption that has been made in relating these simulations to the developmental time course in infancy is that the prediction output accounts for infants knowledge of the statistical structure of the speech input during the first year of life and that the performance of the lexical output simulates the acquisition of mappings from speech to meaning early in the second year. This interpretation is consistent with a role for the prediction task in developing structured representations that support subsequent lexical acquisition.

While the work presented here falls short of providing a detailed account of any single set of experimental data, it seems likely that much of the existing experimental literature can be accounted for within this framework. However, there are some results in the experimental literature that appear to be inconsistent with this framework. One well-publicised result concerns the ability of infants to learn regularities that supposedly

can not be learnt by recurrent network (Marcus, Vijayan, Bandi Rao & Vishton, 1999; though various authors have subsequently demonstrated neural network accounts of exactly this data, e.g. Dominey & Ramus, 2000; Siros, Buckingham & Shultz, 2000; see also Gasser & Colunga this volume). This discussion will instead focus on results that suggest a developmental dissociation between word-form learning and the properties of the systems that map from speech to meaning. These dissociations challenge the account proposed here in which vocabulary acquisition re-uses representations that arise from the acquisition of word forms.

The first result that might challenge the account presented here was reported by Stager and Werker (1997). They observed that infants in the early stages of vocabulary acquisition (at around 14 months), are unable to learn that phonological neighbours such as “bih” and “dih” refer to two distinct objects. Infants did not pay increased attention to trials in which the word-object association was switched for these two highly similar names – although they did show a novelty preference for trials that involved switching two phonologically distinct words such as “lif” and “neem”. This result is surprising since it is inconsistent with a finding first reported by Jusczyk and Aslin (1995) – and replicated by Stager and Werker (1997) – that even at 9 months, infants that are familiarised with word forms in connected speech can readily distinguish between minimal pairs like “bih” and “dih”.

At face value these results suggest that the processes involved in mapping speech to meaning do not have access to as detailed a representation of the speech stream as the system involved in learning word-forms. Such a result may be difficult to reconcile within the model presented here in which both mappings make use of the same internal representation of the speech input. One tempting conclusion might be to

assume that the systems involved in learning word forms and mapping to meaning operate on separate representations of the speech input. However, such a conclusion appears to condemn much of the research on early word-form learning as being irrelevant to vocabulary acquisition. It is unclear what function a system for representing word-forms might serve if it does not assist in acquiring the mapping from speech to meaning. Stager and Werker themselves suggest that ignoring the full detail of speech in mapping to meaning may somehow assist the infant in learning language. However, if this 'less-is-more' interpretation is to be convincing it must be backed up by detailed simulations that illustrate the advantage that can be gained by ignoring potentially informative detail in one learning situation but not in another. Further experimental investigations of the abilities of infants at different ages may be enlightening in this respect.

A further empirical challenge to the account that has been developed here focuses on the abilities of 6-month-olds who have yet to master the segmentation of words from connected speech. By the account that has been proposed here, it would not be expected that these infants could map from speech to meaning since they do not yet have fully formed representations of word-forms. However, results reported by Tincoff and Jusczyk (1999) demonstrate that 6-months-old show precocious knowledge of the meaning of two particular words – "mummy" and "daddy"[4] – indicated by increased looking time towards a video of the named parent (though not to images of an unfamiliar male or female). Although "mummy" and "daddy" are clearly exceptional words because of their extremely high frequency in infant-directed speech and the salience of parents in the lives of infants, this result clearly challenges any simple

account in which word forms can only be attached to meaning after they have been segmented from connected speech.

It remains to be seen whether the model presented here could account for the early acquisition of words like “mummy” that are of high frequency and salience to infants. If these results do reflect the operation of the same system that maps word-forms to meaning in older infants, then these precociously learnt words may provide a valuable insight into the functioning of the immature system. Further investigations of the infants responses to these words (e.g. sensitivity to noisy or mispronounced tokens) may be especially valuable.

At face value, therefore, the results reported by Stager and Werker (1997) and Tincoff and Jusczyk (1999) suggest some separation of the processes that allow word forms to be mapped to meaning and systems that are involved in discovering words in the speech stream. It is unclear at present whether the modelling framework presented here could account for these apparent dissociations. Further simulations to explore these seemingly contradictory aspects of the behavioural literature should be carried out.

In conclusion, the work that we have presented here provides a modelling framework within which to explore a variety of important issues in lexical segmentation and vocabulary acquisition. While these simulations fall short of capturing the scale and complexity of the acquisition problem faced by infants, the mechanisms proposed appear to be sufficiently general to merit further investigation. Further simulations that include the specific tasks and materials used in the developmental literature would be valuable.

Acknowledgements

This work was supported by EPSRC research studentship number 94700590 and by an MRC Programme Grant to William Marslen-Wilson and Lorraine Tyler. I would like to thank Gareth Gaskell, Gary Cottrell, Morten Christiansen, Tom Loucas, Billi Randall and Ingrid Johnsrude for comments and suggestions on this work and to thank Julian Pine and Philip Quinlan for useful feedback on a previous draft of this manuscript. All simulations were carried out using the Tlearn simulator developed by Jeff Elman of the Centre for Research in Lanuage, University of California, San Diego. Finally, I would like to thank my niece Isobel for providing such a concrete illustration of the wonderful abilities of infants during the first two years.

References

Allen, J., & Christiansen, M. H. (1996). Integrating multiple cues in word segmentation: a connectionist model using hints. In G. W. Cottrell (Ed.), Proceedings of the 18th Annual Cognitive Science Society Conference. Mahwah, NJ: LEA.

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

Baillargeon. R. (1995) Physical reasoning in infancy. In Gazzaniga, M. S. (Ed). The Cognitive Neurosciences. (pp.181-204). Cambridge, MA: MIT Press.

Baldwin, D. A. (1991) Infants' contribution to the achievement of joint reference. Child Development, 62, 875-890.

Baldwin, D. A. (1993) Infants' ability to consult the speaker for clues to word reference. Journal of Child Language, 20, 395-418.

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

Bloom, P. (1994). Possible names: the role of syntax-semantics mappings in the acquisition of nominals. Lingua, 92, 297-329.

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

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

Bullinaria, J.A. (2001). Simulating the evolution of modular neural systems. In Moore, J. D. & Stenning, K. Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society p.146-153.

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

Charles-Luce, J, & Luce, P. A. (1995) An examination of similarity neighbourhoods in young children's receptive vocabularies. Journal of Child Language, 22, 727-735.

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

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

Crick, F. H. C. (1989) The recent excitement about neural networks. Nature, 337, 129-132.

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

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

Davis, M. H., Gaskell, M. G., & Marslen-Wilson, W. D. (1997) Recognising embedded words in connected speech: context and competition. In J. Bullinaria, D.

Glasspool, & G. Houghton (Eds), Proceedings of the Fourth Neural Computation in Psychology Workshop. London: Springer-Verlag.

Davis, M.H., Marslen-Wilson, W. D. & Gaskell, M. G. (2002) Leading up the lexical garden-path: Segmentation and ambiguity in spoken word recognition. Journal of Experimental Psychology: Human Perception and Performance, 28, 218-244.

Dominey, P.F., & Ramus, F. (2000) Neural network processing of natural language: I. Sensitivity to serial, temporal and abstract structure of language in the infant. Language and Cognitive Processes, 15, 87-127

Eimas, P. D., Siqueland, E. R., Jusczyk, P. & Vigorito, J. (1971) Speech perception in early infancy, Science, 171, 304-306.

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

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

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

Fahlman, S. E., & Lebiere, C. (1990) The cascade correlation learning architecture. in D. Touretzky (Ed.), Advances in Neural Information Processing, 2, Morgan-Kaufman.

Fernald, A. (1985) Four month old infants prefer to listen to motherese. Infant Behaviour and Development, 8, 181-195.

Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the second year. Psychological Science, 9, 228-231.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. Journal of Child Language, 16, 477-501.

Fenson, L., Dale, P. S., Reznik, J. S., Bates, E., Thal, D. J., & Perthick, S. J. (1994) Variability in early communicative development. Monographs of the Society for Research in Child Development, 59, Number 242.

Gaskell, M. G., & Marslen-Wilson. (1997). Integrating form and meaning: a distributed model of speech perception. Language and Cognitive Processes, 12, 613-656.

Gaskell, M. G., & Marslen-Wilson, W. D. (1999). Ambiguity, competition and blending in spoken word recognition. Cognitive Science, 23, 439-462.

Gasser, M. (1994). Modularity in a connectionist model of morphology acquisition. Proceedings of the International Conference on Computational Linguistics, 15, 214-220.

Gleitman, L. R. (1994) Words, words, words. Philosophical Transactions of the Royal Society of London. Series B, 346, 71-77.

Goldowsky, B. N., & Newport, E. L. (1993). Modeling the effects of processing limitations on the acquisition of morphology: the less is more hypothesis. In E. V. Clark (Ed.), Proceedings of the 24th Annual Child Language Forum. Stanford, CA: CSLI.

Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., Gordon, L. (1987). The eyes have it: lexical and syntactic comprehension in a new paradigm. Journal of Child Language, 14, 23-45.

Harrington, J., Watson, G., & Cooper, M. (1989). Word boundary detection in broad class and phoneme strings. Computer Speech and Language, 3, 367-382.

Harris, Z. S. (1955). From phoneme to morpheme. Language, 31, 190-222.

Jusczyk, P. W. (1997) The discovery of spoken language. Cambridge, MA, MIT Press.

Jusczyk, P. W. (1999) How infants begin to extract words from speech. Trends in Cognitive Science, 3, 323-328.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. Cognitive Psychology, 29, 1-23.

Jusczyk, P. W., Cutler, A. & Redanz, N. (1993a) Preference for the predominant stress patterns of English words. Child Development, 64, 675-687.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y. & Jusczyk, A. M. (1993b) Infants' sensitivity to the sound patterns of native language words. Journal of Memory and Language, 32, 402-420.

Jusczyk, P. W., & Hohne, E. A. (1997) Infants' memory for spoken words. Science, 277, 1984-1985.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. Cognitive Psychology, 39, 159-207.

Korman, M. (1984) Adaptive aspects of maternal vocalizations in differing contexts at ten weeks. First Language, 5, 44-45.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N. & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. Science, 255, 606-608.

Lehiste, I. (1960). An acoustic-phonetic study of internal open juncture. Phonetica, 5(supplement), 5-54.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. Cognitive Psychology, 18, 1-86.

Marcus, G. F., Vijayan, S., Bandi Rao, S., Vishton, P.M. (1999) Rule learning by seven-month-old infants. Science, 283, 77-80.

Mareschal, D. (2000) Object knowledge in infancy: current controversies and approaches. Trends in Cognitive Science, 4, 408-416.

Marslen-Wilson, W.D. (1984). Function and processing in spoken word recognition: a tutorial review. In H. Bouma & D. G. Bouwhuis (Eds.), Attention and Performance X: Control of Language Processing. Hillsdale NJ: Erlbaum.

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

Maskara, A. & Noetzel, W. (1993) Sequence recognition with recurrent neural networks. Connection Science, 5, 139-152.

Monsell, S., & Hirsh, K. W. (1998). Competitor priming in spoken word recognition. Journal of Experimental Psychology: Learning, Memory and Cognition, 24, 1495-1520.

Norris, D. (1990). A dynamic-net model of human speech recognition. In G. T. M. Altmann (Ed.), Cognitive Models of Speech Processing . Cambridge, MA: MIT Press.

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

Oviatt, S. L. (1980) The emerging ability to comprehend language: An experimental approach. Child Development, 51, 97-106.

O'Reilly, R.C. (1996) Biologically plausible error-driven learning using local activation differences: the generalized recirculation algorithm. Neural Computation, 8, 895-938.

Page, M. (2000). Connectionist modelling in psychology: A localist manifesto. Behavioural and Brain Sciences, 23, 443-512.

Pinker, S. (1989) Learnability and cognition. Cambridge, MA: MIT Press.

Plunkett, K., Sinha, C., Møller, M. F., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. Connection Science, 4, 293-312.

Plaut, D. C., & Kello, C. T. (1998). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In B. MacWhinney (Ed.), The Emergence of Language. Mahweh, NJ: Erlbaum.

Quinn, P., Eimas, P., & Rosenkrantz, S. (1993). Evidence for representations of perceptually similar natural categories by 3- and 4-month-old infants. Perception, 22, 463-475.

Roy, D. K., & Pentland, A. P. (2002) Learning words from sights and sounds: A computational model. Cognitive Science, 26, 113-146.

Rueckl, J. G., Cave, K. R. & Kosslyn, S. M. (1989) Why are "what" and "where" processed by separate cortical visual systems? A computational investigation. Journal of Cognitive Neuroscience, 1, 171-186.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical language learning by 8 month olds. Science, 274, 1926-1928.

Schafer, G., & Plunkett, K. (1998) Rapid word-learning by 15 month olds under tightly-controlled conditions. Child Development, 69, 309-320.

Schafer, G., Plunkett, K. & Harris, P. L. (1999) What's in a name? Lexical knowledge drives infants' visual preferences in the absence of referent input. Developmental Science, 2, 187-194.

Servan-Schrieber, D., Cleeremans, A. & McClelland, J. L. (1991) Graded state machines: The representation of temporal contingencies in simple recurrent networks. Machine Learning, 7, 161-193.

Shastri, L., & Ajjanagadde, V. (1993) From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. Behavioural and Brain Sciences, 16, 417-494

Shrager, J., & Johnson, M. H. (1996) Dynamic plasticity influences the emergence of function in a simple cortical array. Neural Networks, 9, 1119-1129.

Sirois, S., Buckingham, D., & Shultz, T. R. (2000). Artificial grammar learning by infants: An auto-associator perspective. Developmental Science, 4, 442-456.

Siskind, J. M. (1996) A computational study of cross-situational techniques for learning word-to-meaning mappings. Cognition, 61, 39-91.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. Artificial Intelligence, 46, 159-216.

Sougné, J. (1998). Connectionism and the problem of multiple instantiation. Trends in Cognitive Science, 2, 183-189.

Stager, C. L. & Werker, J. F. (1997) Infants listen for more phonetic detailed in speech perception than in word-learning tasks. Nature, 388, 381-382.

St. John, M. F., & McClelland, J. L. (1990) Learning and applying contextual constraints in sentence comprehension. Artificial Intelligence, 46, 217-257.

Swingley, D., Pinto, J. P., & Fernald, A. (1999). Continuous processing in word recognition at 24 months. Cognition, 71, 73-108.

Thomas D. G., Campos J. J., Shucard D. W., Ramsay D. S., & Shucard J. (1981) Semantic comprehension in infancy: A signal detection analysis. Child Development, 52, 798-903.

Tincoff, R. & Jusczyk, P. W. (1999) Some beginnings of word comprehension in 6-month-olds. Psychological Science, 10, 172-175.

Walley, A. (1993) The role of vocabulary development in children's spoken word recognition and segmentation ability. Developmental Review, 13, 286-350.

Waxman, S. & Markow, D. (1996) Words as an invitation to form categories: Evidence from 12- to 13-month-olds. Cognitive Psychology, 29, 257-302.

Werker, J. R. & Tees (1984) Cross-language speech perception: Evidence for perceptual reorganisation during the first year of life. Infant Behaviour and Development, 7, 49-63.

Wolff, J. G. (1977). The discovery of segmentation in natural language. British Journal of Psychology, 68, 97-106.

Figure Captions

Figure 1: Simple recurrent network architecture used for Simulation 1, showing a snapshot of training activations during the segment /d/ in the sequence "*lid cap lock*". Throughout each training sequence, the target for the network is to activate a representation of all the words in the sequence, not just the current word. Solid arrows show trainable connections, the dotted arrow shows fixed one-to-one connections that store a copy of the hidden unit activations at the previous time step.

Figure 2: Lexical activation for target words and competitors during the sequence "lid cap lock" averaged over 10 fully trained networks from Simulation 1. Error bars show +/- 1 standard error.

Figure 3: Correct recognition (left axis) and recognition point (right axis) throughout training. Results averaged over ten networks from Simulation 1. Error bars show +/- 1 standard error.

Figure 4: Recognition point for items with cohort-competitors (sharing consonant and vowel – e.g. lick – lid) and items without cohort competitors (e.g. knit – knot) in Simulation 1. Results averaged over ten networks. Error bars show +/- 1 standard error.

Figure 5: Simple recurrent network architecture used for Simulation 2, showing a snapshot of training activations during the segment /d/ in the sequence "*lid cap lock*".

The network is identical to Simulation 1 (Figure 1) except for the addition of a output units trained to predict the input at the next time step.

Figure 6: (a) Correct recognition and (b) Recognition point for networks trained with (Simulation 2) and without (Simulation 1) the prediction task. Results averaged over ten networks in each simulation. Error bars show +/- 1 standard error

Figure 7: RMS output error for the both output tasks from Simulation 2 (Lexical task with prediction task and prediction task) compared to output error for Simulation 1 (Lexical task). Results averaged over ten networks. Error bars show +/- 1 standard error.

Figure 8: Magnitude of change (Euclidean distance) in hidden unit states within words and between words for networks trained without the prediction task (Simulation 1) and with a prediction task (Simulation 2). Results averaged over ten networks. Error bars show +/- 1 standard error.

Footnotes

[1] Not all words in speech are content words; infants must also learn the role played by articles, prepositions and other function words in speech. However, for the purposes of the current paper we will focus our attention solely on how infants learn their first words – typically concrete nouns.

[2] Our discussion of the use of metrical stress and phonotactic information focuses on cues that support word-boundary detection in English. One strength of the statistical mechanisms proposed here is that they may be sufficiently flexible to account for segmentation in other languages in which different phonotactic and metrical cues operate. For reasons of space, the current chapter will concentrate on lexical segmentation and vocabulary acquisition in English.

[3] This value provides a suitable threshold to ensure that only correct identifications are made. Results showed a similar pattern (when false-identifications were excluded) with a lower threshold.

[4] Infants were actually tested on the form of these two words that parents reported as being most frequently used around them. These are assumed to be "mummy" and "daddy" for clarity.

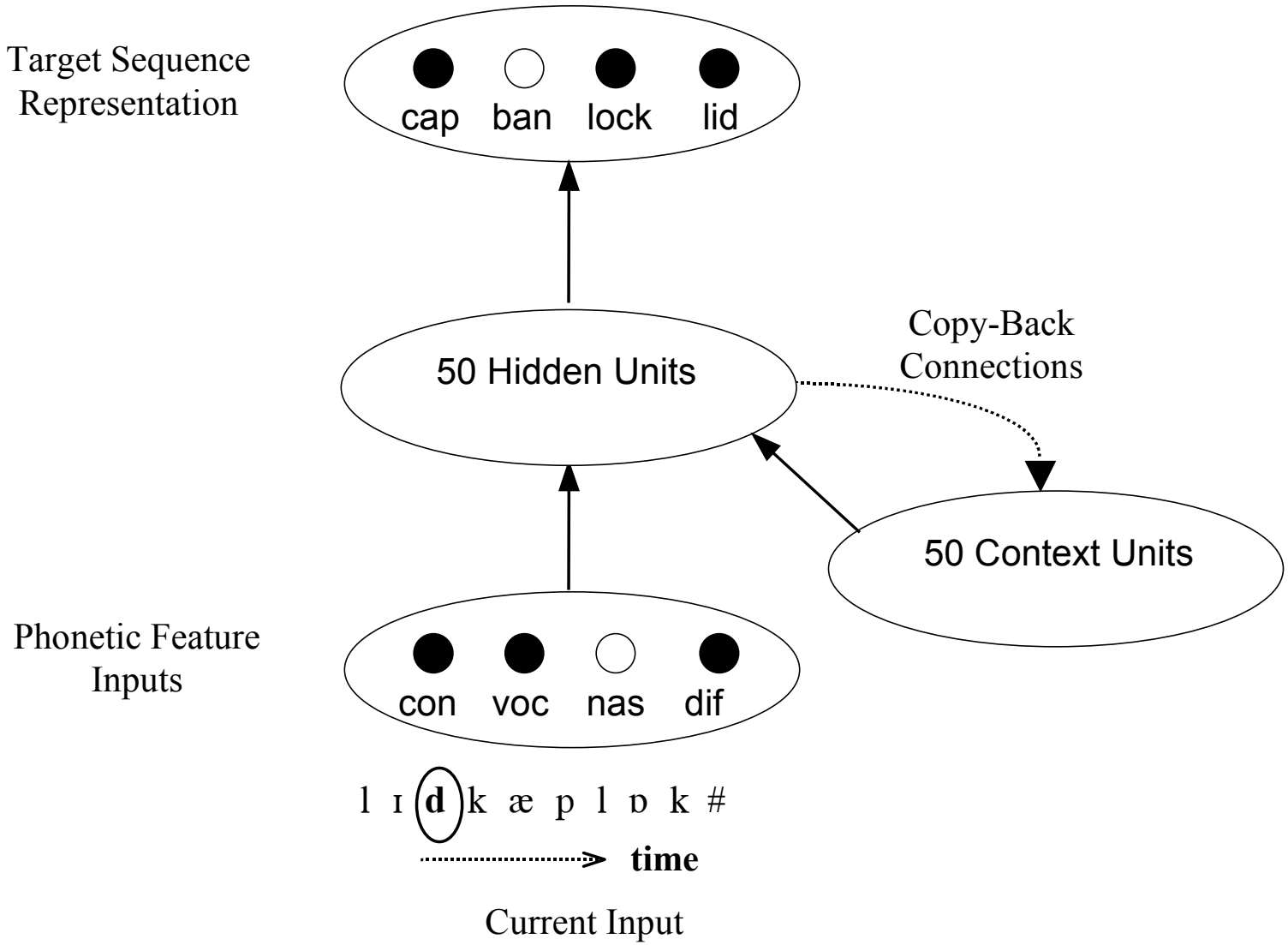
Target Sequence Representation
cap ban lock lid
50 Hidden Units
Copy-Back Connections
50 Context Units
Phonetic Feature Inputs
con voc nas dif
l ɪ d k æ p l ɒ k #
time
Current Input

Figure 1

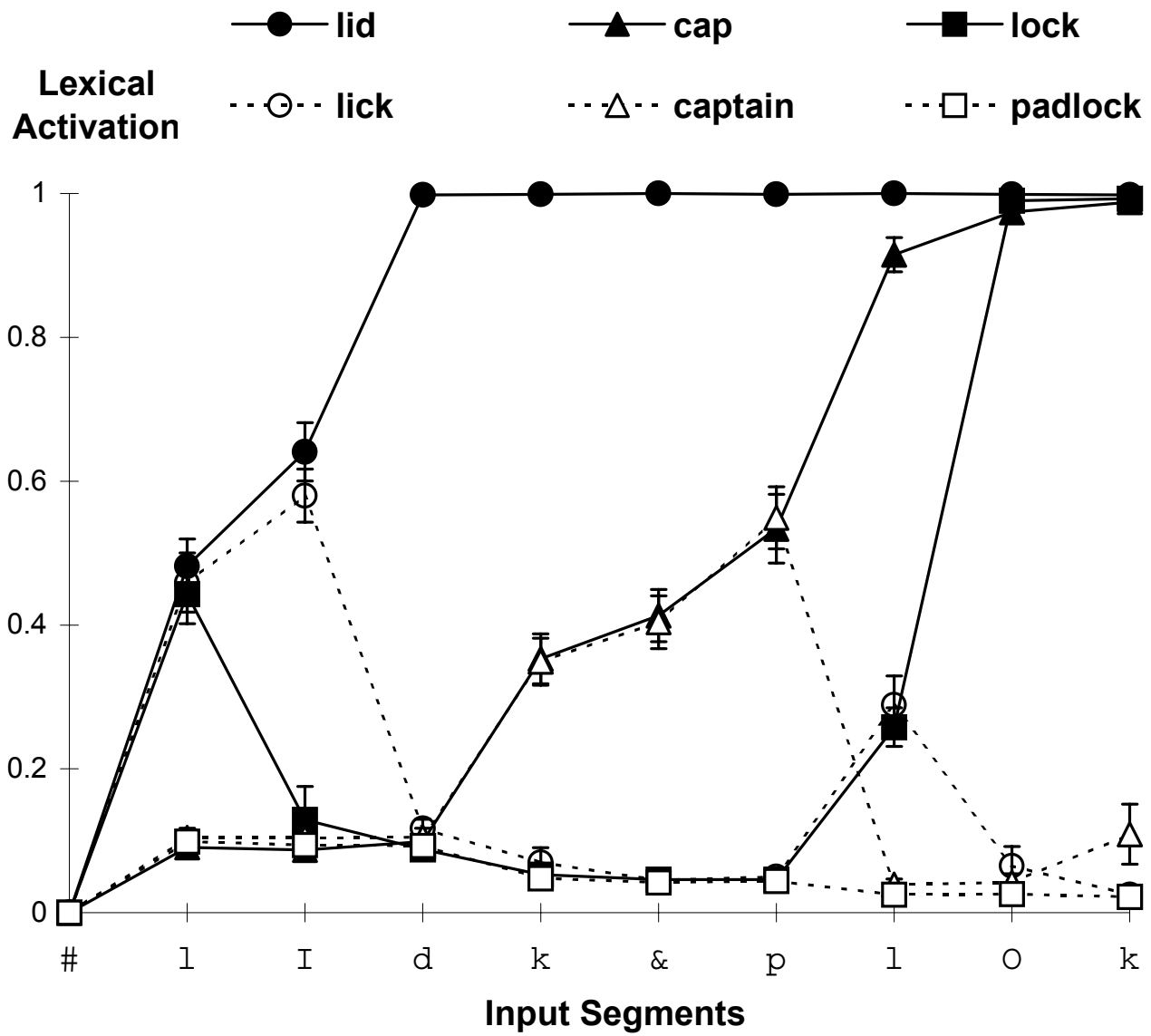
lid
cap
lock
lick
captain
padlock
Lexical Activation
1
0.8
0.6
0.4
0.2
0
l I d k & p l O k
Input Segments

Figure 2

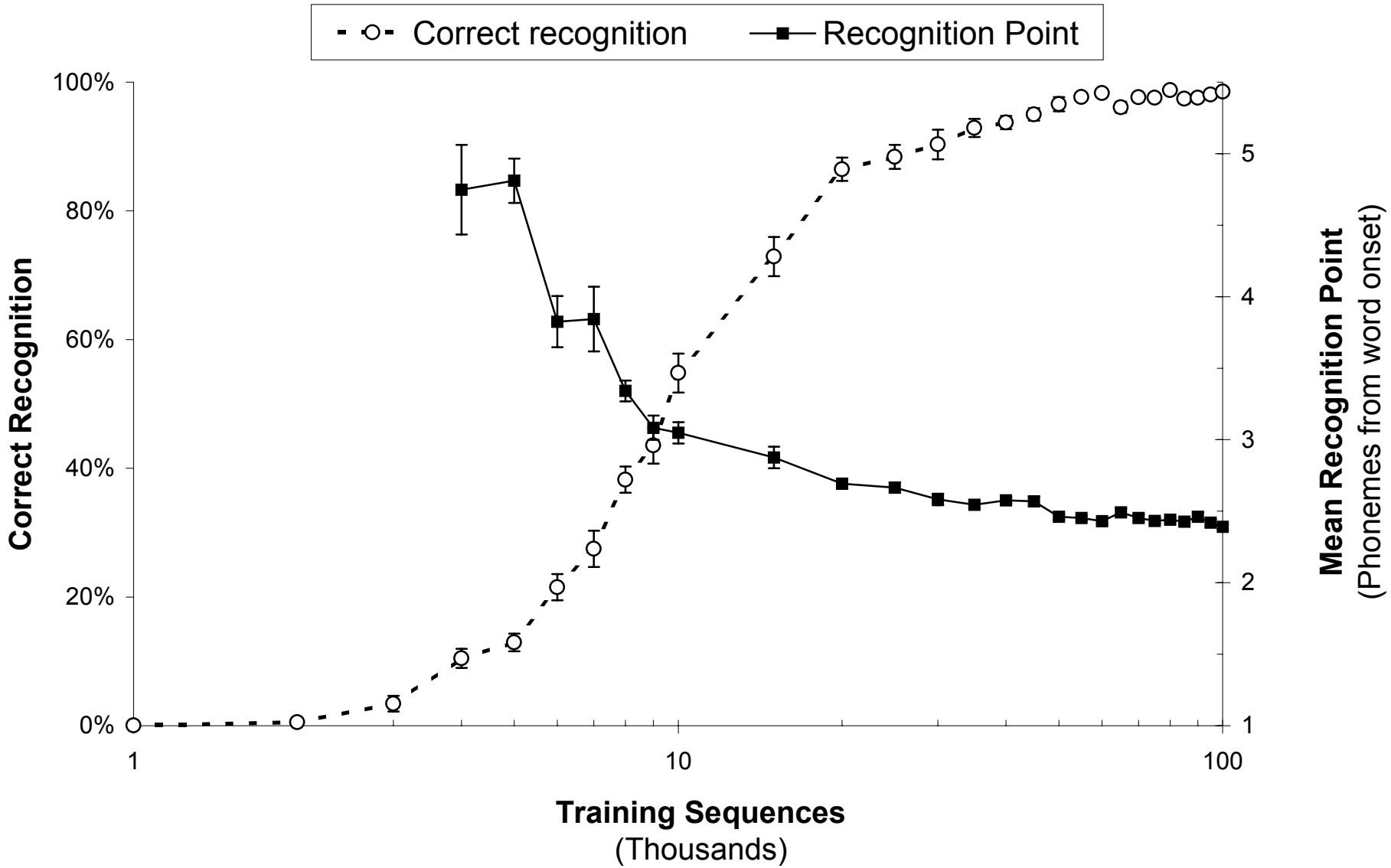
Correct recognition
Recognition Point
100%
80%
60%
40%
20%
0%
Correct Recognition
5
4
3
2
1
Mean Recognition Point
(Phonemes from word onset)
1
10
100
Training Sequences
(Thousands)

Figure 3

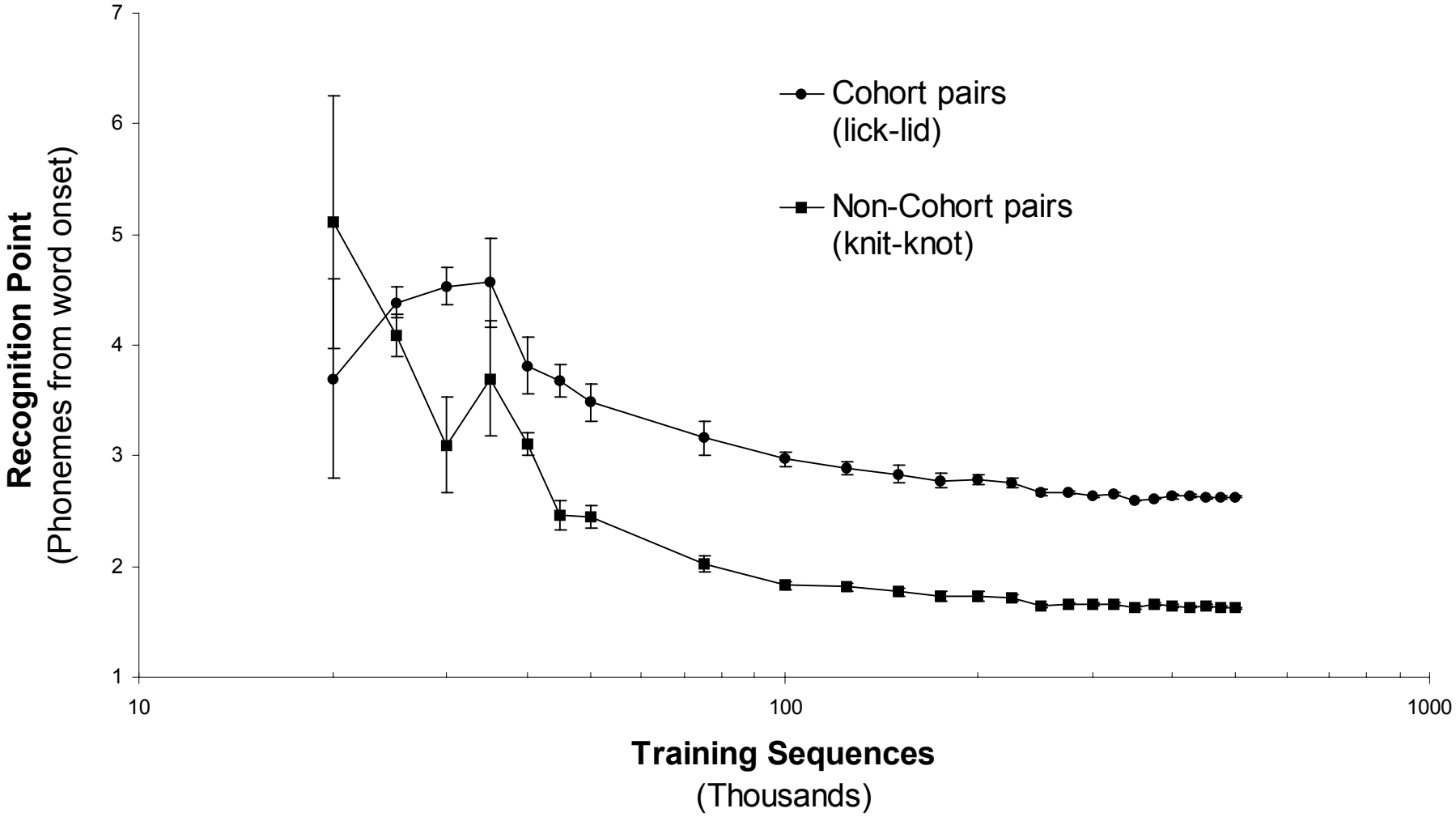
Recognition Point
(Phonemes from word onset)
7
6
5
4
3
2
1
10
100
1000
Training Sequences
(Thousands)
Cohort pairs
(lick-lid)
Non-Cohort pairs
(knit-knot)

Figure 4

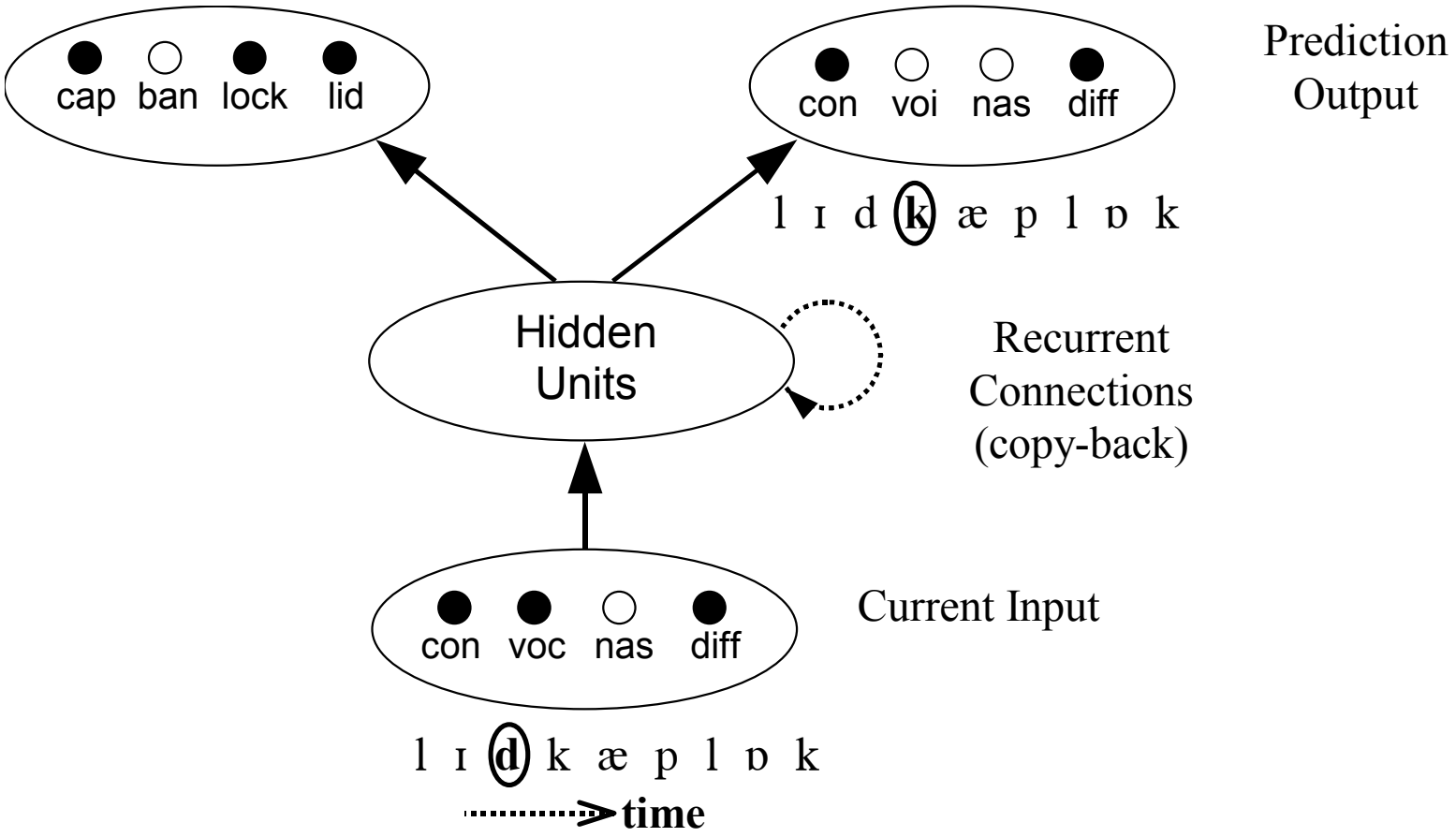
cap ban lock lid
con voi nas diff
Prediction
Output
l ɪ d k æ p l ɒ k
Hidden
Units
Recurrent
Connections
(copy-back)
con voc nas diff
Current Input
l ɪ d k æ p l ɒ k
time

Figure 5

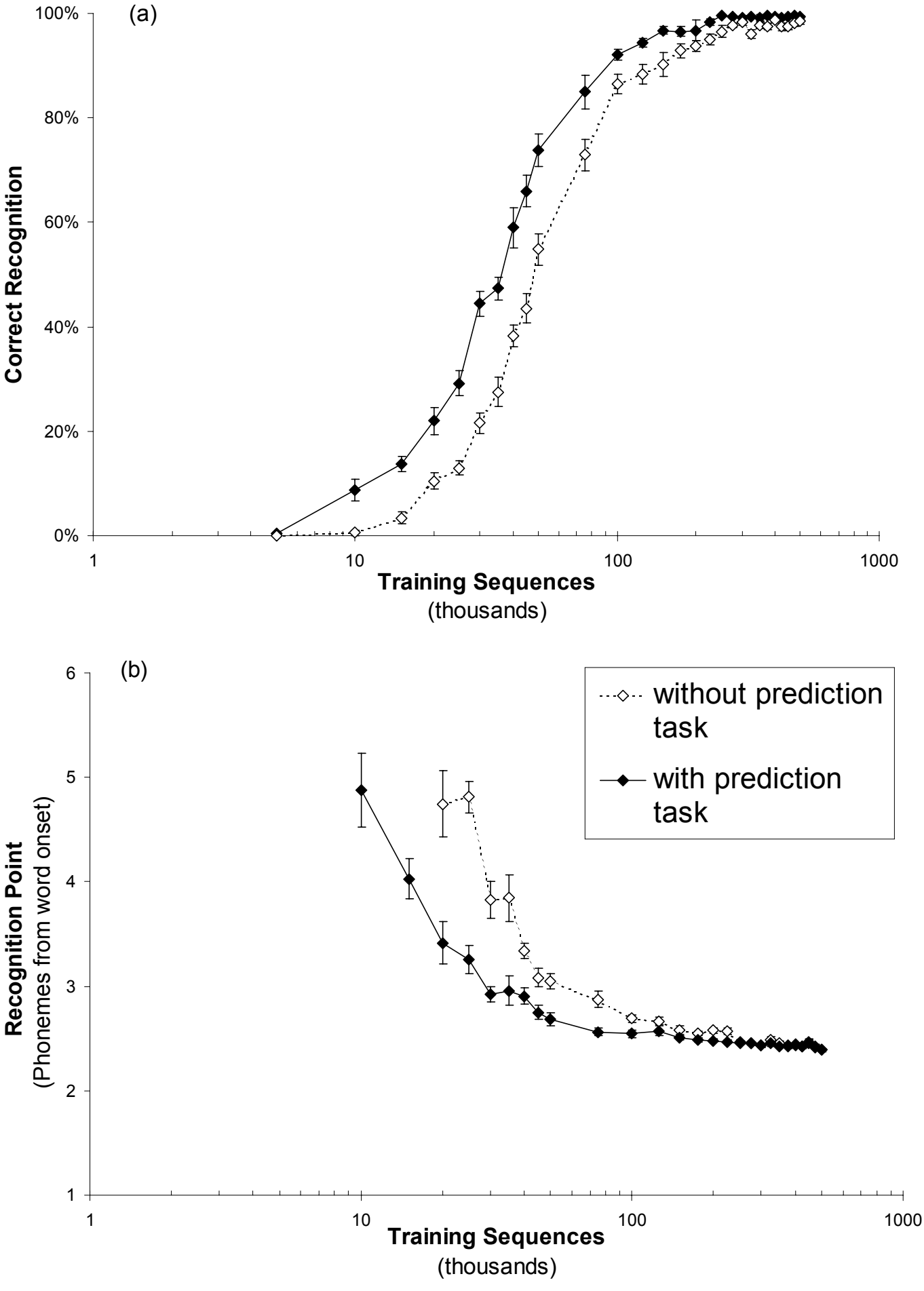
(a)
100%
80%
60%
40%
20%
0%
Correct Recognition
1
10
100
1000
Training Sequences
(thousands)
(b)
6
5
4
3
2
1
Recognition Point
(Phonemes from word onset)
without prediction task
with prediction task
1
10
100
1000
Training Sequences
(thousands)

Figure 6

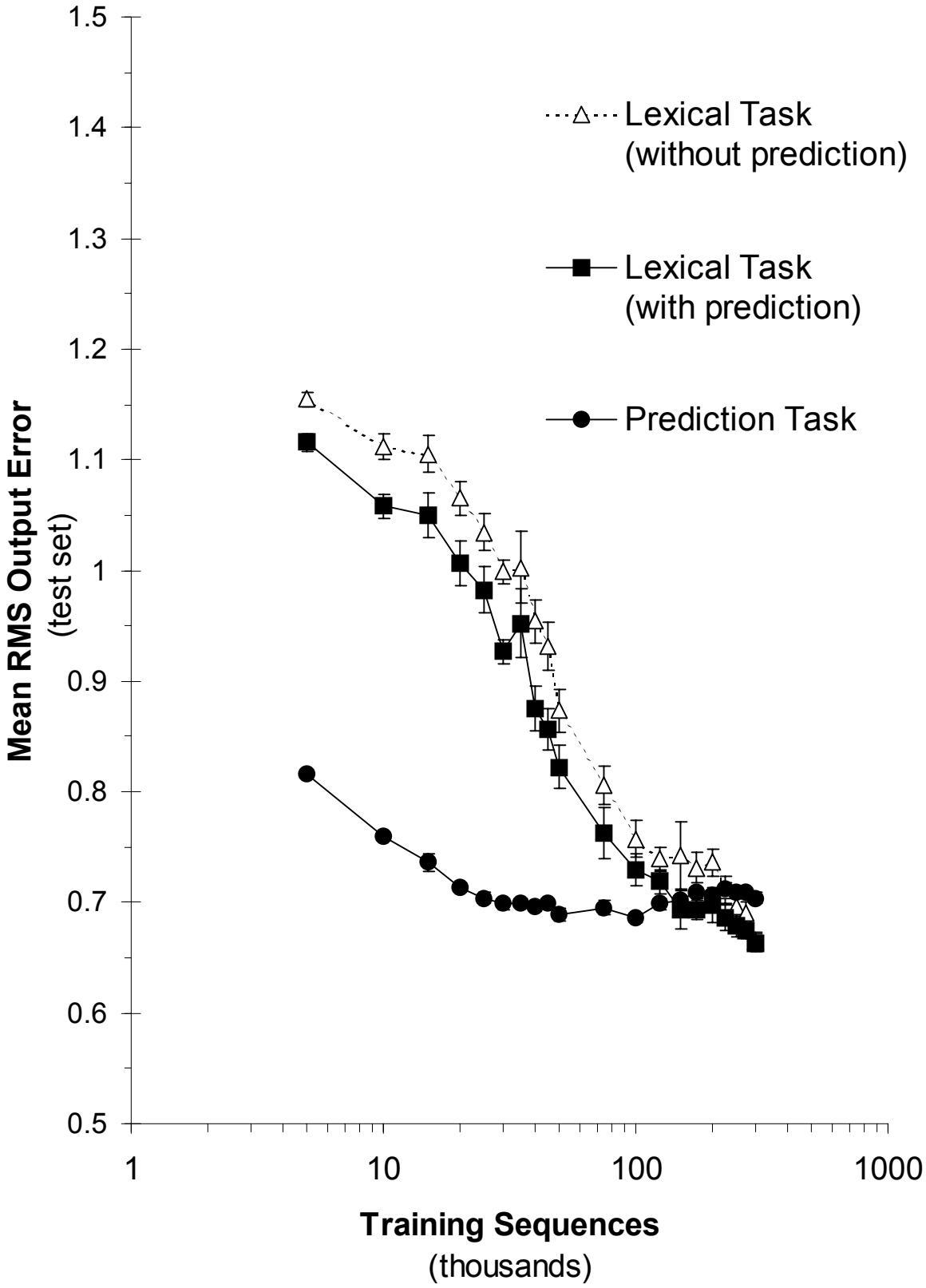
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
Mean RMS Output Error
(test set)
Lexical Task
(without prediction)
Lexical Task
(with prediction)
Prediction Task
1
10
100
1000
Training Sequences
(thousands)

Figure 7

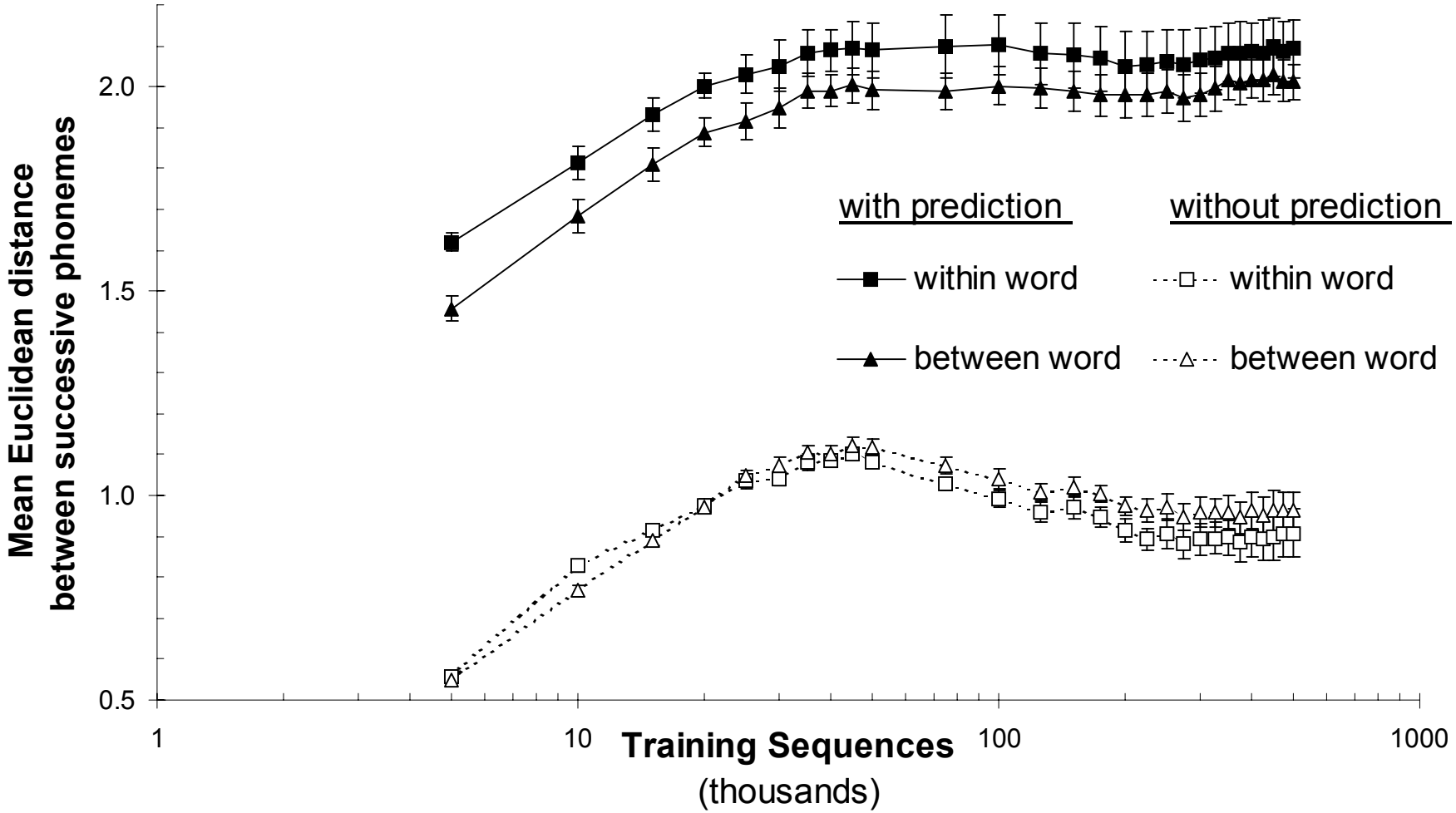
Mean Euclidean distance
between successive phonemes
2.0
1.5
1.0
0.5
1
10
100
1000
Training Sequences
(thousands)
with prediction
without prediction
within word
between word
within word
between word

Figure 8