LANGUAGE AND COGNITIVE PROCESSES, 1998, *13* (2/3), 221–268

Learning to Segment Speech Using Multiple Cues: A Connectionist Model

Morten H. Christiansen, Joseph Allen, and
Mark S. Seidenberg

Program in Neural, Informational, and Behavioral Sciences, University of Southern California, USA

Considerable research in language acquisition has addressed the extent to which basic aspects of linguistic structure might be identified on the basis of probabilistic cues in caregiver speech to children. This type of learning mechanism presents classic learnability issues: there are aspects of language for which the input is thought to provide no evidence, and the evidence that does exist tends to be unreliable. We address these issues in the context of the specific problem of learning to identify lexical units in speech. A simple recurrent network was trained on a phoneme prediction task. The model was explicitly provided with information about phonemes, relative lexical stress, and boundaries between utterances. Individually these sources of information provide relatively unreliable cues to word boundaries and no direct evidence about actual word boundaries. After training on a large corpus of child-directed speech, the model was able to use these cues to reliably identify word boundaries. The model shows that aspects of linguistic structure that are not overtly marked in the input can be derived by efficiently combining multiple probabilistic cues. Connectionist networks provide a plausible mechanism for acquiring, representing, and combining such probabilistic information.

Requests for reprints should be addressed to Morten H. Christiansen, Program in Neural, Informational, and Behavioral Sciences, University of Southern California, University Park MC-2520, Los Angeles, CA 90089-2520. USA.

We would like to thank Michael Brent, Gary Cottrell, James Morgan, and two anonymous reviewers for their comments on a previous version of this article. We are also grateful to Michael Brent for providing access to a computational implementation of the DR algorithm. The research presented here is partially supported by NIMH grant 47566 and an NIMH Research Scientist Development Award to MSS.

INTRODUCTION

In recent years there has been renewed interest in and new insights about the statistical properties of language and their possible role in acquisition and comprehension. There is particular interest in the idea that the child's entry into language—the initial identification of relevant phonological and lexical units and basic syntactic information such as grammatical categories—is driven by analyses of the statistical properties of input. Such properties include facts about the distributions of words in contexts and correlations among different types of linguistic information. These properties of language have been largely excluded from investigations of grammatical competence and language acquisition since Chomsky (1957).

Several factors have conspired to make this new era of statistical research more than merely a revival of classical structural linguistics. First, there have been new discoveries concerning the statistical structure of language (e.g. Aijmer & Altenberg, 1991) that have led to some impressive results in applied areas such as automatic speech recognition and text tagging (Brill, 1993; Church, 1987; Marcus, 1992). Second, there have been important discoveries concerning aspects of the input to the child that may provide reliable cues to linguistic structure (Echols & Newport, 1992; Jusczyk, 1993; Morgan, 1986). Third, there has been the development of connectionist learning procedures suitable for acquiring and representing such information efficiently (Rumelhart & McClelland, 1986). These procedures are considerably more powerful than the behaviourist learning rules available to the structural linguists of the 1950s, and, more interestingly, they are coupled with a theory of knowledge representation that permits the development of abstract, underlying structures (Elman, 1991; Hinton, McClelland, & Rumelhart, 1986). The considerable progress in these related areas provides a strong motivation for reconsidering questions about language learning that many linguists assume were settled long ago.

The theory that statistical properties of language play a central role in acquisition faces two classic learnability problems. First, there is the question as to how children might learn specific aspects of linguistic structure for which there is no direct evidence. One of the central claims of modern theoretical linguistics is that languages exhibit properties that must be known innately to the child because experience provides no evidence for them (Crain, 1991). A second observation about language acquisition is that the input affords unlimited opportunities to make reasonable but false generalisations, yet children rapidly converge on the grammars of the languages to which they are exposed without seeming to pursue these many alternatives (Hornstein & Lightfoot, 1981). Thus, the input is said to provide both too little evidence concerning properties of the target language and too much evidence consistent with irrelevant analyses. Innate forms of linguistic knowledge and constraints on learning are seen as providing the solutions to these learnability puzzles, which are thought to have established strong limitations on the role of experience in language acquisition. Both of these issues are relevant to approaches to acquisition that rely on statistical properties of the input. With regard to the first claim, it must be determined whether the picture concerning "linguistic structures for which there is no evidence" changes when we consider statistical properties of language that have previously been ignored. With regard to the second claim, languages can be statistically analysed in innumerable ways and therefore the problem as to how the child could know which aspects to attend to is a serious one. There is a further problem insofar as statistical properties of language provide cues to linguistic structure that are probabilistic at best. How such partial, unreliable cues to linguistic structure could facilitate language learning is unclear.

In this article, we explore systems that are capable of learning and representing statistical properties of language such as the constellations of overlapping, partially predictive cues increasingly implicated in research on language development (e.g. Morgan, & Demuth, 1996). Such cues tend to be probabilistic and violable, rather than categorical or rule-governed. Importantly, these systems incorporate mechanisms for combining different sources of information, including cues that may not be highly constraining when considered in isolation. We explore the idea that conjunctions of these cues provide evidence about aspects of linguistic structure that is not available from any single source of information, and that this process of integration reduces the potential for making false generalisations. Thus, the general answer we adopt to both of the classical learnability questions is that there are mechanisms for efficiently combining cues of even very low validity, that such combinations of cues are the source of evidence about aspects of linguistic structure that would be opaque to a system insensitive to such combinations, and that these mechanisms are used by children acquiring languages (for a similar view, see Bates & MacWhinney, 1987). These mechanisms also play a role in skilled language comprehension and are the focus of so-called constraint based theories of processing (MacDonald, Pearlmutter & Seidenberg, 1994; Trueswell & Tanenhaus, 1994) that emphasise the use of probabilistic sources of information in the service of computing linguistic representations. Since the learners of a language grow up to use it, investigating these mechanisms provides a link between language learning and language processing. In the remainder of this article we explore these ideas as they apply to the problem of segmenting utterances into words. Although we concentrate here on the relevance of combinatorial information to this specific aspect of acquisition, our view is that similar mechanisms are likely to be relevant to other aspects of acquisition and to skilled performance.

Derived Linguistic Structure: A Theoretical Framework

In the standard learnability approach, language acquisition is viewed in terms of the task of acquiring a grammar. We propose an alternative view in which language acquisition can be seen as involving several simultaneous tasks. The *primary task*—the language learner's goal—is to comprehend the utterances to which she is exposed for the purpose of achieving specific outcomes. In the service of this goal the child attends to the linguistic input, picking up different kinds of information, subject to perceptual and attentional constraints. There is a growing body of evidence that as a result of attending to sequential stimuli, both adults and children incidentally encode statistically salient regularities of the signal (e.g. Cleeremans 1993; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996). The child's *immediate task*, then, is to update its representation of these statistical aspects of language. Our claim is that knowledge of other, more covert aspects of language is derived as a result of how these representations are combined. Linguistically relevant units (e.g., words, phrases, and clauses) emerge from statistical computations over the regularities induced via the immediate task. On this view, the acquisition of knowledge about linguistic structures that are not explicitly marked in the speech signal—on the basis of information, that is—can be seen as a third *derived task*. In the research described later, the immediate task is to encode statistical regularities concerning phonology, lexical stress, and utterance boundaries. The derived task is to integrate these regularities in order to identify the boundaries between words in speech.

The Segmentation Problem

Comprehending a spoken utterance requires segmenting the speech signal into words. Discovering the locations of word boundaries is a nontrivial problem because of the lack of a direct marking of word boundaries in the acoustic signal the way that white spaces mark boundaries on a page. The segmentation problem provides an appropriate domain for assessing our approach insofar as there are many cues to word boundaries, including prosodic and distributional information, none of which is sufficient for solving the task alone.

Early models of spoken language processing assumed that word segmentation occurs as a by-product of lexical identification (e.g. Cole & Jakimik, 1978; Marslen-Wilson & Welsh, 1978). More recent accounts hold that adults use segmentation procedures in addition to lexical knowledge (Cutler, 1996). These procedures are likely to differ across languages, and presumably include a variety of sublexical skills. For example, it is well known that adults are sensitive to phonotactic information, and make consistent judgments about whether a sound string is a "possible" native word (Greenburg & Jenkins, 1964). This type of knowledge could assist in adult segmentation procedures (Jusczyk, 1993). Cutler (1994) presents evidence from perceptual studies suggesting that adults know about and utilise language specific rhythmic segmentation procedures in processing utterances.

It seems reasonable to assume that children are not born with the knowledge sources that appear to subserve segmentation processes in adults. They have neither a lexicon nor knowledge of the phonological or rhythmic regularities underlying the words in the language being learned. The important developmental question concerns how the child comes to achieve steady-state adult behaviour. Intuition suggests that children might begin to add to their lexicon by hearing words in isolation. A single word strategy whereby children adopted entire utterances as lexical candidates would seem to be viable very early in acquisition. In the Bernstein–Ratner corpus (1987) and the Korman corpus (1984), 22–30% of child-directed utterances are made up of single words. However, many words will never occur in isolation. Moreover, this strategy on its own is hopelessly underpowered in the face of the increasing size of utterances directed toward infants as they develop. Instead, the child must develop viable strategies that will allow him or her to detect utterance internal word boundaries regardless of whether or not the words appear in isolation. A better suggestion is that a bottom-up process exploiting sublexical units allows the child to bootstrap the segmentation process (Morgan & Demuth, 1996). This bottom-up mechanism must be flexible enough to function despite cross-linguistic variation in the constellation of cues relevant for the word segmentation task.

Cooper and Paccia-Cooper (1980) and Gleitman, Gleitman, Landau, and Wanner (1988) proposed the existence of strategies based on prosodic cues (including pauses, segmental lengthening, metrical patterns, and intonation contour), which they held to be likely cross-linguistic signals to the presence of word boundaries. More recent proposals concerning how infants detect lexical boundaries have focused on statistical properties of the target language that may be exploited in early segmentation. Two of the cues to segmentation we utilise in our model (sequential phonological regularities and lexical stress) have both received considerable attention in recent investigations of language development.

Cutler and her colleagues (e.g. Cutler & Mehler, 1993) have emphasised the potential importance of rhythmic strategies to segmentation. They have suggested that skewed stress patterns (e.g. the majority of words in English have strong initial syllables) play a central role in allowing children to identify likely boundaries. Evidence from speech production and perception studies with prelinguistic infants supports the claim that infants are sensitive to rhythmic structure and its relationship to lexical segmentation by nine months (Jusczyk, Cutler, & Redanz, 1993). A second potentially relevant source of information which could be useful in deriving the locations of boundaries is the phonological regularities in the language being learned. A recent study by Jusczyk, Friederici, and Svenkerud (1993) suggests that infants develop knowledge of phonotactic regularities in their language between six and nine months. Furthermore, there is evidence that both children and adults are sensitive to and can utilise such information to segment the speech stream. Saffran, Newport, and Aslin (1996) show that adults are able to use phonotactic sequencing to determine possible and impossible words in an artificial language after only 20 minutes of exposure. They suggest that learners may be computing the transitional probabilities between sounds in the input and using the strengths of these probabilities to hypothesise possible word boundaries. Similarly, there is now evidence that infants as young as eight months show the same type of sensitivity (Saffran, Aslin, & Newport, 1996). Thus, children appear to be sensitive to the statistical regularities of potentially relevant sublexical properties of their languages such as stress and phonotactics, consistent with the hypothesis that these cues could play a role in bootstrapping segmentation.

The remainder of this article is organised as follows. First, we discuss prior computational work on word segmentation, including our previous work on the integration of two cues, phonology and utterance boundary information, in an artificial language learning task (Allen & Christiansen, 1996). The penultimate section presents the results of our new simulations in which we use a corpus of child-directed speech as well as an additional cue encoding relative lexical stress, and comparisons are made with other approaches. Finally, in the General Discussion, we discuss implications of the simulation results for theories of language acquisition.

COMPUTATIONAL APPROACHES TO WORD SEGMENTATION

There have been several attempts to develop computational approaches to the related problems of segmenting and recognising words in the speech stream. Most attention has focused on the identification of isolated words, using models that already possess knowledge of lexical items. For example, the influential TRACE model of speech perception (McClelland & Elman, 1986), was an interactive activation model with layers of units corresponding to phonetic features, phonemes, and words. These layers were interconnected such that excitatory activation could flow between layers and inhibitory activation within layers. This model was successful in accounting for a variety of speech perception data and led to predictions about coarticulation effects which were subsequently confirmed experimentally

(Elman & McClelland, 1988). Theoretically, the force of the model was to suggest that a combination of top-down lexical feedback and bottom-up phonetic information was necessary to account for human performance. Later models were proposed in which the flow of activation is entirely bottom-up (e.g. Norris, 1993, 1994; Shillcock, Lindsey, Levy, & Chater, 1992). Both the TRACE model and the bottom-up models were intended as models of adult word recognition and not of developmental word segmentation. Both include lexical information that is not available to an infant embarking on language acquisition.[1]

Other connectionist models have addressed the issue of learning to segment the speech stream. Elman (1990) trained a Simple Recurrent Network (SRN) on a small artificial corpus (1270 words tokens/15 types) with no explicit indication of word boundaries. After training, the error for each item in the prediction task was plotted, revealing that error was generally high at the onset of words but decreased as more of the word was processed. Elman suggested that sequential information in the signal could thus serve as a cue to word boundaries, with peaks in the error landscape indicating the onset of words. In a similar vein, Cairns, Shillcock, Chater, and Levy (1994) also considered peaks in the error score as indications of word boundaries (with "peak" defined in terms of a cut-off point placed varying numbers of standard deviations above the mean). Their model, a recurrent network trained on a corpus of conversational speech using back-propagation through time (Rumelhart, Hinton, & Williams, 1986), was able to predict word boundaries above chance. Most recently, Aslin, Woodward, LaMendola, and Bever (1996) trained a feed-forward network on small corpora of child-directed speech using triplets of phonetic feature bundles to capture the temporal structure of speech. An additional unit was activated at utterance boundaries. The output layer consisted of a single unit representing the existence of an utterance boundary. This representational scheme allowed the network to acquire knowledge of utterance final phonological patterns which with some success could then be used to identify utterance internal word boundaries.

Arguably, the most successful computational demonstration of word segmentation is found in the work of Brent and colleagues (Brent, 1996; Brent & Cartwright, 1996; Brent, Gafos, & Cartwright, 1994). They employ a statistical algorithm based on the Minimal Description Length principle (Rissanen, 1983). This algorithm determines an optimised characterisation of a corpus by calculating the minimal vocabulary necessary for describing the input. This procedure is used to build a lexicon from scratch and to

[1] The model by Shillcock et al. (1992) did not include a lexicon but the focus of this work was to simulate the putative top-down lexical effects of Elman and McClelland (1988)—in particular, the compensation for coarticulation—rather than word segmentation.

segment the speech stream. The success of this Distributional Regularity (DR) algorithm on a corpus of child-directed speech is increased significantly when the description procedure is constrained by built-in knowledge of legal word-initial and word-final consonant clusters as well as the requirement that all words must contain a vowel. The DR model is an abstract description of a system sensitive to statistical regularities that may be relevant to segmentation, but the current implementation abstracts away from issues of psychological plausibility. For example, knowledge of additional constraints such as boundary phonotactics are currently independent "add-ons" to the basic algorithm. This results in a model in which knowledge of phonotactics is in place before any segmentation takes place. A more natural solution would allow the acquisition of phonotactics to proceed hand in hand with the development of the segmentation process.

The focus of the present work is on the integration of cues, and how such integration can facilitate the discovery of derived linguistic knowledge. Our aim is a psychologically plausible mechanism which, unlike the model of Brent and colleagues, incorporates the *simultaneous* learning and integration of cues. In contrast to earlier connectionist work on the segmentation of the speech stream, we also seek the integration of *multiple* cues (covering both distributional and acoustic cues). Finally, we approach the problem of word segmentation (and the problem of language acquisition in general) through our theoretical notion of immediate versus derived tasks. We now outline our preliminary work within this framework, before turning to the new simulations.

A Simplified Model

Allen and Christiansen (1996) conducted a series of simulations that demonstrated how distributional information reflecting sequential phonological regularities in a language may interact with information regarding the ends of utterances to inform the word segmentation task in language acquisition. They compared the performance of two SRN models by varying the information available to them. Incorporating the observation by Saffran, Newport, and Aslin (1996) that adults are capable of acquiring sequential information about syllable combinations in an artificial language, Allen and Christiansen trained the first network on a set of 15 trisyllabic (CVCVCV) words—the "vtp" (variable transitional probabilities) vocabulary—in which the word internal transitional probabilities were varied so as to serve as a potential source of information about lexical boundaries. In this vocabulary some syllables occurred in more words, and in more locations within words, than others. A second network was trained on a "flat" vocabulary made up of 12 words with no "peaks" in the word internal syllabic probability distribution; that is, the probability of a given syllable following any other syllable was the same for all syllables, and each syllable was equally likely to appear in any position within a word. Training corpora were created by randomly concatenating 120 instances of each of the words in a particular vocabulary set into utterances ranging between two and six words. Although word boundaries were not marked, a symbol marking the utterance boundary was added to the end of each utterance. (For details see Allen & Christiansen, 1996.)

Figure 1 shows the SRN employed in the simulations. This network is essentially a standard feed forward network with an extra component (the context units) allowing it to process temporal sequences. Originally developed by Elman (1988), the SRN provides a powerful tool with which to model the learning of many aspects of language ranging from speech processing (Norris, 1993; Shillcock, Lindsey, Levy, & Chater, 1992) to the modelling of a mapping from meaning to sound (Cottrell & Plunkett, 1991) to syntactic structure (Christiansen, in preparation; Christiansen & Chater, 1994; Elman, 1991, 1993). Fairly extensive studies have also been conducted on their computational properties (Chater, 1989; Chater & Conkey, 1992; Christiansen & Chater, in press; Cottrell & Tsung, 1993; Maskara & Noetzel, 1993; Servan-Schreiber, Cleeremans, & McClelland, 1989, 1991).

The SRN is typically trained on a prediction task in which the net has to predict the next item in a sequence. The SRNs used by Allen and Christiansen (1996) and in the new simulations reported here were trained to predict the next phoneme in a sequence. Consider, for example, as input

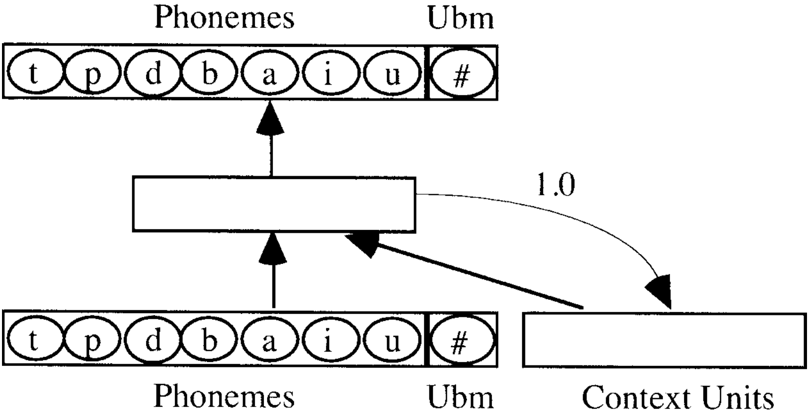

FIG. 1. Illustration of the SRN used in the simulations of Allen and Christiansen (1996). The network consisted of eight input/output units and thirty units in the hidden and context layers. Thick arrows indicate trainable weights, whereas the thin arrow denotes the copy-back weights (which are always 1).

the word /k&t/ (*cat*) (a key to the phonological transcription is found in Appendix A). At time *t* the unit representing /k/ is activated and activation propagates forward to the output units. Only /&/ is meant to be active on the output, and an error is calculated with respect to how much the network's output deviated from this target. This error is then used to adjust the weights according to the back-propagation learning rule (Rumelhart et al., 1986). The activation of the hidden unit layer is then copied to the context layer. At the next time step, $t + 1$, /&/ is activated on the input and activation propagates forward from this unit as well as from the context units. This time the net is expected to predict a /t/. This cycle is repeated for the whole training set. Following the final phoneme of an utterance, the target is an utterance boundary marker, corresponding to the silences between utterances. As a result of training, the boundary unit is more likely to be activated following phoneme sequences that tend to precede an utterance boundary. Since these sequences will also precede the ends of words within an utterance, the net is expected also to activate the utterance boundary unit at the ends of words not occurring at utterance boundaries.

By varying the syllabic probability distribution within words in the vtp vocabulary, Allen and Christiansen (1996) changed the likelihood that an utterance boundary marker will follow any particular sequence. In other words, a syllable that appears with higher probability at the ends of words than at other positions is more likely to appear prior to an utterance boundary than a syllable that occurs with equal probability at any position within the word. Similarly, a syllable that only appears at the beginning of words is unlikely to be followed by an utterance boundary. It was expected that the network trained on the vtp vocabulary would exploit this difference in determining the likelihood of an utterance boundary after any phoneme sequence. In the flat vocabulary, on the other hand, all syllables are equally likely to be followed by any other syllable, and equally likely to appear in any position in the word. Since no syllable is more likely than another to appear prior to an utterance boundary, the syllabic probability distribution cannot serve as an information source for the boundary prediction task.

In Allen and Christiansen (1996), full training sets were presented seven times to each network (a total of 78,771 tokens). Results showed that the network trained on the vtp vocabulary predicted a boundary with significantly higher confidence at lexical boundaries than at word internal positions. The network trained on the flat vocabulary, on the other hand, demonstrated almost no discrimination between end-of-word and non-end-of-word positions. Predictions about boundary locations were uniform across syllables for the latter net, and never reached the level of activation achieved by the net trained on the vtp vocabulary. The net trained on both utterance boundary locations and the vtp vocabulary learned to differentiate ends of words from other parts of words, whereas the network trained on boundary markers and the flat vocabulary failed to do so. Thus, variation in transitional probabilities between syllables increased accuracy on the segmentation task by lowering the probability that a boundary will appear after some syllables and raising the probability that one will appear after other syllables.

The model shows how sensitivity to variability in transitional probabilities between syllables, a statistical property of language, may allow learners to identify probable points at which to posit word boundaries. This idea differs from the suggestion by Saffran, Newport, and Aslin (1996) that people are sensitive to the differences between the transitional probabilities that exist within words (which are usually higher) and those that exist between words (which are lower). The network of Allen and Christiansen (1996) suggests that differences in transitional probabilities within words may provide an additional source of information, one that can interact with information about the locations of utterance boundaries to inform the segmentation process.

Allen and Christiansen (1996) also tested the extent to which utterance boundary information may facilitate the learning of the phoneme prediction task. A first set of simulations involved a comparison between two nets trained on the vtp vocabulary with and without utterance boundary information. No difference was found between the two training conditions, presumably due to a ceiling effect brought about by the "strong" sequential regularities of the vtp vocabulary. In a second set of simulations, two networks were similarly trained on the flat corpus. When tested on the 12 words in the flat vocabulary set (with no word or utterance boundaries marked), the network trained with boundary markers had a significantly lower error on the phoneme prediction task than the net trained without boundary markers. The presence of boundary markers in the input significantly altered the outcome of learning, such that the net trained with boundary markers was better able to learn the sequential regularities which were present in the corpus. This result relied on the fact that successfully predicting boundaries required the network to rely on longer sequences than a network required only to predict the next phoneme. The SRN architecture allows the network to discover the particular distributional window by which it can perform the entire task optimally.

These results suggest that interaction between information sources influences learning in significant ways. The comparison between nets trained on the vtp and the flat vocabulary shows how knowledge required for the immediate task of predicting the next phoneme in a sequence can be used in the derived task of predicting word boundaries (as suggested in the Introduction). The comparison between the nets trained with and without utterance boundary information shows that because the two tasks are learned together, the presence of the boundary prediction task alters the solution applied to the phoneme prediction task. Next, we shall see how the same principles apply to an SRN trained on a corpus of child-directed speech.

LEARNING TO SEGMENT CHILD-DIRECTED SPEECH

The simplified model of Allen and Christiansen (1996) outlined in the previous section abstracted away from a number of important aspects of natural language. For example, all words were of the same length (trisyllabic) and had the same syllable structure (CV). In addition, the vocabulary contained only 15 words. A pressing question is therefore whether the approach of Allen and Christiansen will scale up when faced with input that has many of the same characteristics as the language infants are exposed to. In order to address this question, we exposed a scaled-up model to a phonetically transcribed corpus of child-directed speech.

Method

In addition to the phonological and utterance boundary information used as input in the previous simulations, we added information about the relative stress patterns of the input words. The latter was included as a possible extra cue for the net to use in the derived task of word segmentation. Like the notion of probable boundary location, the notion of stress itself is also an emergent one. The multiple physical and phonetic cues for what is perceived as stress include chest pulses, amplitude, duration, and pitch contours, many of which play an independent phonemic role in numerous languages. In addition to being realised at more than one level of strength, English manifests stress at the level of the word, the intonational phrase, and possibly at other levels as well. In our simulations we chose only to encode a citation form of lexical stress[2] manifested on syllables, following a long tradition of viewing the stress-bearing unit in stress languages as the syllable (Jacobsen, 1962).

In speech between adults, monosyllabic content words are in general differentiated from monosyllabic function words in terms of lexical stress. Vowel segments in the latter tend to be reduced as is typical for weakly stressed syllables in English. However, it is not clear that the same stress differentiation exists in speech directed at young children. Bernstein-Ratner

[2] In our simulations, a particular word will always receive the same stress pattern independently of the context it occurs in. This is, of course, an idealisation as it is well known that stress may shift left in English depending on context (*thirte'en* → *thi'rteen me'n*) in the service of avoiding stress clash. Nevertheless, it is a reasonable idealisation because most of the bisyllabic words in our corpus are stress initial, and thus not subject to leftward stress shift.

(1987) conducted an analysis of speech to infants and found that vowel reduction in English function words is far less frequent in child-directed speech than in speech between adults. Further evidence presented by Morgan, Shi and Allopenna (1996) suggests that vowel quality taken alone is statistically a very poor cue to discovering function words (although it may become valuable when combined with information about vowel duration). Thus, early English motherese may not involve a significant differentiation of function words from content words—at least not at the level of stress that we are concerned with. In our simulations, we therefore represent all monosyllabic words as having primary stress.

Input

We used a part of the Korman (1984) corpus as input to our model. This corpus is included in the CHILDES database (MacWhinney & Snow, 1990) and consists of speech directed to infants between the ages of 6 and 16 weeks. Korman's original research focused on the interactions between mothers and their preverbal infants. The data were collected in the UK. We found this corpus particularly appropriate because speech addressed to preverbal infants is likely to be different with respect to type–token frequency and utterance length compared with speech addressed to children who have already begun to utter multiword utterances (also cf. Aslin et al., 1996).

The Korman corpus consists of 37,549 words distributed over 11,376 utterances. It contains 1888 types of different words and has a type–token ratio of 0.05. Since the corpus is transcribed in CHAT format we needed to transform it into a phonological form with the addition of relative lexical stress patterns. For this purpose a dictionary was compiled from the MRC Psycholinguistic Database available from the Oxford Text Archive. This database includes relative stress information for a large number of words. Monosyllabic words are listed as unstressed, but we coded them as having primary stress (cf. the earlier discussion). Multisyllabic words were coded according to their rhythmic patterns in the database. The compiled dictionary listed orthographic, phonological (British pronunciation), and relative stress information for 9170 words (with a frequency of occurrence higher than 3). An additional 67 words (which were highly frequent in the Korman corpus) were coded by hand and added to the dictionary, resulting in a total of 9237 words. Using this dictionary the orthographic word forms in the Korman corpus were replaced with a citation form of their phonological counterpart and information about their relative stress pattern.

The preprocessing of the corpus led to a reduction in the size of the original corpus because some of the words in the Korman corpus were not present in the dictionary. Utterances which included a word that was not found in the dictionary were deleted from the corpus. The resulting corpus consists of 27,467 words distributed over 9108 utterances, leaving out 1058 of original word types (covering 2894 tokens). Many of the word types left out are repetitions of baby babble (e.g. *agooo* and *boobababoobaba*), examples of "baby talk" (e.g. *cupsy* and *ticky*), interjections (e.g. *huh* and *oooooh*) as well as onomatopoeia (e.g. *beep* and *boom*). Brent and Cartwright (1996) suggest leaving such words out of a (training) corpus because they occur in isolation more often than ordinary words, their use is highly idiosyncratic, and there is no standardised orthographic spelling for many of them.

The preprocessed corpus was then separated into a training corpus and a test corpus by removing every 10th utterance and placing it in the test corpus. The remaining 8181 utterances constituted the training corpus whose distributional characteristics can be found in Table 1. The training corpus is highly biased towards monosyllabic words in that 86.8% of all words are monosyllables, compared with 12.3% bisyllabic words and 0.9% trisyllabic words. Aslin et al. (1996) also found a similarly high proportion of monosyllabic words in their corpus of speech to 12-month-olds. On average each utterance has a length of 3.0 words and contains 9.0 phonemes. Average length of a word is 3.0 phonemes. The type–token ratio in the training set is 0.03 indicating a considerable amount of repetition in the input; that is, the same words are used again and again. Finally, it is worth noticing that 77.3% of the bisyllabic words follow a strong–weak stress pattern as do 77.6% of all multisyllabic words (patterns 20 and 200).

Table 2 contains the distributional characteristics of the test corpus. As in the training corpus, there is a strong bias towards monosyllabic words. Thus,

TABLE 1
Distribution of Word Tokens and Types Across the Nine Stress Patterns Found in the Training Corpus

Word Form	*Stress*	*Tokens*	*Types*
Monosyllabic	2	21,402	523
Bisyllabic	20	2,338	210
	22	389	7
	02	284	40
	00	15	1
Total		3,026	258
Trisyllabic	200	182	24
	020	33	7
	202	3	1
	210	2	1
Total		220	33
All Forms	9	24,648	814

"2" indicates primary stress, "1" secondary stress, "0" no stress.

TABLE 2
Distribution of Word Tokens and Types Across the Eight Stress Patterns Found in the Test Corpus

Word Form	*Stress*	*Tokens*	*Types*
Monosyllabic	2	2,438	263
Bisyllabic	20	273	84
	22	49	2
	02	35	17
	00	1	1
Total		358	104
Trisyllabic	200	20	9
	020	2	2
	202	1	1
Total		23	12
All Forms	9	2,819	379

86.5% of the words are monosyllabic, whereas 12.7% are bisyllabic and 0.7% are trisyllabic. The type–token ratio is 0.13, indicating less repetition in the test corpus than in the training corpus. Of the 814 word types in the training corpus, 55.4% of these (451 types) do not occur in the test corpus. Of the 379 word types in the test corpus, 4.2% of these (16 types) do not occur in the training corpus.

Model and Task

The SRN employed in the present simulations resembles the simplified model described in the previous section. Figure 2 provides an illustration of the scaled-up SRN. Like the previous model, the current model takes single phonemes as input—but this time represented as the activation of bundles of 11 phonetic features (using the feature scheme found in Appendix A)—or an utterance boundary represented by a single unit (marked "Ubm"). Moreover, the current model also has two additional input units which are used to represent relative lexical stress. When both units are off (i.e. "0 0") the syllable has no stress, whereas if the first unit is on (i.e. "1 0") the syllable has an intermediate level of stress, and a strong stress level when the second unit is on (i.e. "0 1"). For monosyllabic words the second unit is always on, but for a word such as /tU-mi/ (*tummy*) with a strong–weak stress pattern (2–0) the second unit will be on for the duration of the first syllable and both units off during the second. Since utterance boundaries are construed as silences, both stress units and phonological features are off when the boundary unit is on. The output layer consists of a set of 36 phonological units, an utterance boundary unit, and two units coding for stress. We used

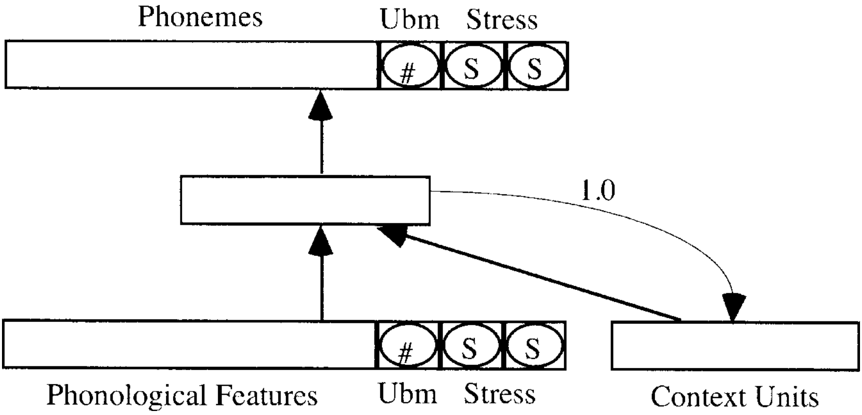

FIG. 2. The SRN trained on the Korman (1984) corpus. Thick arrows indicate trainable weights, whereas the thin arrow denotes the copy-back weights (which are always 1). Note that the size of the layers does not indicate the number of units they contain.

localist representations of phonemes as output to facilitate performance assessments and analyses of segmentation errors. The network in addition has 80 hidden units and 80 context units, resulting in a 14–80–39 SRN.

The SRN was trained under five training conditions that varied the combination of cues provided as input:

1. *phon-ubm-stress* phonological, utterance boundary, and stress information
2. *phon-ubm* phonological and utterance boundary information
3. *phon-stress* phonological and stress information
4. *phon-only* phonological information
5. *stress-ubm* stress and utterance boundary information.

The task of the SRN was to predict for the next time-step the same combination of cues that it received as input (but with the phonological information coded in terms of phonemes instead of features). For example, in the phon-ubm-stress condition the network was required to predict the next phoneme or utterance boundary as well as the appropriate relative stress level. It is important to note that in none of the training conditions were word boundaries explicitly marked in the input. Each network was trained on a single pass through the training corpus (that is, 83,130 phoneme tokens for the conditions involving utterance boundary information and 73,947 phonemes for the conditions without utterance boundary information[3]). Initial explorations of the learning parameter space indicated that the best performance was to be found using the same settings as in the simulations involving the simple model. We used a learning rate of 0.1, random initialisation of starting weights within the interval (−0.25, 0.25), and a momentum of 0.95. Identical learning parameters and an identical set of initial weights were used for all training conditions.

Results

Following training, the nets were tested on the test corpus as well as on corpora consisting of novel words, nonwords (illegal strings such as /slrf/), and two sets of 500 bisyllabic words with strong–weak and weak–strong stress patterns, respectively.

General Performance

In the following we only report the results found when the trained nets were tested on the test corpus. However, very similar performance was obtained when the trained nets were tested on their respective training corpora.

Assessing the Acquisition of Phonology. Before considering the primary issue concerning word boundaries, we need to assess how the models performed on the prediction task (i.e. the immediate task). Successful performance would mean that a model had encoded information about phonological regularities. We evaluated this by asking how well the network was able to predict the next target phoneme based on the prior sequence. In order to test how different combinations of cues affected the learning of phonology, we calculated the mean squared error (MSE) for the phonological output units only. This permitted us to compare four of the five training conditions. The error scores fell into two groups.[4] Given the results from our simplified model, it was not surprising to find that the network

[3] This length difference is simply due to the addition of the boundary marker at the end of utterances. This difference does not alter the relevant statistics concerning the learning of the phonological regularities in the training corpus since the network on these occasions only receives error with respect to the utterance boundary marker.

[4] Given that the activation over the phonological units typically adds up to 1, the range of error tends to be between 0 and 1. On this scale the reported differences in error are important as indications of how well the nets perform on the prediction task given different cue combinations. It should be noted, however, that the prediction task is not deterministic (e.g. following /k/ we may get /&/, /e/, /I/, etc.). Because the net tries to accommodate this indeterminacy, it will tend to encode the probabilities of the set of next possible phonemes given previous context. This leads to a better representation of the phonological regularities, but also to a relatively high level of error on the task of predicting the exact target phoneme. Thus, the error will never reach 0, even under ideal conditions.

trained in the phon-only condition had the highest error: 0.923. The net trained using a combination of phonological and stress information (phon-stress) had essentially the same error: 0.922. These two nets formed the first group with no reliable difference between the two error scores, $t(16{,}884) = 0.084$, $P > 0.9$. In the second group, we found the network trained on phonology and utterance boundary information (phon-ubm) had an MSE of 0.822, and the network trained using all three kinds of cues (phon-ubm-stress) produced the lowest error score, 0.802. The difference between these two error scores was marginally significant, $t(18{,}740) = 1.814$, $P < 0.07$. More importantly, there was significant difference between the two groups (as exemplified by a comparison of the error between the phon-stress condition and the phon-ubm condition: $t(17{,}812) = 9.024$, $P < 0.0001$). This result replicated a finding from the simplified model, demonstrating that the addition of utterance boundary information during training helped learning the phonology of the training corpus.[5] Although the error scores indicated that adding relative lexical stress was only marginally helpful in the immediate task of learning phonology, we shall see below that stress did have a significant positive effect on the learning of the derived task of word segmentation—especially generalisation to novel words.

We also plotted the error over time in order to determine whether the error score was high at the onset of a word and gradually fell as more of the word was processed as reported in Elman (1990). Although error tended to decline in cases where the phonological context was very constraining, we did not find that it reliably decreased during the processing of a word. Instead, error declined over phonological subclusters where the sequence of phonemes was highly predictable, and this phenomenon did not correlate well with word boundaries. For example, error dropped across subsequences such as /he/, where /e/ is extremely likely given /h/,[6] but not across words such as /bjut@ fUI/ (*beautiful*), which presumably is highly constraining at the end of the word. Our failure to replicate Elman's results may be a result of both using a much larger number of words (thus, a particular sequence is less likely to accurately determine which phoneme comes next) and the fact that the input is coded in terms of features rather than phonemes (the contributions of the weights from a single feature are less constraining of the

[5] As mentioned earlier, Allen and Christiansen did not find a difference in performance between the two nets trained on the vtp corpus with and without utterance boundary markers. It was hypothesised that this was due to a ceiling effect brought about by the strong phonological regularities of the vtp vocabulary. As suggested by Allen and Christiansen, when a net is trained on "real" language these regularities are less strong, permitting the utterance boundary cue to be more useful both for the learning of phonology and for the learning of word segmentation. This suggestion was corroborated by the results discussed previously.

[6] /he/ occurs in highly frequent words such as *hey* and *hello*. The transitional probability of /e/ given /h/ is 0.41.

following phoneme than the set of weights from a particular phoneme because each feature participates in many phonemes). Also, in the case of shorter legal words embedded within longer words, the network tends to predict a boundary at the end of such an embedded word, hence increasing the error at this position in the longer word. Cairns et al. (1994) appear to get similar results in that many of their false alarms correlated with phoneme clusters that were well-formed given phonotactic considerations; that is, they marked legal word initial or word final segment sequences (e.g. placing a boundary ("#") as in /r&#ptOr/ (*raptor*) creates the phonotactically illegal onset /pt/, whereas placing the boundary as in /r&p#tOr/ creates the legal offset /p/ and the legal onset /t/). Thus, variations in the error landscape appear to be a poor source of information regarding lexical boundaries, but may point to phoneme subclusters that constitute phonotactically well-formed units.

Indicating Word Boundaries. We now turn to word segmentation performance measured in terms of the activation of the boundary unit between words. Recall that in none of the training conditions did the nets receive explicit information about what counted as word boundaries. Rather, it was expected that the networks receiving utterance boundary information would learn the regularities concerning which phoneme clusters occurred at the ends of utterances and then generalise this knowledge to ends of words inside an utterance. Figure 3 depicts the activation of the boundary unit during the processing of the first four utterances in the training corpus by the phon-ubm-stress trained SRN, an example on which the model was quite successful at this task. The results presented in subsequent sections describe the behaviour of the networks more generally.

Notice that in all but one instance the activation of the boundary unit is always higher at lexical boundaries (black bars) than at word internal positions (grey bars) across the four utterances. In general, both word frequency and the phonological uniqueness of a particular word appear to affect how much the boundary unit is activated at the end of a word (and not word internally). For example, /ju/ (*you*) is by far the most frequent word in the training corpus (1551 tokens), resulting reliably in the activation of the boundary unit following the input sequence /ju/. In contrast, /hed/ (*head*) is much less frequent (20 tokens) but this sequence of phonemes always occurs at the ends of words in the training corpus, allowing the net to reliably predict a boundary following /hed/. Another important factor is whether a particular word occurs at the end of utterances or not. The last example, /hed/, often occurs at the end of utterances (as in Fig. 3) and this provides additional strength to the activation of the boundary unit (see later). A

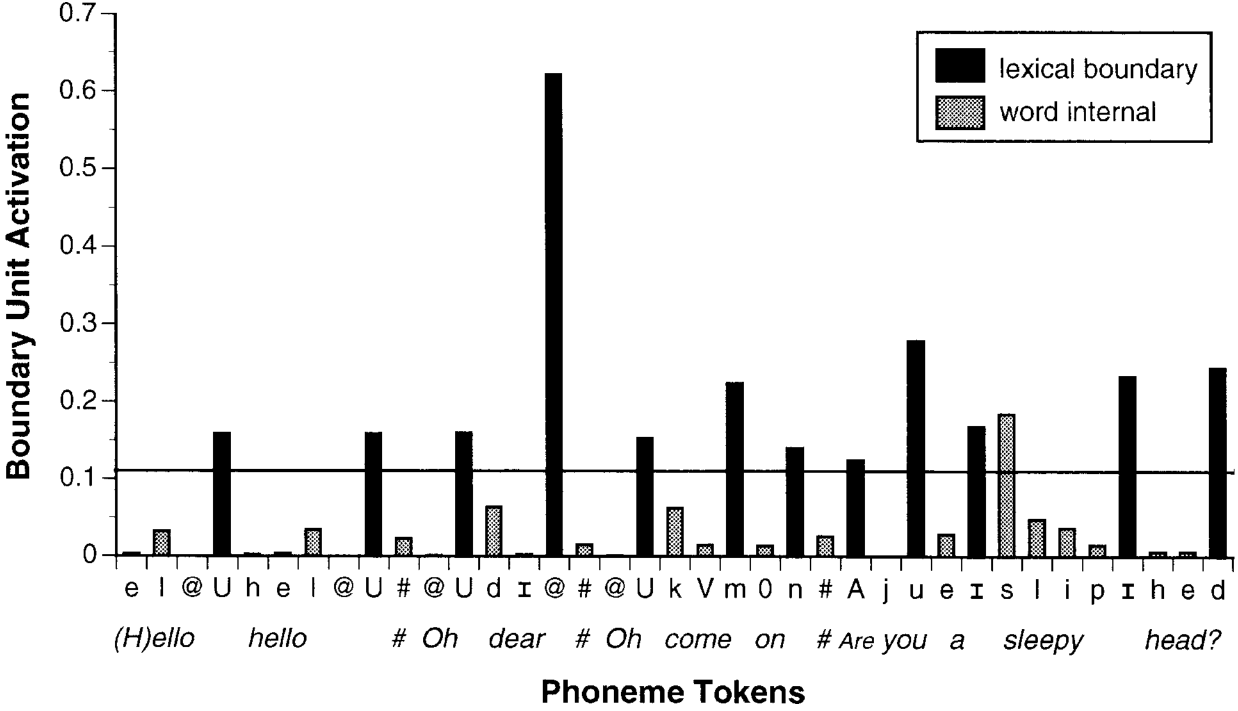

FIG. 3. The activation of the boundary unit during the processing of the first 37 phoneme tokens in the training corpus (for the SRN trained under the phon-ubm-stress condition). Activation of the boundary unit at a particular position corresponds to the network's hypothesis that a boundary follows this phoneme. Black bars indicate the activation at lexical boundaries, whereas the grey bars correspond to activation at word internal positions. The horizontal line indicates the mean boundary unit activation across the whole corpus. A gloss of the input utterances is found beneath the input phoneme tokens (with "#" denoting an utterance boundary).

sequence such as /eI/ (*a*[7]) never occurs at the end of utterances, so here the network activates the boundary unit to a lesser degree. In this case, the network is relying on the similarity of /eI/ to other words, such as /pleI/ (*play*), /heI/ (*hey*), and /deI/ (*day*), which do occur at the end of utterances. However, generalisation from these words is made difficult because /eI/ also occurs internally in words such as /beI-bI/ (*baby*), /meIk/ (*make*), and /feIs/ (*face*). In fact, it is because of the existence of words such as /feIs/ that the network is making the mistake of activating the boundary unit following the sequence /eIs/. Still, the network manages to activate the boundary unit higher at the end of /eI/ than the overall activation mean (indicated by the horizontal line).

One way of quantifying the results pertaining to the activation of the boundary unit across a corpus is to look at the average activation at word internal positions, lexical boundaries, and utterance boundaries. Figure 4 shows these mean activations for the phon-ubm-stress trained SRN tested on the test corpus. Notice that the activation of the boundary unit at lexical boundaries is five times higher than at word internal positions. This replicates a similar finding in the simplified model (trained on the vtp corpus) where the mean boundary unit activation at lexical boundaries (0.204) was five times higher than at word internal positions (0.04). In the feed-forward network of Aslin et al. (1996: Fig. 8.5) the activation of the boundary unit at lexical boundaries is roughly 2.5 times higher than the activation at word internal positions. The proportional difference between activation of the boundary unit at the ends of words compared with activation word internally provides an indication of how well the net can differentiate between them.

The activation at utterance boundaries is 36% higher than at lexical boundaries, indicating that it is easier for the net to predict a boundary for words occurring at the end of utterances (as opposed to predicting boundaries utterance internally). This is because some words (e.g. /hed/, as mentioned earlier) tend to occur more often at the end of utterances than in positions within an utterance. Other words (e.g. /eI/) never occur at the end of utterances and will therefore tend to lower the activation at lexical boundaries within an utterance. The closer the activation at utterance boundaries is to that at lexical boundaries, the better the network is at generalising knowledge about sequences at the end of utterances to lexical boundaries within utterances. In the simulations by Aslin et al. (1996), the boundary unit activation at utterance boundaries is nearly twice as high at lexical boundaries, suggesting somewhat better generalisation in our model.

[7] In contrast to many other dialects of English, the determiner *a* is indeed transcribed as the diphthong /eI/ in the particular British dialect that was used for the phonological encoding of words in the MRC Psycholinguistic Database.

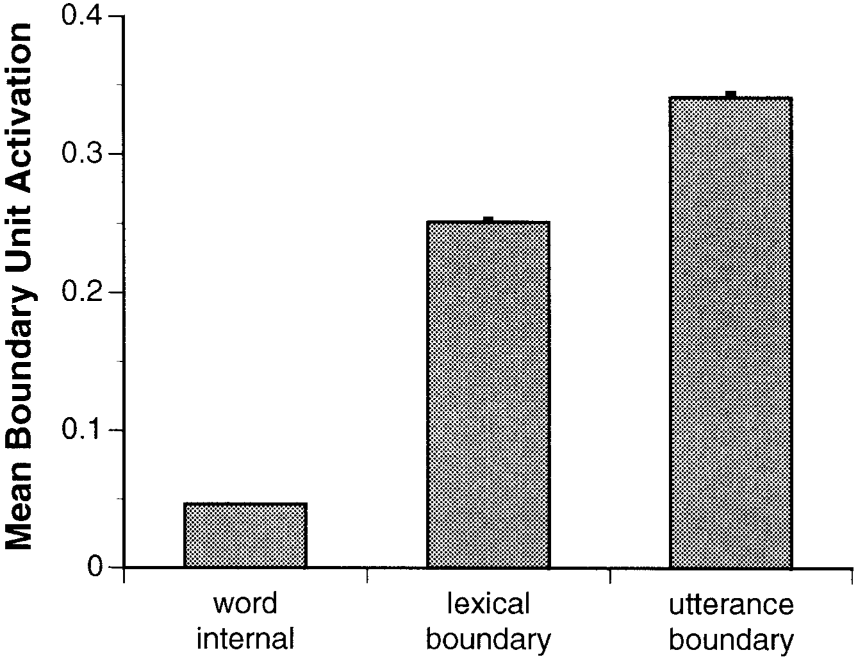

FIG. 4. The mean activation of the boundary unit at word internal positions, lexical boundaries, and utterance boundaries. Results are shown for the phon-ubm-stress trained network tested on the test corpus. Error bars indicate standard error of the mean.

The results from the phon-ubm-stress net indicate that it has acquired a very good ability to distinguish between phonological sequences occurring at positions inside words and at lexical boundaries. The simulation results reported in Allen and Christiansen (1996) indicated that combining the phonological and utterance boundary cues improved performance significantly. Nevertheless, we saw earlier that adding the stress cue only marginally facilitated the immediate task of learning phonology. In order to investigate the effect of the stress cue on the activation of the boundary unit, we calculated standardised scores for the mean boundary unit activations at lexical boundaries and at word internal positions across the whole test corpus. Figure 5 shows the *z*-scores for the three training conditions involving utterance boundary information. The difference between the *z*-scores at lexical boundaries and word internal positions indicates the network's ability to distinguish sequences that end words from sequences that do not. A comparison between the phon-ubm-stress condition and the phon-ubm condition shows that the two nets differentiate the activation of the boundary unit at positions within and between word to an equal degree.

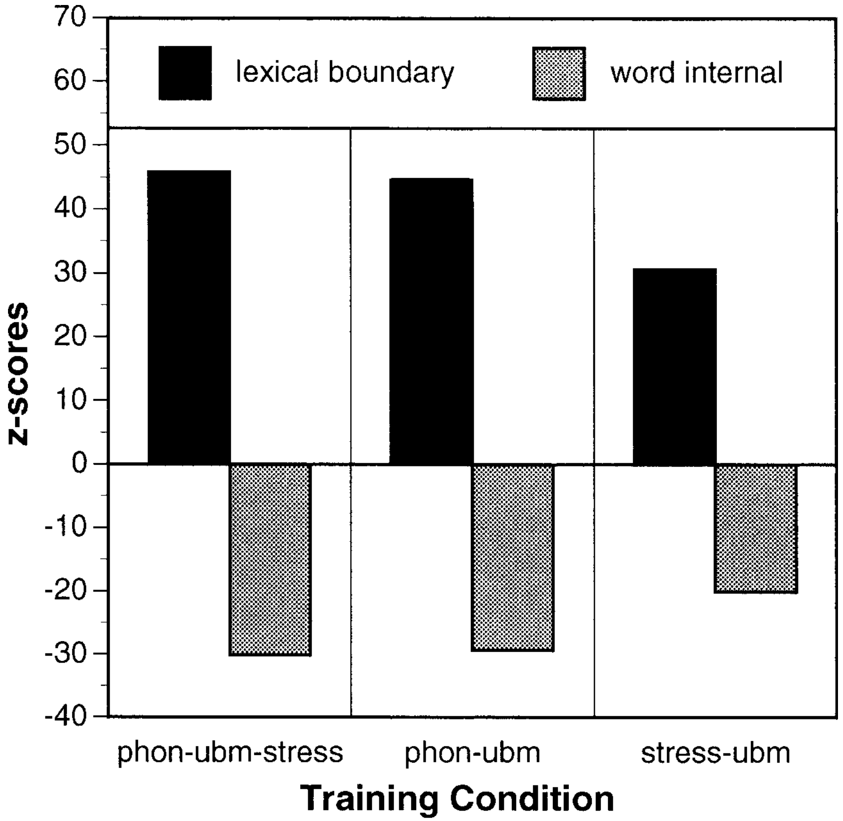

FIG. 5. The z-scores for the mean boundary unit activations at lexical boundaries and word internal positions standardised against the mean for the whole test corpus. Results are shown for the test corpus given the three training conditions involving utterance boundary information.

The net trained on the stress-ubm condition, on the other hand, reaches a level of performance 33% below the nets trained under the two other conditions. Although the results shown so far indicate that relative lexical stress is not a very informative cue, we shall see next that stress does facilitate the derived task of discovering words in speech when assessed in terms of individual word predictions.

Segmenting the Speech Stream. There are many ways to assess the performance of the networks on the task of detecting the boundaries of individual words. For the sake of comparison, we adopt the measure used by Brent and colleagues (Brent & Cartwright, 1996; Brent et al. 1994). They suggest that segmentation be assessed in terms of *accuracy* and *completeness*:

$$\text{Accuracy} = \frac{\text{Hits}}{\text{Hits} + \text{False Alarms}}$$

$$\text{Completeness} = \frac{\text{Hits}}{\text{Hits} + \text{Misses}}$$

Accuracy determines the proportion of correct predictions out of all the predictions that a system is making. Completeness, on the other hand, determines the proportion of correct predictions that the system actually made out of the set of possible correct predictions. Of course, these measures depend on what counts as a hit, a false alarm, or a miss. Using the approach of Aslin et al. (1996), we stipulate that a net has predicted a word boundary when the activation of the boundary unit is above the mean activation for that unit calculated across the whole corpus. A hit is recorded when the activation of the boundary unit is above the mean at a lexical boundary, whereas a miss is recorded when it is not. A false alarm is recorded if the activation is above the mean when there is no actual lexical boundary in the input. For example, in Fig. 3 we find 12 hits because all 12 black bars (indicating activations at lexical boundaries) are above the mean (indicated by the horizontal line). We also find a single false alarm because the grey bar for /s/ following /eI/ is above the mean when there is in fact no lexical boundary.

In order to assess word level performance additional definitions are needed. Following Brent and Cartwright (1996), a word level hit requires that the system correctly predicts both the word initial and the word final boundary (without any false alarms within the word). A miss occurs if the system does not segment a word at its appropriate boundaries, and a false alarm stems from segmenting a word incorrectly. As an example, consider the string:

#the#dog#s#chase#thec#at#

where "#" corresponds to the prediction of a boundary. At the boundary level we have five correct boundaries (hits), one miss, and two false alarms, resulting in 71.4% accuracy and 83.3% completeness. Turning to the word level, we have two hits (*the*, *chase*), three misses (*dogs*, *the*, *cat*), and four false alarms (*dog*, *s*, *thec*, *at*), corresponding to an accuracy of 33.3% and a completeness of 40.0%.

TABLE 3
Percent Accuracy and Completeness for the Three Nets Trained with the Utterance Boundary Cue, for an Algorithm that Treats Utterances as Words, and for a Pseudo-random Algorithm that Predicts Lexical Boundaries Given the Mean Word Length

Training Condition/ Program	*Words*		*Boundaries*	
	Accuracy	*Completeness*	*Accuracy*	*Completeness*
phon-ubm-stress	42.71	44.87	70.16	73.71
phon-ubm	37.31	40.40	65.86	71.34
stress-ubm	8.41	18.02	40.91	87.69
utterances as words	30.79	10.15	100.00	32.95
pseudo-random	8.62	8.56	33.40	33.15

Results are shown for the tasks of segmenting words and correctly predicting lexical boundaries given the test corpus.

The results from the previous section suggested that the stress cue did not provide any additional help in differentiating lexical boundaries from word internal positions. Table 3 presents results from an analysis of how well the nets trained under the different conditions performed on the tasks of segmenting words and predicting boundaries. These results suggest that stress information is a valuable cue to learning the task of lexical boundary prediction and subsequently to word segmentation. More than 42% of the words predicted by the net trained under phon-ubm-stress condition are correctly segmented words—a level of performance which is significantly better than that of the phon-ubm trained net ($\chi^2 = 18.27$, $P < 0.001$). The value of the stress cue for the word segmentation task is also reflected in a significantly higher rate of completeness (compared with the phon-ubm net: $\chi^2 = 11.51$, $P < 0.001$). In terms of the prediction of lexical boundaries, less than 30% of the predicted boundaries were incorrect, compared to 34% in the phon-ubm condition ($\chi^2 = 12.69$, $P < 0.001$). The same pattern is also evident concerning the completeness of the lexical boundary predictions ($\chi^2 = 4.00$, $P < 0.05$). Another way to approach the completeness of lexical boundary predictions is to look at the hit–miss ratio. This ratio is 2.8 for the phon-ubm-stress trained net and 2.5 for the net trained under the phon-ubm condition. In comparison, Cairns et al. (1994) report a lexical boundary completeness of 21% and a hit–miss ratio of 0.3.[8] Aslin et al. (1996, Fig. 8.7) with a lexical boundary completeness of about 62% obtain a hit–miss ratio of 1.6. Turning to the net trained under the stress-ubm condition, we see that it performs poorly on the word segmentation task, but somewhat better on the

[8] Nonetheless, this rather poor result should be seen in the light of fact that they trained their model on a corpus of adult conversational speech, which is likely to have significantly different statistics than a child-directed corpus.

task of predicting lexical boundaries. The high level of completeness on the latter task is because the net is over-segmenting, predicting a word boundary for 64% of the possible positions in the test corpus. The network therefore only misses about 12% of the actual lexical boundaries, but then also has 44% more false alarms than hits.

As nonconnectionist benchmarks for the performance of the networks, two programs were created. The first program can be seen as an implementation of a simplified version of the single word segmentation strategy; that is, the idea that the child bootstraps into the word segmentation task by focusing on single word utterances in the speech stream. Here the program simply treats every utterance as a word and predicts lexical boundaries appropriately. The second program randomly predicts lexical boundaries given knowledge of the mean word length and its standard deviation (as well as the assumption of normally distributed word lengths). This program thus has knowledge about the words in the speech stream that is not available to the networks nor to the "utterances as words" program. Because of the stochastic nature of the pseudo-random program, results are averaged over 100 runs. Results from both programs are reported in Table 3.

Somewhat surprisingly, the "utterances as words" program does fairly well on the word-level segmentation task. The relatively high level of accuracy is due to the fact that almost a third of all utterances in the test corpus are single word utterances. Of course, the program under-segments many utterances, which is why word level completeness is around 10%. The perfect score on the lexical boundary accuracy is an artifact of the program only predicting lexical boundaries at utterance boundaries, which means that all its predicted boundaries will be correct by definition. The completeness score, on the other hand, reflects its failure to predict any lexical boundaries at utterance internal positions, a serious drawback. The pseudo-random program is doing worse on the word segmentation task than the "utterances as words" program, despite its built-in knowledge about the distribution of word lengths. It does, however, reach a level of performance comparable to the "utterances as words" program in terms of lexical boundary completeness.

In comparison with the performance of the nets trained under the phon-ubm-stress and phon-ubm conditions, both programs fare worse on all accounts (save the artificially high lexical boundary accuracy score for the "utterances as words" program). Thus, although the single word segmentation strategy as implemented here does have some credibility, its usefulness declines as the child grows older and the number of single word utterances decreases. Conversely, the performance of the networks is likely to be able to accommodate such changes in the input (see the later Discussion).

A broader perspective on the performance of the networks on our corpus can be found by considering the work of Brent and colleagues. Simulations were conducted in which we ran an on-line version of the DR algorithm on our training corpus.[9] Whereas the implementation in Brent and Cartwright (1996) used batch learning in which the program had access to the entire corpus at once, the on-line version segments one utterance at a time and updates the lexicon before processing the next utterance. Unlike the SRN which was tested after training, the assessment of the DR program is based on output generated while processing the corpus. Using similar definitions of accuracy and completeness as in the SRN simulations, the average performance of the DR program across the training corpus resulted in an word segmentation accuracy of 46.50% to and a completeness of 51.89%. When the vowel constraint and information about consonant clusters occurring at initial and final positions in utterances was incorporated into the DR program, performance was improved significantly, resulting in 72.06% accuracy and 65.05% completeness. The results suggest that our model is performing well in comparison with the level of performance that can be achieved via the abstract strategy of the DR algorithm.

In comparing the results from the SRN and DR simulations it should be kept in mind that the two models are not performing the same task. Our SRN model learns to segment speech while being trained on a segmental prediction task, whereas the DR model is building a lexicon and using it to inform the segmentation process. The DR model as implemented takes whole utterances as input, generates all possible segmentations for that utterance, and applies a set of evaluation criteria to select the optimal segmentation from the candidate set. These criteria seek to minimise the number of words in a segmentation, select segmentations with high frequency words over low frequency words, and minimise the number and length of new lexical entries resulting from a segmentation. As such, the DR model provides a description of an abstract strategy for the calculation of an optimised description of the input. In contrast, the SRN model suggests a psychological mechanism by which children may integrate multiple cues in the service of word segmentation.

Processing with Partial Information. The previous sub-sections have demonstrated that having three cues available during training is better than having only two or fewer when it comes to the derived task of predicting lexical boundaries. Another way of looking at the informativeness of individual cues is to supply a trained network with partial information. Here,

[9] Previous results from using this on-line implementation were presented in Brent (1996). In order to run the DR program on our training corpus it was necessary to remove utterances involving vowel-less interjections, such as /hm/. The resulting corpus contained 7900 utterances.

we focus on the SRN trained under the phon-ubm-stress condition and study the effects of receiving partial information after learning.

Figure 6 depicts standardised scores for the mean boundary unit activations at lexical boundaries and at word internal positions for the phon-ubm-stress trained net tested on the test corpus using different combinations of cues. The differences between lexical boundary and word internal *z*-scores decrease as the cue combinations become less informative. Best performance is obtained when all three cues are provided to the network. The phon-stress cue combination follows in close succession. Next,

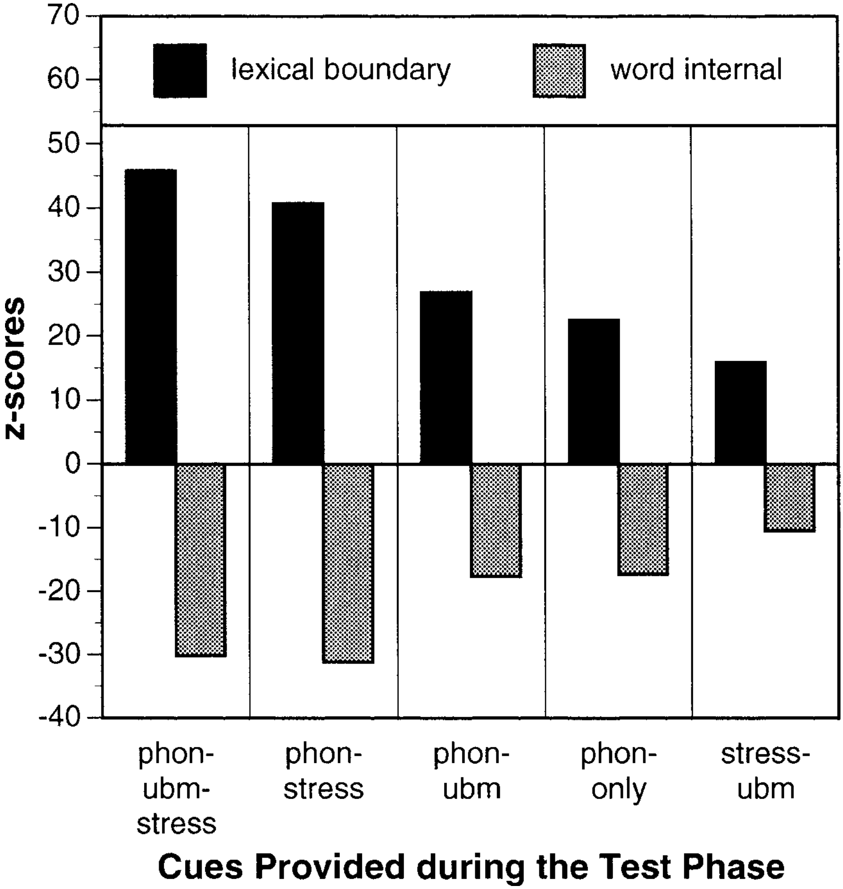

FIG. 6. The *z*-scores for the mean boundary unit activations at lexical boundaries and word internal positions standardised against the mean for the whole test corpus. Results are shown for the network trained using all three cues but tested on the test corpus using different combinations of cues.

the combination of phonology and utterance boundary information and the single phonology cue group together at a slightly lower level of performance. As was the case during learning, the combination of stress and utterance boundary information (stress-ubm) is also a relatively poor test cue to detecting lexical boundaries. When combined with phonology during testing, stress information appears to be a more valuable cue than utterance boundary information. Presumably, this is because the relative stress patterns may be helpful for the segmentation of multisyllabic words and, as we shall see later, for generalisation to novel words. In contrast, utterance boundary information only allows the net to "reset" itself; that is, it then knows that a new utterance is about to begin. Still, receiving this cue is slightly better than not, as indicated by the small difference between the phon-ubm and phon-only test condition.

This analysis is further supported by the accuracy and completeness scores at lexical boundaries and words as presented in Table 4. Although the scores for the phon-stress test condition is fairly close to the phon-ubm-stress condition, the three cue test condition results in a reliably better performance on word accuracy ($\chi^2 = 10.06$, $P < 0.01$), word completeness ($\chi^2 = 8.76$, $P < 0.01$), and lexical boundary accuracy ($\chi^2 = 4.21$, $P < 0.05$), but not on lexical boundary completeness ($\chi^2 = 2.80$, $P < 0.09$). As a group, these two test conditions result in a significantly higher level of performance than the group consisting of the phon-ubm and the phon-only test conditions (p's < 0.001). There was no reliable difference between the accuracy and completeness scores in the second group (p's > 0.9). Thus, when it comes to the prediction of lexical boundaries and word level boundaries (post-training), stress is a more valuable cue than utterance boundary information. When a net has become sensitive to lexical stress patterns during training, the absence of the stress cue in the testing phase is likely to be perceived as if all of the input have a zero level of stress. Predictions will therefore be made according to the network's knowledge of phonological sequences with zero

TABLE 4
Percent Accuracy and Completeness for the Net Trained Using all Three Cues, but Tested on the Test Corpus Using Different Combinations of these Cues

Testing/ Condition	*Words*		*Boundaries*	
	Accuracy	*Completeness*	*Accuracy*	*Completeness*
phon-ubm-stress	42.71	44.87	70.16	73.71
phon-stress	38.67	40.97	67.69	71.73
phon-ubm	15.45	13.37	47.09	40.76
phon-only	14.60	12.38	46.80	39.69
stress-ubm	8.10	7.80	35.91	34.59

Results are shown for the tasks of segmenting words and for correctly predicting word boundaries.

stress patterns, leading to many erroneous predictions. Receiving a combination of stress and utterance boundary information, as in the final test condition, results in the poorest performance. The accuracy and completeness scores are significantly worse than in the phon-only test condition (p's < 0.001), indicating the importance of the phonological cue to the segmentation task. The implication of the results presented in this section is that the cues that influence performance after training may not correspond to those that play an important role in learning.

So far, we have shown that the combination of phonological, stress, and utterance boundary information together provide valuable cues to the derived task of word segmentation. The testing of the SRNs included only a fraction (4.2%) of previously unseen words. Next, we tested the trained nets on a set of novel words and a set of nonwords, the latter violating phonotactic constraints found in the training corpus.

Performance on Novel Words and Nonwords

In order to assess how the types of phonological knowledge obtained in the networks constrained generalisation, we tested two of the networks on a set of 50 novel words and 50 nonwords. Our novel words are real words of English that the net had not been trained on. They are a pseudo-random sample drawn from the MRC Psycholinguistic Database, with two constraints. First, we only included words with final syllables that did not appear in the training set, in order to test the network's ability to generalise to novel sequences. Second, the bisyllabic novel words were constrained to a stress distribution similar to that of the training corpus as a whole. Of the bisyllabic novel words 80% were stress initial. In the training corpus, 77% of the multisyllabic tokens were stress initial. The set of novel words we used appears in Appendix B.

Our 50 nonwords consisted of two types of words. The first 20 nonwords consisted of monosyllabic sequences with offsets that did not occur in the MRC database (i.e. either very low frequency or non-existent in English). An offset is defined here as the set of consonantal segments following the final vocalic segment of the word. These were constructed by computing a list of all offsets that appear in the database, and comparing this list to a computed set of all possible offsets of one, two, or three phonemes. Nonwords such as /skiSD/ (*skeeshth*) were then created by concatenating a legal onset and vocalic segment (e.g. /ski/) with one of the clusters (e.g. /SD/) drawn randomly from the resulting list of non-existent offsets. A second set of 10 nonwords consisted of sequences without vocalic segments. These words were created by concatenating two, three, or four legal consonantal sequences. For example, the nonword /slrm/ consists of the three phonotactically legal sequences /sl/, /lr/, and /rm/. All three of these legal sequences appear in the training corpus, but the sequence as a whole cannot constitute a word. An additional set of 20 nonwords was created by prefixing a legal initial syllable (/tIm/) to 20 of the monosyllables. The bisyllabic nonwords were given stress on the first syllable. The set of nonwords also appears in Appendix B.

We presented each of the novel words and nonwords one at a time to both the network trained on the phon-ubm-stress condition and the net trained on the phon-ubm condition. We measured the activation of the boundary unit word internally and word finally for all words. Both networks differentiated the ends of words from word internal positions when tested on the novel words, in that they produced a significantly higher activation of the boundary unit at the ends of words than word internally: Phon-ubm-stress net, $t(98) = 6.75$, $P < 0.0001$; phon-ubm net: $t(98) = 5.59$, $P < 0.0001$. Neither net showed a reliable difference for nonwords: Phon-ubm-stress net, $t(98) = 0.566$, $P > 0.5$; phon-ubm net: $t(98) = 1.94$, $P > 0.09$). These results indicate that both nets acquired gross phonotactics, in that they showed the capacity to differentiate between possible and impossible novel phonetic strings. The novel words produced higher activation of the boundary unit at word boundaries—the nonwords did not.[10]

The results suggest that the network trained with stress and that trained without stress perform in a similar way, but a closer look shows that the net trained on all three cues is better at correctly identifying words than the net trained on two cues. When the mean activation across the word served as the criterion for whether a boundary was posited at a given position, the net trained on all three cues identified 23 of the 50 novel words correctly in isolation (completeness = 46%), while the net trained on only two cues identified only 11 of the 50 novel words (completeness = 22%). On this measure, the three-cue net performs significantly better than the two-cue net ($\chi^2 = 4.23$; $P < 0.05$). This result shows that adding sensitivity to stress to the training regimen allows the net to perform better on the task of identifying the boundaries of completely novel words, even though stress on its own is not a good predictor of the locations of word boundaries.

Looking at the types of error made when the networks did not make correct identifications is also quite instructive—here we focus on the phon-ubm-stress trained net. The incorrect boundary predictions (on 27 words) largely correspond to legal (i.e. existing) word boundary sequences in the language being learned. Of the nineteen syllabic sequence types after

[10] These results are consistent with those reported in Allen and Christiansen, where the vtp network was tested on novel words and nonwords. The results also have an important parallel in work by Saffran, Aslin and Newport (1996) and Saffran, Newport, and Aslin (1996) showing, respectively, that adults and eight-month-old infants as a consequence of exposure to structured input have the ability to differentiate learned words from unknown words in speech based on distributional information.

which the network incorrectly predicted a boundary, only five of these did not occur at the end of a word in the training corpus, and only three of these did not occur word-finally in the MRC database. For example, the network predicts a boundary after two phonemes of the novel word /eIdZ/ (*age*). Although this is considered an error, it is a reasonable one considering that /eI/ is the phonological encoding on which the network was trained for the word "*A*", and that is embedded within the longer word "*age*". The three sequences that do not appear at the ends of any words in the dictionary are /brA/, /bU/, and /l0/, each of which ends in a vowel that appears at the end of a large number of the words in the training set. Thus, mis-segmentations are generally constrained by what counts as a legal offset in English.

A broader test of novel word segmentation was also performed to assess whether the network had, in a sense, developed a preference for the stress initial pattern prevalent in English bisyllabic items. We therefore presented 500 novel bisyllabic words to the three cue network. Of these 28.6% (143) were accurately recognised using the criterion described earlier. The same 500 words were then presented to the network with their stress encoding reversed; that is, the stress was placed on the final rather than the initial syllable. Under these conditions, the network recognised only 14.6% (73) of the novel words—49% fewer than when the same words were presented with stress on the initial syllable ($\chi^2 = 22.6$, $P < 0.001$). This shows that the stress cue is playing a significant role in the segmentation of novel words. As a result of encoding how stress in interaction with phonological information was correlated with boundary locations, the network appears to develop a preference for stress initial bisyllabic words (see General Discussion).

Discussion

Our simulations quite closely replicate the results reported for the simplified model of Allen and Christiansen (1996). SRNs appear to have the properties relevant for the integration of multiple cues in the segmentation of speech. They are able to learn not only the basic phonology and phonotactics of the training corpus necessary for carrying out the immediate task of predicting the next element in a phonological sequence, but also to combine the cues relevant for the derived task of word segmentation. Of the three cues, relative lexical stress turned out to be more helpful to the derived task than to the immediate task—in particular, for the generalisation to novel words. In contrast, the utterance boundary cue appears to be useful for the learning of both the immediate and the derived tasks.

Limitations and possible improvements. When looking at the word level segmentation performance (Table 3) it is clear that although the model is doing quite well, people are obviously able to detect more than 44% of the words in the speech stream. However, it should be kept in mind that (1) we are modelling the initial acquisition of the segmentation process rather than steady-state performance, and (2) skilled word recognition is likely to rely on other higher-level processes. Moreover, recall that the nets are trained on speech directed to infants between the ages of 6 and 16 weeks. It is as of yet unknown how well infants at this stage of development are able to segment speech. Although evidence of word segmentation has not been found for children younger than 7.5 months (Jusczyk & Aslin, 1995), it seems clear that the model as it is now will not be able to account for adult-level word segmentation without additional cues, training, and/or architectural augmentations. We therefore discuss ways in which the model could come to "grow" with the task.

As mentioned earlier, the net was trained on only a single pass through a training set consisting of 24,648 words. A rough estimate suggests that a child is exposed to about 1000–1500 words per day of direct speech[11] (leaving out a sizeable number of words that the child is exposed to in the general linguistic environment). Thus, we trained the model on less than a month's worth of data, leaving plenty of room for additional training. Such additional training on new corpora would expose the net to a greater variety of words. This is likely to lead to a better encoding of the phonotactic regularities of English. It may also allow the network to better capture the effects of individual word frequency. A possible complication may be that corpora of speech addressed to older children tend to contain a greater number of long sentences (in contrast to the short sentences of early motherese). Aslin et al. (1996) report that when they trained their network on longer (six–eight word) utterances it failed to perform successfully on the word segmentation task. However, this complication may be overcome when other aspects of sentential prosody than the marking of utterance boundaries is taken into account. Whereas the prosody of the short sentences/phrases of early motherese typically does not encode much utterance internal information (because of their minimal length), the longer utterances of late motherese often contain additional pauses which in many cases mark clausal or phrasal units (e.g. Fernald, Taeschner, Dunn, Papousek, Boysson-Bardies, & Fukui, 1989). Although these pauses do not appear be very reliable as cues to the acquisition of syntactic structure (e.g. Fernald & McRoberts, 1996; Fisher & Tokura, 1996), they none the less may provide strong cues to the endings of words (in the same way that the pauses at utterance boundaries are hypothesised to). Thus, for the purpose of word segmentation, the additional pauses are likely to play the same role as utterance boundaries, and may help solve the problem concerning longer utterances by dividing them into more

[11] The number of words per day was calculated from the Korman (1984) corpus in which each session is recorded over 24 hours (with a voice-activated microphone placed near the child).

manageable portions. A system relying generally on pauses as a potential cue to the word segmentation problem may therefore not suffer from the kind of "growing pains" suggested by the results of Aslin et al. Indeed, recent simulations by Christiansen and Allen (1997) indicate that pauses within utterances has a positive effect on the performance of an SRN, but that the benefit stems from allowing the net to "reset" itself following each pause (because whatever follows a pause is most likely the start of the a new word).

Lexical stress was primarily included as a possible cue to the word segmentation task, but a secondary motivation was to investigate whether the network would develop a preference to trochaic (i.e. strong–weak) sequences over iambic (weak–strong) sequences. Such a preference appears in infants acquiring English between the ages of six and nine months (Jusczyk et al., 1993). Further evidence suggests that nine-month-olds are able to integrate such preferences with acquired knowledge about distributionally legal sequences in the input (Morgan & Saffran, 1995). Although we chose to represent relative lexical stress directly, as an alternative one could focus on vowel quality; that is, the tendency to reduce vowels in weakly stressed syllables. This approach is taken by Cairns et al. (1994) in their connectionist model of word segmentation. However, as mentioned previously, analyses of speech directed to infants by Bernstein-Ratner (1987) and Morgan et al. (1996) have suggested that vowel quality taken alone is statistically a very poor cue to discovering unstressed segments. Our simulation results indicate that relative lexical stress is a valuable cue when integrated with phonological and utterance boundary information. Specifically, the results obtained through the manipulation of the stress patterns of the bisyllabic novel words show that the network was able develop a preference for the strong–weak sequences. Importantly, the network developed this strong–weak preference from the statistics of the input, without having anything like the "periodicity bias" of Cutler and Mehler (1993) built in.

It could be objected that representing stress across whole syllables instead of just vowel segments could both lead to a less realistic representation of stress and make the segmentation task easier for the net. On the issue of realism, we note that insofar as the acoustic correlates of stress can be measured in terms of amplitude and duration, the effects of stress are manifest on consonantal segments as well as vocalic ones. The extent to which stress is manifest on vocalic versus consonantal segments is surely different, but this probably results from the nature of the segments being produced: Vocalic segments are more open to changes such as duration than plosives, for example, because the identity of a plosive is inherently related to its duration. Thus, it is unclear whether representing stress on vowel segments only is more realistic than representing it across the whole syllable.

Nevertheless, we did run additional simulations with stress represented on vowel segments only. The results showed that the phon-ubm-stress net reached the same level of performance independently of whether stress was represented on the vowels only or across whole syllables (word accuracy comparison:$\chi^2 = 3.14, P > 0.09$; word completeness comparison: $\chi^2 = 0.32, P > 0.9$). It therefore seems unlikely that the segmentation task was unduly facilitated by our syllabic representation of stress. However, it is clear that our present implementation of stress as a step-wise function could be improved. An arguably better implementation would involve a smooth function reflecting the continuous nature of the acoustic parameters recognised as stress. We are currently pursuing more realistic implementations of relative lexical stress as well as considering additional ways of incorporating stress information.

In the present model, we have abstracted away from the acoustic variability that characterises fluent speech. The fact that we used a segmental representation of the input should not be taken as evidence that we assume that children have access to an innate phonemic inventory for his or her language. Learning to make the phonemic distinctions relevant for ones native language appears to occur within the first six months after birth (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992) and thus overlaps with early segmentation. We have abstracted away from this aspect of the acquisition process because we wanted to focus on segmentation—although we do believe that the two processes are likely to affect each other. In common with other computational models of word segmentation (e.g. Aslin et al., 1996; Brent & Cartwright, 1996), we used a corpus in which every instance of a particular word always had the same phonological form. In recent work (Christiansen & Allen, 1997), we have taken the first steps towards including acoustic variability, training SRNs on a corpus involving coarticulation; that is, segmental variation determined by the surrounding material. These simulations also include a novel way of incorporating input variability in terms of different phonetic realisations of individual segments. Earlier models, such as Cairns et al. (1994), modelled this variation by flipping random features with a certain probability. However, the variation in acoustic realisation does not vary randomly, rather for any segment certain features are more susceptible to change than others, and this is what the approach of Christiansen and Allen is meant to capture.

Other Potential Cues. A potentially important cue to word segmentation not treated here is the correlation of words with objects/situations/events in the immediate environment of the child. Although initial segmentation may take place based on information found in the linguistic input, later stages of word segmentation are likely to benefit from paying attention to correlations between speech input and nonlinguistic stimuli. For example, an infant observing that the sequence /bAl/ (*ball*) tends to occur in the presence of roundish things in the immediate environment could possibly use this information to segment /bAl/ out of longer sequences such as /hi3zeIbAl/ (*here's a ball*) (instead of into, say, the phonotactically legal sequences /hi3/, /zeIb/, and /Al/). Thus, infants may use such correlations to confirm (and reinforce) word candidates derived from the speech input and to ignore others.[12]

Other hypothesised cues relate to the use of the context and frequency with which a particular phonological subsequence occurs to determine whether it constitutes a good word candidate. For example, the DR algorithm of Brent and Cartwright (1996) is more likely to treat a subsequence as a word if it occurs in a variety of contexts and is familiar in the sense of occurring frequently in the input. An SRN may develop something akin to such context effects by picking up the differences in transitional probabilities occurring within and between words. That is, when a particular subsequence occurs in many contexts the transitional probabilities at its boundaries are going to be lower than at positions within it (because the context changes whereas the internal constitution of the subsequence does not). Frequency effects may arise because repeated exposure to the same subsequence will bias the internal transitional probabilities towards this particular phonological sequence. Thus, additional training on a larger variety of words may allow the network to adopt a somewhat similar strategy (although the effect is likely not to be as clearcut as in the DR algorithm—but it may fit better with the way context and frequency manifest themselves in infants).

GENERAL DISCUSSION

In this article, we have demonstrated how the integration of multiple cues in a connectionist model can allow it to learn a task for which there appears to be no explicit information in the input. When trained on a corpus of child-directed speech given phonological, utterance boundary, and stress information an SRN can learn to segment the speech input rather well. Utterance boundary information taken alone is not a reliable cue to word boundaries (although the "utterances as words" program was able to do better than the pseudo-random program); and likewise for stress. Even when combined, utterance boundary and stress information do not provide a useful cue to the segmentation process. Nevertheless, when combined with

[12] However, it should be noted that establishing such correlations is in itself a nontrivial task (see Christiansen & Chater, 1992, for discussion). Nevertheless, the integration of multiple cues approach may also be relevant for solving this problem (e.g. de Sa, 1994).

phonological information we saw that the three cues together provided a reasonably reliable basis for segmenting speech. Given the right kind of computational mechanism, independently unreliable cues can be integrated to yield a significantly more reliable outcome. Our results suggest that interactions between cues may form an additional source of information—that is, the integration of cues involves more than just a sum of the parts. The results also demonstrate that neural networks may provide the right kind of computational mechanisms for solving language learning tasks requiring the integration of multiple, partially informative cues.

This still leaves the question of what counts as a good cue. One way of approaching this question is to note the possible connection between language learning and the learning of complex sequential structure in general (for an overview of the latter, see Berry & Dienes, 1993). Connections between language acquisition and the learning of artificial languages have been suggested in the literature for both adults (e.g. Morgan, Meier, & Newport, 1987; Morgan & Newport, 1981; Saffran, Newport, & Aslin, 1996) and infants (e.g. Morgan & Saffran, 1995; Saffran, Aslin, & Newport, 1996), but perhaps the most important tie for our purposes is the use of SRNs to model both sequence learning (e.g. Servan-Schreiber, Cleeremans, & McClelland, 1989) and the learning of linguistic structure (e.g. Christiansen, in preparation; Elman, 1991, 1993). Cleeremans (1993) successfully applied SRNs to model the results from a number of sequential learning experiments. Analyses revealed a specific architectural limitation in relation to the prediction task: SRNs tend only to encode information about previous subsequences if this information is locally relevant for making subsequent predictions. For example, compare the phonological strings /heI/ (*hey*) and /peI-strI/ (*pastry*). In our training corpus, both strings are uniquely identifiable following the substring /eI/. Nevertheless, it is likely that the SRN would not be able to distinguish between the two if they occurred with the same probability in the corpus. Instead of predicting a word boundary following /heI/ (by activating the boundary unit) and an /s/ (by activating the /s/ unit) following /peI/, the back-propagation learning algorithm would probably drive the SRN to activate these two units equally in both cases. However, this limitation may be alleviated to some degree if the set of training items has a nonuniform probability distribution. Thus, in our simulations the network can learn to distinguish between /heI/ and /peI-strI/ because they have different frequencies of occurrence in the training corpus (590 and 4, respectively), forcing the net to encode the previous context. Fortunately, many aspects of natural language similarly involve probability distributions that are characterised by nonuniformity. For example, English motherese (and our training set) is skewed strongly towards monosyllabic words, and the stress of multisyllabic words is biased heavily towards a strong-weak pattern.

Focusing on this inherent limitation of SRNs, we may consider what would constitute a good cue for the net (and *mutatis mutandis*, for an infant). As a first approximation, we suggest that a good cue is one that involves a nonuniform probability distribution such that the net is forced to rely on more subtle aspects of the input in order to make correct predictions than it would without that cue. This will insure a deeper encoding of the structural regularities found in the input. In our simulations, this amounts to a better representation of the phonological regularities for which there is evidence in the input, and, in turn, a better basis for solving the derived task of word segmentation. The phonological cue on its own allows for a decent level of performance, but the net will tend to rely on fairly short sequences of phonemes in order and to make reasonable predictions. As evidenced by the results of Allen and Christiansen (1996) using the flat vocabulary, the addition of utterance boundary information forces the net to represent longer sequences of previous input tokens in order to reduce error on the prediction task. Compared with the net trained on the flat corpus without utterance boundary markers, the former net achieves a significantly better performance because it is forced to encode more of the regularities underlying the input.

But how might the stress cue help to improve performance when the net also has to predict stress patterns? Given the nonuniform distribution of stress patterns across multisyllabic words, the SRN could largely make correct predictions about stress by focusing on strong–weak patterns. Within both stressed and unstressed syllables the net can simply predict the current stress level as the next target. The crucial point occurs when the stress changes (step-wise) from strong to weak as the syllable boundary is straddled. If it was the case that offsets occurring at the end of the first syllable are different from those occurring at the end of the second syllable, then the net could potentially use that information to make the right predictions about most stress patterns (and thus reduce its error on the prediction task). We tested this prediction via a statistical analysis of the training corpus. Importantly, we found that the range of sequences at the ends of initial syllables in multisyllabic words is quite restricted. There are 11 consonantal offset types for stressed initial syllables, all of which end in single phonemes. For final, unstressed syllables there were 42 types, only 12 of which (28%) had this character, e.g. /brek-f@st/ (*breakfast*). Moreover, for monosyllables there were 52 types, only 17 of which (32%) ended in single consonantal phonemes (Appendix C). What this means is that a complex cluster could signal to the net that a word boundary is imminent. Without utterance boundary information as a cue to which clusters may end words, the stress cue becomes much less salient because otherwise only distributional regularities can point to the ends of words. However, once the network with utterance boundary information available is required to predict which phonemes will bear stress, encoding the differences between the ends of stressed syllables and the ends of unstressed syllables allows the network to predict the end of a stressed sequence.

This process of cue integration in neural networks has the additional advantage that given the right set of cues a network may avoid making unwanted over-generalisations. For any finite set of examples there will always exist numerous hypotheses that are consistent with that input. Without additional constraints[13] on this hypothesis space, a learning mechanism cannot reliably learn the regularities underlying the input; that is, in case of language acquisition the child cannot reliably learn the knowledge necessary to become a competent speaker of his or her language. The same problem arises in the acquisition of the individual cues to various aspects of language. Since each cue on its own is not a reliable source of information regarding the particular aspect of language that it is relevant to, many hypotheses may explain the regularities underlying each cue. However, if a gradient descent learning mechanism is forced to capture the regularities of many semicorrelated cues within the same representational substrate, it then becomes necessary for it to only represent hypotheses that are consistent with all the cues provided. Consider the conceptual illustration of three hypothesis spaces consistent with three different information sources (i.e. cues A, B, and C) in Fig. 7. If a network was only to learn the regularities underlying one of the cues, say A, then it could form a representation supporting any of the hypotheses in A. However, if the network is also required to learn to the regularities characterised by the B cue, it would have to settle on a representation which would accommodate the regularities found in both cues. Given that gradient descent learning works by stepwise reduction of the error, the network would have to settle on a solution that will minimise the error concerning the processing of both cues. This essentially means that the network has to settle on the set of hypotheses which can be found in the intersection of the hypothesis spaces A and B. Unless A and B are entirely overlapping (in which case they would not be separate cues anyway) or are disjoint (in which case one of them would not be a cue because of lack of correlation), this will constrain the overall set of hypotheses that the network will entertain. If the net has to pay attention to additional cues (e.g. C) then the available set of hypotheses will be constrained further. Thus, the integration of multiple cues in learning systems, such as SRNs, may constrain over-generalisation.

The fact that our model is able to achieve a quite high level of performance on a task for which there is no single reliable cue may have ramifications

[13] These constraints have typically been envisaged as being innate (e.g. Crain, 1991) or as arising out of (negative) feedback allowing the learning system to revise hypotheses which lead to over-generalisations. The integration of cues may provide a third possibility.

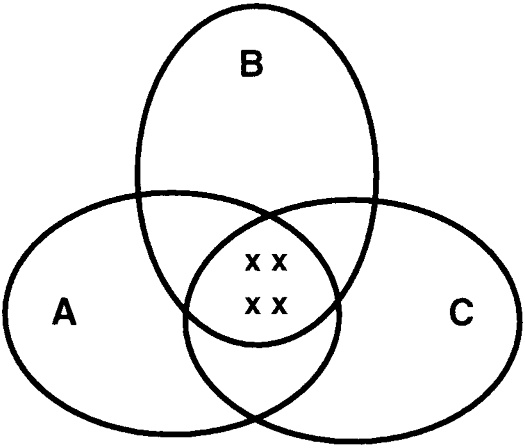

FIG. 7. A conceptual illustration of three hypothesis spaces given the information provided by the cues A, B, and C. The "x"s correspond to hypotheses that are consistent with all three cues.

outside the domain of speech segmentation. During language development, children readily learn aspects of their language for which traditional theories suggest that evidence in the input is degenerate or nonexistent (e.g. Crain, 1991). The classical answer to this problem of the "poverty of the stimulus" is to assume that knowledge of these aspects of language is not learned, but rather form a specifically linguistic innate endowment prewired into the child before birth. Our results suggest that the value of this answer may diminish when hitherto ignored statistical properties of the input and learning mechanisms capable of integrating such properties are taken into account. The networks used in our simulations were not specifically prewired for the detection of word boundaries, instead the architecture of the SRN has a bias towards the learning of highly structured sequential information. This bias forced the net to focus on the relevant aspects of the input signal, indicating that the question of how the child knows which aspects of the signal to pay attention to may not be a serious one. Finally, our analyses suggest how combinations of unreliable cues become reliable, constraining each other through mutual interaction.

While our results pertain to the specific task of word segmentation, we submit that the same principles are likely to support the learning of other kinds of linguistic structure as well. This hypothesis is supported by the growing number of studies finding potential cues to learning of higher level language phenomena, for example, grammatical category (Kelly, 1992); clause structure (Hirsh-Pasek, Kemler Nelson, Jusczyk, Wright Cassidy, Druss, & Kennedy, 1987; grammatical function (Grimshaw, 1983; and argument structure (Pinker, 1989). Adults have been shown also to integrate multiple sources of probabilistic information when computing linguistic representations (e.g. MacDonald et al., 1994; Trueswell & Tanenhaus, 1994). Thus, it would appear that there is an abundance of cues that an infant may integrate in the process of overcoming the *apparent* poverty of the stimulus and that adults use in normal processing. What we have shown is that combining such cues allows for an interaction which in itself is an important source of information. Until we have exhausted the possibilities of such integration processes as the basis for learning "linguistic structures for which there is no evidence", it would seem premature and ill-advised to assume that knowledge thereof must necessarily be innate.

REFERENCES

Aijmer, K., & Altenberg, B. (Eds). (1991). *English corpus linguistics*. New York: Longman.

Allen, J., & Christiansen, M.H. (1996). Integrating multiple cues in word segmentation: A connectionist model using hints. In *Proceedings of the 18th annual Cognitive Science Society conference* (pp. 370–375). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Aslin, R.N., Woodward, J.Z., LaMendola, N.P., & Bever, T.G. (1996). Models of word segmentation in fluent maternal speech to infants. In J.L. Morgan & K. Demuth (Eds), *From signal to syntax* (pp. 117–134). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 157–193). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Bernstein-Ratner, N. (1987). The phonology of parent–child speech. In K. Nelson & A. van Kleeck (Eds.), *Children's language* (Vol. 6, pp. 159–174). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Berry, D.C., & Dienes, Z. (1993). *Implicit learning: Theoretical and empirical issues*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Brent, M.R. (1996). *Lexical acquisition and lexical access: Are they emergent behaviors of a single system?* Paper presented at the ninth annual CUNY Conference on Human Sentence Processing, March 21–23.

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

Brent, M.R., Gafos, A., & Cartwright, T.A. (1994). Phonotactics and the lexicon: Beyond bootstrapping. In E. Clark (Ed.), *Proceedings of the 1994 Stanford Child Language research forum*. Cambridge, UK: Cambridge University Press.

Brill, E. (1993). A corpus based approach to language learning. PhD Dissertation, Department of Computer and Information Science, University of Pennsylvania, PA.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1994). Lexical segmentation: The role of sequential statistics in supervised and un-supervised models. In *Proceedings of the 16th annual conference of the Cognitive Science Society* (pp. 136–141). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Chater, N. (1989). *Learning to respond to structures in time*. Technical Report No. RIPRREP/1000/62/89. Research Initiative in Pattern Recognition, St Andrews Road, Malvern, UK.

Chater, N., & Conkey, P. (1992). Finding linguistic structure with recurrent neural networks. In *Proceedings of the 14th annual meeting of the Cognitive Science Society* (pp. 402–407). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Chomsky, N. (1957). *Syntactic structures*. The Hague, Netherlands: Mouton.

Christiansen, M.H. (in preparation). Recursive sentence structure in connectionist networks.

Christiansen, M.H., & Allen, J. (1997). Coping with variation in speech segmentation. In A. Sorace, C. Heycock & R. Shillcock (Eds.), *Proceedings of the GALA '97 Conference on Language Acquisition: Knowledge Representation and Processing* (pp. 327–332). University of Edinburgh.

Christiansen, M.H., & Chater, N. (1992). Connectionism, meaning and learning. *Connection Science*, *4*, 227–252.

Christiansen, M.H., & Chater, N. (1994). Generalization and connectionist language learning. *Mind and Language*, *9*, 273–287.

Christiansen, M.H., & Chater, N. (in press). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*.

Church, K.W. (1987). Phonological parsing and lexical retrieval. *Cognition*, *25*, 53–69.

Cleeremans, A. (1993). *Mechanisms of implicit learning: Connectionist models of sequence processing*. Cambridge, MA: MIT Press.

Cole, R.A., & Jakimik, J. (1978). How words are heard. In G. Underwood (Ed.), *Strategies of information processing* (pp. 67–117). London: Academic Press.

Cooper, W.E., & Paccia-Cooper, J.M. (1980). *Syntax and speech*. Cambridge, MA: Harvard University Press.

Cottrell, G.W., & Plunkett, K. (1991). Learning the past tense in a recurrent network: Acquiring the mapping from meanings to sounds. In *Proceedings of the 13th annual meeting of the Cognitive Science Society* (pp. 328–333). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Cottrell, G.W., & Tsung, F.-S. (1993). Learning simple arithmetic procedures. *Connection Science*, *5*, 37–58.

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, *14*, 601– 699.

Cutler, A. (1994). Segmentation problems, rhythmic solutions. *Lingua*, *92*, 81–104.

Cutler, A. (1996). Prosody and the word boundary problem. In J. Morgan & K. Demuth (Eds), *From signal to syntax* (pp. 87–99). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Cutler, A., & Mehler, J. (1993). The periodicity bias. *Journal of Phonetics*, *21*, 103–108.

de Sa, V.R. (1994). Unsupervised classification learning from cross-modal environmental structure. PhD Dissertation, Department of Computer Science. University of Rochester, New York.

Echols, C.H., & Newport, E.L. (1992). The role of stress and position in determining first words. *Language Acquisition*, *2*, 189–220.

Elman, J.L. (1988). *Finding structure in time*. Technical Report No. CRL-8801. Center for Research in Language, University of California, San Diego, CA.

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

Elman, J.L. (1991). Distributed representation, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–225.

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

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

Fernald, A., & McRoberts, G. (1996). Prosodic bootstrapping: A critical analysis of the argument and the evidence. In J.L. Morgan & K. Demuth (Eds), *From signal to syntax* (pp. 365–388). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., 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.

Fisher, C., & Tokura, H. (1996). Prosody in speech to infants: Direct and indirect acoustic cues to syntactic structure. In J.L. Morgan & K. Demuth (Eds), *From signal to syntax* (pp. 343–363). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Gleitman, L.R., Gleitman, H., Landau, B., & Wanner, E. (1988). Where learning begins: Initial representations for language learning. In F.J. Newmeyer (Ed.), *Linguistics: The Cambridge survey, Vol. 3* (pp. 150–193). Cambridge, UK: Cambridge University Press.

Greenburg, J.H., & Jenkins, J.J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, *20*, 157–177.

Grimshaw, J. (1981). Form, function, and the language acquisition device. In C.L. Baker & J.J. McCarthy (Eds.), *The logical problem of language acquisition* (pp. 165–182). Cambridge, MA: MIT Press.

Hinton, G., McClelland, J.L., & Rumelhart, D.E. (1986). Distributed representations. In D.E. Rumelhart & J.L. McClelland (Eds), *Parallel distributed processing, Vol. I* (pp. 77–109). Cambridge, MA: MIT Press.

Hirsh-Pasek, K., Kemler Nelson D.G., Jusczyk, P.W., Wright Cassidy, K., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for prelinguistic infants. *Cognition*, *26*, 269–286.

Hornstein, N., & Lightfoot, D. (Eds). (1981). *Explanation in linguistics: The logical problem of language acquisition*. New York: Longman.

Jacobsen. R. (1962). *Selected writings: I. Phonological studies*. The Hague, Netherlands: Mouton.

Jusczyk, P.W. (1993). From general to language-specific capacities: The WRAPSA model of how speech perception develops. *Journal of Phonetics*, *21*, 3–28.

Jusczyk. P.W., & Aslin, R.N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *28*, 1–23.

Jusczyk, P.W., Cutler, A., & Redanz, N.J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development*, *64*, 675–687.

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

Kelly, M.H. (1992). Using sound to solve syntactic problems The role of phonology in grammatical category assignments. *Psychological Review*, *99*, 349–364.

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.

MacDonald, M.C., Pearlmutter, N.J., & Seidenberg, M.S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*, 676–703.

MacWhinney, B., & Snow, C. (1990). The child language data exchange system: An update. *Journal of Child Language*, *17*, 457–472.

Marcus, M. (1992). New trends in natural-language processing—statistical natural-language processing. In *Proceedings of the National Academy of Sciences of the United States of America*, *92*, 10052–10059.

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

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

McClelland, J.L. & Elman, J.L. (1986). Interactive processes in speech perception: The TRACE model. In J.L. McClelland, & D.E. Rumelhart, (Eds), *Parallel distributed processing, Vol. II* (pp. 58–121). Cambridge, MA: MIT Press.

Morgan, J.L. (1986). *From simple input to complex grammar*. Cambridge, MA: MIT Press.

Morgan, J.L., & Demuth, K. (1996). Signal to syntax: An overview. In J. Morgan & K. Demuth (Eds), *From signal to syntax* (pp. 1–22). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Morgan, J.L., Meier, R.P., & Newport, E.L. (1987). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology*, *19*, 498–550.

Morgan, J.L., & Newport, E.L. (1981). The role of constituent structure in the induction of an artificial language. *Journal of Verbal Learning and Verbal Behavior*, *20*, 67–85.

Morgan. J.L., & Saffran, J.R. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development*, *66*, 911–936.

Morgan, J.L., Shi, R., & Allopenna, P. (1996). Perceptual bases of rudimentary grammatical categories: Toward a broader conceptualization of bootstrapping. In J. Morgan & K. Demuth (Eds), *From signal to syntax* (pp. 263–281). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Norris, D.G. (1993). Bottom-up connectionist models of "interaction". In G. Altmann & R. Shillcock (Eds), *Cognitive models of speech processing*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Norris, D.G. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 189–234.

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

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length principle. *Annals of Statistics*, *11*, 416–431.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by error propagation. In J.L. McClelland & D.E. Rumelhart (Eds), *Parallel distributed processing, Vol. 1* (pp. 318–362). Cambridge, MA: MIT Press.

Rumelhart, D.E., & McClelland, J.L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1*. Cambridge, MA: MIT Press.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical cues in language acquisition: Word segmentation by infants. In *Proceedings of the 18th annual Cognitive Science Society conference* (pp. 376–380). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Saffran, J.R., Newport, E.L., & Aslin, R.N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.

Servan-Schreiber, D., Cleeremans, A., & McClelland, J.L. (1989). Learning sequential structure in simple recurrent networks. In D. Touretzky (Ed.), *Advances in neural information processing systems, Vol. 1* (pp. 643–653). Palo Alto, CA: Morgan Kaufman.

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

Shillcock, R., Lindsey, G., Levy, J., & Chater, N. (1992). A phonologically motivated input representation for the modelling of auditory word perception in continuous speech. In *Proceedings of the 14th annual conference of the Cognitive Science Society* (pp. 408–413). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Trueswell, J.C., & Tanenhaus, M.K. (1994). Towards a lexicalist framework of constraint-based syntactic ambiguity resolution. In C. Clifton, L. Frazier, & K. Rayner (Eds), *Perspectives on sentence processing* (pp. 155–179). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

APPENDIX A

Phonological key:

&	*back*
0	*what*
@	*cot*
A	*answer*
I	*pick*
O	*caught*
U	*took*
V	*but*
a	*father*
e	*mate*
i	*beet*
o	*own*
u	*boot*
3	r as in British *absurd*
9	ng as in *ring*
D	*the*
S	*shoot*
T	*three*
Z	*vision*
b	*bag*
d	*dog*
f	*foo*
g	*girl*
h	*hi*
j	*you*
k	*cow*
l	*lamb*
m	*mum*
n	*no*
p	*pick*
r	*roll*
s	*stop*
t	*top*
v	*value*
w	*wind*
z	*zed*

The Phonemes from the MRC Psycholinguistic Database and Their Feature Representations.

Phoneme	*son.*	*cons.*	*voice*	*nasal*	*degree*	*labial*	*pal.*	*phar.*	*l.lip*	*tongue*	*rad.*
&	1	0	1	0	0	0	0	1	0	0	1
3	1	0	1	0	0	0	0	1	1	0	0
9	0	1	1	1	1	0	0	0	0	0	0
@	1	0	1	0	1	0	0	0	0	0	0
A	1	0	1	1	0	0	0	1	0	0	0
D	0	1	1	0	0	0	1	0	0	0	0
I	1	0	1	0	0	0	0	0	0	0	0
O	1	0	1	0	0	0	0	0	1	0	0
0	1	0	1	0	0	0	0	0	1	1	0
S	1	1	0	0	0	0	0	0	0	0	0
T	0	1	0	0	0	0	1	0	0	0	0
U	1	0	1	0	0	0	0	0	1	0.5	0
V	1	0	1	0	0	0	0	0	0	0	0
Z	1	1	1	0	0	0	0	0	0	0	0
a	1	0	1	0	0	0	0	1	0	0	0
b	0	1	1	0	1	1	0	0	1	0	0
d	0	1	1	0	1	0	1	0	0	1	0
e	1	0	1	0	0	0	0	0	0	0	1
f	0	1	0	0	0	0	1	0	1	0	0
g	0	1	1	0	1	0	0	0	0	0	0
h	0	1	0	0	0	0	0	1	0	0	0
i	1	0	1	0	0	0	0	0	0	0	1
j	0	0	1	0	0	0	0	0	0	0	1
k	0	1	0	0	1	0	0	0	0	0	0
l	0	0	1	0	0	0	1	0	0	1	0
m	0	0	1	1	1	1	0	0	1	0	0
n	0	0	1	1	1	0	1	0	0	1	0
o	1	0	1	0	0	0	0	0	1	0	1
p	0	1	0	0	1	1	0	0	1	0	0
r	0	0	1	0	0	0	0	1	1	0	0
s	0	1	0	0	0	0	1	0	0	1	0
t	0	1	0	0	1	0	1	0	0	1	0
u	1	0	1	0	0	0	0		1	0.5	1
v	0	1	1	0	0	0	1	0	1	0	0
w	0	0	1	0	0	1	0	0	1	0	0
z	0	1	1	0	0	0	1	0	0	1	0

Son. = sonorant; cons. = consonantal; pal. = palatal; phar. = pharyngeal; l.lip = lower lip; rad. = radical.

Two mistakes were discovered in the feature coding of the phonemes used in the simulations reported in this article. The consonantal feature of /9/ was coded as "0" and the voice feature of /g/ was coded as "0". Subsequent simulations have confirmed that this mistake did not significantly alter our results.

APPENDIX B

Novel words:

Orthography	*Phonology*	*Stress*	*Orthography*	*Phonology*	*Stress*
adjust	@-dZVst	02	*bogus*	b@U-g@s	20
brassy	brA-sI	20	*bully*	bU-lI	20
cadet	k@-det	02	*consul*	k0n-s@l	20
deadlock	ded-l0k	20	*echo*	e-k@U	20
glamour	gl&-m@	20	*legal*	li-g@l	20
add	&d	2	*age*	eIdZ	2
boot	but	2	*calf*	kAf	2
deck	dek	2	*deep*	dip	2
den	den	2	*fad*	f&d	2
fig	fIg	2	*gap*	g&p	2
geese	gis	2	*host*	h@Ust	2
hulk	hVlk	2	*hut*	hut	2
Jill	dZIl	2	*joke*	j@Uk	2
lace	leIs	2	*lag*	l&g	2
lent	lent	2	*lick*	lIk	2
mail	meIl	2	*match*	mAtS	2
peg	peg	2	*race*	raIs	2
reef	rif	2	*rig*	rIg	2
sag	s&g	2	*saint*	seInt	2
salt	sOlt	2	*slap*	sl&p	2
sold	s@Uld	2	*taint*	tent	2
tool	tul	2	*van*	v&n	2
vile	vaIl	2	*volt*	vOlt	2
weave	wiv	2	*week*	wik	2
weep	wip	2	*zeal*	zil	2

Nonwords:

9OT	gOw	dntr	tIm-gOw	tIm-dntr
Zed	fer	snn	tIm-fer	tIm-snn
3ik	svutp	mjp	tIm-svutp	tIm-mjp
wsVk	skiSD	krmb	tIm-skiSD	tIm-krmb
dl@z	prIbv	spml	tIm-prIbv	tIm-spml
j9at	je3	pskt	tIm-je3	tIm-pskt
zmwAD	swamb	sntT	tIm-swamb	tIm-sntT
Svjublz	trAz9l	tStrl	tIm-trAz9l	tIm-tStrl
bgsUg	vugbk	mlkm	tIm-vugbk	tIm-mlkm
pnEg	wUsTr	rl	tIm-wUsTr	tIm-rl

APPENDIX C

Consonantal sequences occurring at the end of monosyllables and final syllables of stress initial bisyllabic words:

9 S bl blz d dZ dl dn dnt f gl k kl kt l m n ns nt p pl s sl sn t tl tn vn vnt z znt

Consonantal sequences occurring at the end of initial syllables of stress initial bisyllabic words:

9 d f k l m n p s t v