SECTION FOUR

Language processes

CHAPTER ELEVEN

Integrating multiple cues in language acquisition: A computational study of early infant speech segmentation

Morten H. Christiansen and Suzanne Curtin
Cornell University, Ithaca, NY, and University of Pittsburgh, PA, USA

INTRODUCTION

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. In this chapter, we examine systems that have the capacity to extract and store various statistical properties of language. In particular, groups of overlapping, partially predictive cues are increasingly attested 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 integrating different sources of information, including cues that may not be very informative 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 generalizations. Thus, we argue 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 sentence processing (Cottrell, 1989; MacDonald, Pearlmutter, &

Seidenberg, 1994; Trueswell & Tanenhaus, 1994) that emphasize 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 (Seidenberg, 1997).

In the standard learnability approach, language acquisition is viewed in terms of the task of acquiring a grammar (e.g. Pinker, 1994; Gold, 1967). 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. Following Christiansen, Allen, & Seidenberg (1998), 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 he/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 through multiple-cue integration. 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*. We address these issues in the specific context of learning to identify individual words in speech. In the research reported below, 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 remainder of this chapter presents on our work on the modelling of early infant speech segmentation in connectionist networks trained to integrate multiple probabilistic cues. We first describe past work exploring the segmentation abilities of our model (Allen & Christiansen, 1996; Christiansen, 1998; Christiansen et al., 1998). 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. Next, we present results from two new sets of simulations.[1]

The first simulation involves a corpus analysis inspired by the Christiansen

et al. (1998) model, and which provides support for the advantage of integrating multiple cues in language acquisition. In the second simulation, we demonstrate the model's robustness in terms of dealing with noisy input beyond what other segmentation models have been shown capable of dealing with. Finally, we discuss how multiple-cue integration works and how this approach may be extended beyond speech segmentation.

THE SEGMENTATION PROBLEM

Before an infant can even start to learn how to comprehend a spoken utterance, the speech signal must first be segmented into words. Thus, one of the initial tasks confronting the child when embarking on language acquisition involves breaking the continuous speech stream into individual words. Discovering word boundaries is a nontrivial problem, as there are no acoustic correlates in fluent speech to the white spaces that separate words in written text. There are, however, a number of sublexical cues that could potentially be integrated in order to discover word boundaries. The segmentation problem therefore 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, adults tend to make consistent judgements about possible legal sound combinations that could occur in their native language (Greenburg & Jenkins, 1964). This type of phonotactic knowledge may aid in adult segmentation procedures (Jusczyk, 1993). Additionally, evidence from perceptual studies suggests that adults know about and utilize language-specific rhythmic segmentation procedures in processing utterances (Cutler, 1994).

The assumption that children are not born with the knowledge sources that appear to subserve segmentation processes in adults seems reasonable, since they have neither a lexicon nor knowledge of the phonological or rhythmic regularities underlying the words of the particular language being learned. Therefore, one important developmental question concerns how the child comes to achieve steady-state adult behaviour. Intuitively, one might posit that children begin to build their lexicon by hearing words in isolation. A single word strategy, whereby children adopted entire utterances as lexical candidates, would appear to be viable very early in acquisition. In the Bernstein-Ratner (1987) and the Korman (1984) corpora, 22–30% of

child-directed utterances are made up of single words. However, many words, such as determiners, will never occur in isolation. Moreover, this strategy 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/her to detect utterance–internal word boundaries, regardless of whether or not the words appear in isolation. A more realistic suggestion is that a bottom-up process, exploiting sublexical units, allows the child to bootstrap the segmentation process. 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.

Strategies based on prosodic cues (including pauses, segmental lengthening, metrical patterns, and intonation contour) have been proposed as a way of detecting word boundaries (Cooper & Paccia-Cooper, 1980; Gleitman, Gleitman, Landau, & Wanner, 1988). Other recent proposals have focused on the statistical properties of the target language that might be utilized in early segmentation. Considerable attention has been given to lexical stress and sequential phonological regularities—two cues also utilized in the Christiansen et al. (1998) segmentation model. In particular, Cutler and her colleagues (e.g. Cutler & Mehler, 1993) have emphasized 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 preverbal infants supports the claim that infants are sensitive to rhythmic structure and its relationship to lexical segmentation by 9 months (Jusczyk, Cutler & Redanz, 1993). A potentially relevant source of information for determining word boundaries is the phonological regularities of the target language. A study by Jusczyk, Friederici, & Svenkerud (1993) suggests that between 6 and 9 months, infants develop knowledge of phonotactic regularities in their language. Furthermore, there is evidence that both children and adults are sensitive to, and can utilize, such information to segment the speech stream. Work by Saffran, Newport, & Aslin (1996) show that adults are able to use phonotactic sequencing to differentiate between 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 hypothesize possible word boundaries. Further research provides evidence that infants as young as 8 months show the same type of sensitivity after only 3 minutes of exposure (Saffran, Aslin, & Newport, 1996). Thus, children appear to have sensitivity to the statistical regularities of potentially informative 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 issue of when infants are sensitive to particular cues,

and how strong a particular cue is to word boundaries, has been addressed by Mattys, Jusczyk, Luce, & Morgan (1999). They examined how infants would respond to conflicting information about word boundaries. Specifically, Mattys et al. (experiment 4) found that when sequences which had good prosodic information but poor phonotactic cues were tested against sequences that had poor prosodic but good phonotactic cues, the 9 month-old infants gave greater weight to the prosodic information. Nonetheless, the integration of these cues could potentially provide reliable segmentation information, since phonotactic and prosodic information typically align with word boundaries, thus strengthening the boundary information.

Segmenting using multiple cues

The input to the process of language acquisition comprises a complex combination of multiple sources of information. Clusters of such information sources appear to inform the learning of various linguistic tasks (see contributions in Morgan & Demuth, 1996). Each individual source of information, or *cue*, is only partially reliable with respect to the particular task in question. In addition to previously mentioned cues—phontactics and lexical stress—utterance boundary information has also been hypothesized to provide useful information for locating word boundaries (Aslin et al., 1996; Brent & Cartwright, 1996). These three sources of information provide the learner with cues to segmentation. As an example, consider the two unsegmented utterances (represented in orthographic format):

Therearenospacesbetweenwordsinfluentspeech#
Yeteachchildseemstograspthebasicsquickly#

There are sequential regularities found in the phonology (here represented as orthography) which can aid in determining where words may begin or end. The consonant cluster *sp* can be found both at word beginnings (*spaces* and *speech*) and at word endings (*grasp*). However, a language learner cannot rely solely on such information to detect possible word boundaries. This is evident when considering that the *sp* consonant cluster also can straddle a word boundary, as in *cats pyjamas*, and occur word-internally, as in *respect*.

Lexical stress is another useful cue to word boundaries. For example, in English most disyllabic words have a trochaic stress pattern with a strongly stressed syllable followed by a weakly stressed syllable. The two utterances above include four such words: *spaces*, *fluent*, *basics*, and *quickly*. Word boundaries can thus be postulated following a weak syllable. However, this source of information is only partially reliable, as is illustrated by the iambic stress pattern found in the word *between* from the above example.

The pauses at the end of utterances (indicated above by #) also provide

useful information for the segmentation task. If children realize that sound sequences occurring at the end of an utterance always form the end of a word, then they can utilize information about utterance final phonological sequences to postulate word boundaries whenever these sequences occur *inside* an utterance. Thus, knowledge of the rhyme *eech*# from the first example utterance can be used to postulate a word boundary after the similar sounding sequence *each* in the second utterance. As with phonological regularities and lexical stress, utterance boundary information cannot be used as the only source of information about word boundaries, because some words, such as determiners, rarely, if ever, occur at the end of an utterance. This suggests that information extracted from clusters of cues may be used by the language learner to acquire the knowledge necessary to perform the task at hand.

A COMPUTATIONAL MODEL OF MULTIPLE-CUE INTEGRATION IN SPEECH SEGMENTATION

Several computational models of word segmentation have been implemented to address the speech segmentation problem. However, these models tend to exploit solitary sources of information, e.g. Cairns, Shillcock, Chater, & Levy (1997) demonstrated that sequential phonotactic structure was a salient cue to word boundaries, while Aslin, Woodward, LaMendola, & Bever (1996) illustrated that a back-propagation model could identify word boundaries fairly accurately, based on utterance final patterns. Perruchet & Vinter (1998) demonstrated that a memory-based model was able to segment small artificial languages, such as the one used in Saffran, Aslin, & Newport (1996), given phonological input in syllabic format. More recently, Dominey & Ramus (2000) found that recurrent networks also show sensitivity to serial and temporal structure in similar miniature languages. On the other hand, Brent & Cartwright (1996) have shown that segmentation performance can be improved when a statistically-based algorithm is provided with phonotactic rules in addition to utterance boundary information. Along similar lines, Allen & Christiansen (1996) found that the integration of information about phonological sequences and the presence of utterance boundaries improved the segmentation of a small artificial language. Based on this work, we suggest that the integration of multiple probabilistic cues may hold the key to solving the word segmentation problem, and discuss a computational model that implements this solution.

Christiansen et al. (1998) provided a comprehensive computational model of multiple-cue integration in early infant speech segmentation. They employed a simple recurrent network (SRN; Elman, 1990), as illustrated in Fig. 11.1. This network is essentially a standard feedforward network equipped with an extra layer of so-called context units. At a particular time

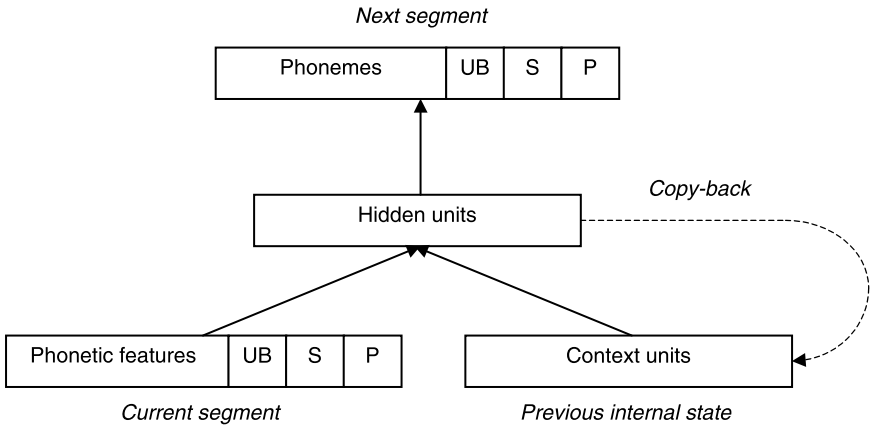

Figure 11.1. Illustration of the SRN used in Christiansen et al. (1998). Arrows with solid lines indicate trainable weights, whereas the arrow with the dashed line denotes the copy-back weights (which are always 1). UB refers to the unit coding for the presence of an utterance boundary. The presence of lexical stress is represented in terms of two units, S and P, coding for secondary and primary stress, respectively. Adapted from Christiansen et al. (1998).

step, t, an input pattern is propagated through the hidden unit layer to the output layer (solid arrows). At the next time step, $t + 1$, the activation of the hidden unit layer at the previous time step, t, is copied back to the context layer (dashed arrow) and paired with the current input (solid arrow). This means that the current state of the hidden units can influence the processing of subsequent inputs, providing a limited ability to deal with integrated sequences of input presented successively.

The SRN model was trained on a *single* pass through a corpus consisting of 8181 utterances of child-directed speech. These utterances were extracted from the Korman (1984) corpus (a part of the CHILDES database; MacWhinney, 2000) consisting of speech directed at pre-verbal infants aged 6–16 weeks. The training corpus consisted of 24,648 words distributed over 814 types and had an average utterance length of 3.0 words (see Christiansen et al., ••••, for further details). A separate corpus, consisting of 927 utterances and with the same statistical properties as the training corpus, was used for testing. Each word in the utterances was transformed from its orthographic format into a phonological form and lexical stress was assigned using a dictionary compiled from the MRC Psycholinguistic Database, available from the Oxford Text Archive.[2]

As input, the network was provided with different combinations of three cues, dependent on the training condition. The cues were: (a) phonology, represented in terms of 11 features on the input and 36 phonemes on the output,[3] (b) utterance boundary information, represented as an extra feature

(UB) marking utterance endings; and (c) lexical stress, coded over two units as either no stress, secondary or primary stress (see Figure 11.1). The network was trained on the *immediate task* of predicting the next phoneme in a sequence, as well as the appropriate values for the utterance boundary and stress units. In learning to perform this task, it was expected that the network would also learn to integrate the cues such that it could carry out the *derived task* of segmenting the input into words.

With respect to the network, the logic behind the derived task is that the end of an utterance is also the end of a word. If the network is able to integrate the provided cues in order to activate the boundary unit at the ends of words occurring at the end of an utterance, it should also be able to generalize this knowledge so as to activate the boundary unit at the ends of words which occur *inside* an utterance (Aslin et al., 1996). Fig. 11.2 shows a snapshot of SRN segmentation performance on the first 37 phoneme tokens in the training corpus. 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. Activations above the mean boundary unit activation for the corpus as a whole (horizontal line) are interpreted as the postulation of a word boundary. As can be seen from the figure, the SRN performed well on this part of the training set, correctly segmenting out all of the 12 words save one (/slipI/ = *sleepy*).

In order to provide a more quantitative measure of performance, accuracy

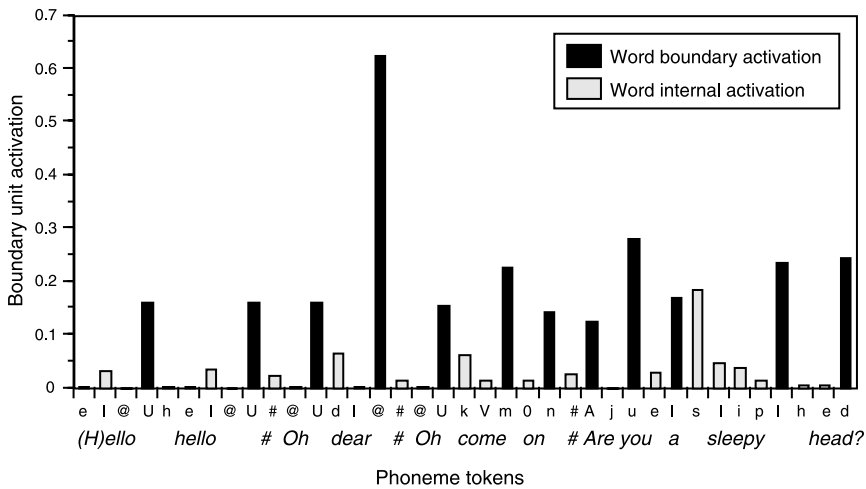

Figure 11.2. The activation of the boundary unit during the processing of the first 37 phoneme tokens in the Christiansen et al. (1998) training corpus. A gloss of the input utterances is found beneath the input phoneme tokens. Adapted from Christiansen et al. (1998).

and completeness scores (Brent & Cartwright, 1996) were calculated for the separate test corpus, consisting of utterances not seen during training:

$$Accuracy = \frac{Hits}{Hits + FalseAlarms}$$

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

Accuracy provides a measure of how many of the words the network postulated were actual words, whereas completeness provides a measure of how many of the actual words the net discovered. Consider the following hypothetical example:

t h e # *d o g* # *s* # *c h a s e* # *t h e c* # *a t*

where # corresponds to a predicted word boundary. Here, the hypothetical learner correctly segmented out two words, *the* and *chase*, but also falsely segmented out *dog, s, thec*, and *at*, thus missing the words *dogs, the*, and *cat*. This results in an accuracy of $\frac{2}{2+4}$ = 33.3% and a completeness of $\frac{2}{2+3}$ = 40.0%.

With these measures in hand, we compare the performance of nets trained using phonology and utterance boundary information—with or without the lexical stress cue—to illustrate the advantage of getting an extra cue. As illustrated by Fig. 11.3, the phon-ub-stress network was significantly more accurate (42.71% vs. 38.67%) and had a significantly higher completeness score (44.87% vs. 40.97%) than the phon-ub network. These results thus demonstrate that having to integrate the additional stress cue with the phonology and utterance boundary cues during learning provides for better performance.

To test the generalization abilities of the networks, segmentation performance was recorded on the task of correctly segmenting novel words. The three-cue net was able to segment 23 of the 50 novel words, whereas the two-cue network was only able to segment 11 novel words. Thus, the phon-ub-stress network achieved a word completeness of 46%, which was significantly better than the 22% completeness obtained by the phon-ub net. These results therefore support the supposition that the integration of three cues promotes better generalization than the integration of two cues. Furthermore, the three-cue net also developed a trochaic bias, and was nearly twice as good at segmenting out novel bisyllabic words with a trochaic stress pattern in comparison to novel words with an iambic stress pattern.

Overall, the simulation results from Christiansen et al. (1998) show that the integration of probabilistic cues forces the networks to develop representations that allow them to perform quite reliably on the task of detecting

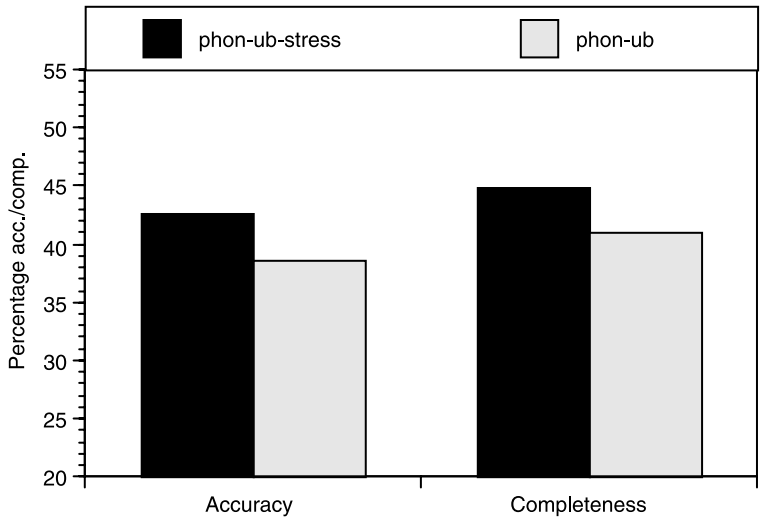

Figure 11.3. Word accuracy (left) and completeness (right) scores for the net trained with three cues (phon-ub-stress; black bars) and the net trained with two cues (phon-ub; grey bars).

word boundaries in the speech stream.[4] This result is encouraging, given that the segmentation task shares many properties with other language acquisition problems that have been taken to require innate linguistic knowledge for their solution, and yet it seems clear that discovering the words of one's native language must be an acquired skill. The simulations also demonstrated how a trochaic stress bias could emerge from the statistics in the input, without having anything like the "periodicity bias" of Cutler & Mehler (1993) built in. Below, in our first simulation, we present a corpus analysis that sheds further light on how the integration of the cues provided by lexical stress and phonology may change the representational landscape to facilitate distributional learning.

SIMULATION 1: STRESS CHANGES THE REPRESENTATIONAL LANDSCAPE

Rhythm is a property of the speech stream to which infants are sensitive at a very young age (Morgan & Saffran, 1995; Jusczyk, 1997; Nazzi, Bertoncini, & Mehler, 1998). Infant research has shown that at age 1–4 months infants are sensitive to changes in stress patterns (Jusczyk & Thompson, 1978). Moreover, English infants have a trochaic bias at age 9 months, yet this preference does not appear to exist at 6 months (Jusczyk, Cutler, & Redanz, 1993), suggesting that at some point during age 6–9 months, infants begin to orientate to the predominant stress pattern of the language. One possible

assumption is that the infant has a rule-like representation of stress that assigns a trochaic pattern to syllables, allowing the infant to take advantage of lexical stress information in the segmentation of speech.

Arguments supporting the acquisition of stress rules are based on child production data that show systematic stages of development across languages and children (Fikkert, 1994; Demuth & Fee, 1995). The consistent nature of stress development supports the postulation of rules in order to account for the production data (Hochberg, 1988). However, the question remains as to what extent this data provides insight into early acquisition processes. We believe that, by drawing on the perceptual and distributional learning abilities of infants, an alternative account emerges, establishing a basis for constraints on stress assignment. We present a corpus analysis investigating how lexical stress may contribute to statistical learning and how this information can help infants group syllables into coherent word units. The results suggest that infants need not posit rules to perform these tasks.

Infants' sensitivity to the distributional (Saffran et al., 1996) and stress-related (Jusczyk & Thompson, 1978) properties of language suggests that infants' exposure to syllables that differ in their acoustic properties (i.e. for lexical stress the change in duration, amplitude, and pitch) may result in differing perceptions of these syllable types. We propose that infants' perceptual differentiation of stressed and unstressed syllables results in a *representational* differentiation of the two types of syllables. This means that the same syllable will be represented differently, depending on whether it is stressed or unstressed. Lexical stress thus changes the representational landscape over which the infants carry out their distributional analysis, and we employ a corpus analysis to demonstrate how this can facilitate the task of speech segmentation.

Simulation details

We used the Korman (1984) corpus that Christiansen et al. (1998) had transformed into a phonologically transcribed corpus with indications of lexical stress. Their training corpus forms the basis for our analyses.[5] We used a whole-syllable representation to simplify our analysis, whereas Christiansen et al. used single phoneme representations.

All 258 bisyllabic words were extracted from the corpus. For each bisyllabic word we created two bisyllabic nonwords. One consisted of the last syllable of the previous word (which could be a monosyllabic word) and the first syllable of the bisyllabic word, and one of the second syllables of the bisyllabic word and the first syllable of the following word (which could be a monosyllabic word). For example, for the bisyllabic word /slipI/ in /A ju eI slipI hed/, we would record the bisyllables /eIsli/ and /pIhed/. We did not record bisyllabic nonwords that straddled an utterance boundary, as they are

not likely to be perceived as a unit. Three bisyllabic words occurred only as single word utterances, and, as a consequence, had no corresponding nonwords. These were therefore omitted from further analysis. For each of the remaining 255 bisyllabic words, we randomly selected a single bisyllabic nonword for a pairwise comparison with the bisyllabic word. Two versions of the 255 word–nonword pairs were created. In one version, the *stress condition*, lexical stress was encoded by adding the level of stress (0–2) to the representation of a syllable (e.g./sli/ → /sli2/). This allows for differences in the representations of stressed and unstressed syllables consisting of the same phonemes. In the second version, the *no-stress condition*, no indication of stress was included in the syllable representations.

Our hypothesis suggests that lexical stress changes the basic representational landscape over which infants carry out their statistical analyses in early speech segmentation. To operationalize this suggestion, we have chosen to use mutual information (MI) as the dependent measure in our analyses. MI is calculated as:

$$MI = \log\left(\frac{P(X,Y)}{P(X)P(Y)}\right)$$

and provides an information theoretical measure of how significant it is that two elements, X and Y, occur together given their individual probabilities of occurrence. Simplifying somewhat, we can use MI to provide a measure of how strongly two syllables form a bisyllabic unit. If MI is positive, the two syllables form a strong unit: a good candidate for a bisyllabic word. If, on the other hand, MI is negative, the two syllables form an improbable candidate for bisyllabic word. Such information could be used by a learner to inform the process of deciding which syllables form coherent units in the speech stream.

Results

The first analysis aimed at investigating whether the addition of lexical stress significantly alters the representational landscape. A pairwise comparison between the bisyllabic words in the two conditions showed that the addition of stress resulted in a significantly higher MI mean for the stress condition ($t(508) = 2.41$, $p < .02$)—see Table 11.1. Although the lack of stress in the no-stress condition resulted in a lower MI mean for the nonwords compared to the stress condition, this trend was not significant, $t(508) = 1.29$, $p > .19$. This analysis thus confirms our hypothesis, that lexical stress benefits the learner by changing the representational landscape in such a way as to provide more information that the learner can use in the task of segmenting speech.

The second analysis investigated whether the trochaic stress pattern provided any advantage over other stress patterns—in particular, the iambic stress pattern. Table 11.2 provides the MI means for words and nonwords for

TABLE 11.1
Mutual information means for words and nonwords in the two stress conditions

Condition	*Words*	*Nonwords*
Stress	4.42	–0.11
No stress	3.79	–0.46

TABLE 11.2
Mutual information means for words and nonwords from the stress condition as a function of stress pattern

Stress pattern	*Words*	*Nonwords*	*No. of words*
Trochaic	4.53	–0.11	209
Iambic	4.28	–0.04	40
Dual	1.30	–1.02	6

the bisyllabic items in the stress condition as a function of stress pattern. The trochaic stress pattern provides for the best separation of words from nonwords, as indicated by the fact that this stress pattern has the largest difference between the *MI* means for words and nonwords. Although none of the differences were significant (save for the comparison between trochaic and dual[6] stressed words, $t(213) = 2.85$, $p < .006$, the results suggest that a system without any built-in bias towards trochaic stress nevertheless benefits from the existence of the abundance of such stress patterns in languages such as English. The results indicate that no prior bias is needed toward a trochaic stress pattern because the presence of lexical stress alters the representational landscape over which statistical analyses are done, such that simple distributional learning devices end up finding trochaic words easier to segment.

The segmentation model of Christiansen et al. (1998) was able to integrate the phonological and lexical stress cues so as to take advantage of the change in the representational landscape that their integration affords. No separate, built-in trochaic bias was needed. Instead, the integration of the three probabilistic cues—phonology, utterance boundary, and lexical stress information—within a single network allowed the trochaic bias to emerge through distributional learning. Of course, both the input to the Christiansen et al. model and the corpus analyses involved idealized representations of speech, abstracting away from the noisy input that a child is faced with in real speech. In the next simulation, we therefore explore the model's ability to segment speech when presented with more naturalistic input, and demonstrate that this type of statistical learning device can in fact cope with noisy input.

SIMULATION 2: COPING WITH CO-ARTICULATION

Ultimately, any model of speech segmentation must be able to deal with the high degree of variation that characterizes natural fluent speech. Our earlier work, as reported above (Allen & Christiansen, 1996; Christiansen, 1998; Christiansen et al., 1998) has established that SRNs constitute viable models of early speech segmentation. These models, like most other recent computational models of speech segmentation (e.g. Aslin et al., 1996; Brent, 1999; Brent & Cartwright, 1996; Perruchet & Vinter, 1998), were provided with idealized input. This is in part due to the use of corpora in which every instance of a word always has the same form (i.e. it is a so-called *citation form*). While this is a useful idealization, it abstracts away from the considerable variation in the speech input that a child is faced with in language acquisition. We therefore now present simulations involving a phonetically transcribed speech corpus that encoded the contextual variation of a word, more closely approximating natural speech. More specifically, we gleaned the adult utterances from the Carterette & Jones (1974) corpus—a part of the CHILDES database (MacWhinney, 2000). These utterances consist of informal speech among American college-aged adults[7].

The goal of the current simulation is to establish whether the success of the word segmentation model discussed here is dependent on the use of the simplified citation form input. Comparisons are made between networks exposed to a corpus incorporating contextual variation (i.e. co-articulation) and networks exposed to a citation form version of the same corpus. If the SRN is to remain a viable model of word segmentation, no significant difference in performance should arise in these comparisons.

Simulation details

The network was provided with the three probabilistic cues, discussed in the previous sections, for possible integration in the segmentation task: (a) *phonology*, represented in terms of an 18 value feature geometry; (b) *lexical stress*, represented as a single separate feature indicating the presence of primary vowel stress; and (c) *utterance boundary information*, represented as a separate feature (UB) which was only activated when pauses occurred in the input.

The simulations involved two training conditions, depending on the nature of the training corpus. In the *co-articulation* condition, the SRN was trained on the phonetically transcribed UNIBET version of the Carterette & Jones corpus. This transcription did not include lexical stress—a cue that contributed significantly to successful SRN segmentation performance in Christiansen et al. (1998). However, lexical stress was indirectly encoded by the use of the reduced vowel *schwa* (/6/ in UNIBET), so we chose to encode all vowels save the *schwa* as bearing primary stress.[8] Utterance boundaries were encoded

whenever a pause was indicated in the transcript. In the *citation form* condition, the SRN was trained on a corpus generated by replacing each word in the orthographic version of the Carterette & Jones corpus with a phonological citation form derived via the Carnegie-Mellon Pronouncing Dictionary (cmudict.0.4)—a machine-readable pronunciation dictionary of North American English which includes lexical stress information. This procedure was similar to the one used to generate training corpora for the models reported in Christiansen et al. (1998). These pronunciations were subsequently translated into UNIBET format. Four vowels which were weakly stressed according to the dictionary were replaced with the UNIBET *schwa* and encoded as stressless, whereas the other vowels were encoded as stressed. Whereas the phonetically transcribed version of the Carterette & Jones corpus included indications where pauses occurred within a single turn, the orthographic version did not include such indications. We therefore counted the number of pauses occurring in each turn in the phonetically transcribed version, and randomly inserted the same number of pauses into the appropriate turn in the citation form version of the corpus.[9]

The overall corpus consisted of 1597 utterances, comprising 11,518 words. Test corpora were constructed by setting aside 10% of the utterances (the same utterances in both training conditions). Thus, the training corpora consisted of 1438 utterances (10,371 words) and the test corpora of 159 utterances (1147 words). In order to provide for more accurate test comparisons between the SRNs trained under the two conditions, utterance boundaries were inserted by hand in the citation form test corpus in the exact same places as found in the co-articulation test corpus. The networks in both training conditions were trained on two passes through their respective training corpora, corresponding to 74,746 sets of weight updates. Identical learning parameters were used in the two training conditions (learning rate, .1; momentum, .95) and the two nets were given the same initial weight randomization within the interval [–.2, .2].

Results

In this simulation, we investigated whether the SRN model of early segmentation could perform as well in the co-articulation condition as in the citation form condition. Fig. 11.4 shows the accuracy and completeness scores for the two networks. The co-articulation SRN obtained an accuracy of 25.27% and a completeness of 37.05%. The citation form SRN reached an accuracy of 24.33% and a completeness of 40.24%. There were no significant differences between the accuracy scores ($\chi^2 = 0.42$, $p > .9$) or the completeness scores ($\chi^2 = 2.46$, $p > .19$). Thus, the SRN model of word segmentation was able to cope

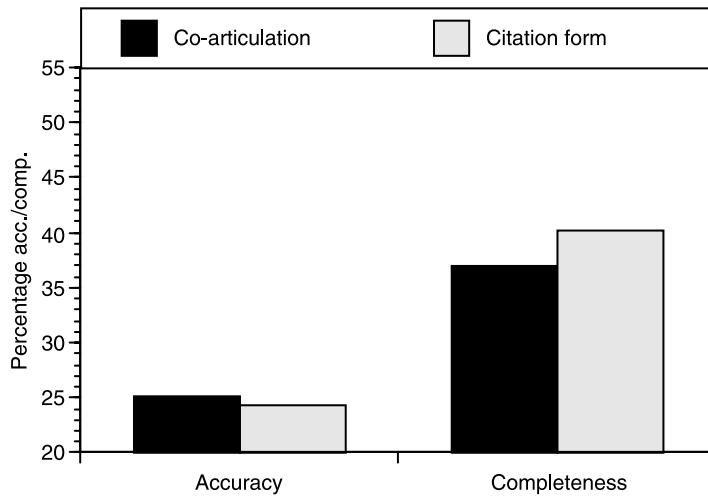

Figure 11.4. Word accuracy (left) and completeness (right) scores for the co-articulation net (black bars) and the citation form net (grey bars).

successfully with variation in the form of co-articulation, suggesting that it provides a good basis for discovering word boundaries in input that is closer to natural speech than the input used in previous computational models.

The results show that our model performs well on the segmentation task—despite being faced with input characterized by considerable variation. This outcome is important, because it demonstrates that the model provides a robust mechanism for the integration of multiple cues, whereas previous models have not been shown to be able to cope satisfactorily with co-articulation. For example, although the connectionist model by Cairns et al. (1997) was trained on a corpus of conversational speech, in which assimilation and vowel reduction had been introduced into the citation forms using a set of rewrite rules, it performed poorly in comparison with the present model (e.g. when pauses were included, their model discovered 32% of the *lexical* boundaries, whereas our model discovered 79% of the lexical boundaries). Our results suggest that connectionist networks provide a useful framework for investigating speech segmentation under less than ideal circumstances. In contrast, it is not clear that other computational frameworks can readily provide the basis for such investigations. For example, statistical optimization models, such as the DR algorithm (Brent & Cartwright, 1996) and the INCDROP model (Brent, 1999), use stored representations of previously encountered lexical units to segment subsequent input. Consequently, these models would end up storing several different phonological versions of the same word in the lexicon if presented with input incorporating co-articulation, as in the above simulation. Likewise, memory-based

segmentation models, such as PARSER (Perruchet & Vinter, 1998), which segments out the longest section of the input that matches a stored unit, would also suffer from similar problems (although the frequency weights attached to such units may provide some relief).

Of course, there is much more to the variation in the speech stream than we have addressed here. For example, the input to our co-articulation nets varied in terms of the individual phonemes making up a word in different contexts, but in real speech co-articulation also often results in featural changes across several segments (e.g. the nasalization of the vowel segment in *can*). Future work must seek to bring the input to segmentation models closer to the actual variations found in fluent speech, and we have sought to take the first steps here.

GENERAL DISCUSSION

In this chapter, we have suggested that the integration of multiple probabilistic cues may be one of the key elements involved in children's acquisition of language. To support this suggestion, we have discussed the Christiansen et al. (1998) computational model of multiple-cue integration in early infant speech segmentation and presented results from three simulations that further underscore the viability of the approach. The corpus analysis in the first simulation showed how lexical stress changes the representational landscape to facilitate word segmentation over a distributional learning device incorporating multiple-cue integration. Previous results obtained from the Christiansen et al. model attest that this model is able to take advantage of such changes in the representational landscape. The second simulation demonstrated that the model is capable of dealing with noisy inputs that are more closely related to the kind of input to which children are exposed. Taken together, we find that the Christiansen et al. model, in combination with the simulations reported here, provide strong evidence in support of multiple-cue integration in language acquisition. In the final part of this chapter, we discuss two outstanding issues with respect to multiple-cue integration—how it works and how it can be extended beyond speech segmentation.

What makes multiple-cue integration work?

We have seen that integrating multiple probabilistic cues in a connectionist network results in more than a just a sum of unreliable parts. But what is it about multiple-cue integration that facilitates learning? The answer appears to lie in the way in which multiple-cue integration can help constrain the search through weight space for a suitable set of weights for a given task (Christiansen, 1998; Christiansen et al., 1998). We can conceptualize the effect that the cue integration process has on learning by considering the

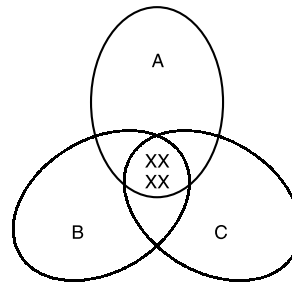

Figure 11.5. An abstract illustration of the reduction in weight configuration space that follows as a consequence of accommodating several partially overlapping cues within the same representational substrate. Adapted from Christiansen et al. (1998).

following illustration. In Fig. 11.5, each ellipse designates for a particular cue the set of weight configurations that will enable a network to learn the function denoted by that cue. For example, the ellipse marked A designates the set of weight configurations that allow for the learning of the function *A* described by the A cue. With respect to the simulations reported above, A, B, and C can be construed as the phonology, utterance boundary, and lexical stress cues, respectively.

If a network using gradient descent learning (e.g. the back-propagation learning algorithm) was only required to learn the regularities underlying, say, the A cue, it could settle on any of the weight configurations in the A set. However, if the net was also required to learn the regularities underlying cue B, it would have to find a weight configuration that would accommodate the regularities of both cues. The net would therefore have to settle on a set of weights from the intersection between A and B in order to minimize its error. This constrains the overall set of weight configurations that the net has to choose between—unless the cues are entirely overlapping (in which case there would not be any added benefit from learning this redundant cue) or are disjunct (in which case the net would not be able to find an appropriate weight configuration). If the net furthermore had to learn the regularities associated with the third cue, C, the available set of weight configurations would be constrained even further.

Turning to the engineering literature on neural networks, it is possible to provide a mathematical basis for the advantages of multiple-cue integration. Here multiple-cue integration is known as *"learning with hints"*, where hints provide additional information that can constrain the learning process (e.g. Abu-Mostafa, 1990; Omlin & Giles, 1992; Suddarth & Holden, 1991). The type of hints most relevant to the current discussion is the so-called *"catalyst hints"* type. This involves adding extra units to a network, such that additional correlated functions can be encoded (in much the same way as the lexical stress units encode a function correlated with the information pro-

vided by the phonological input with respect to the derived task of word segmentation). Thus, catalyst hints are introduced to reduce the overall weight configuration space that a network has to negotiate. This reduction is accomplished by forcing the network to acquire one or more additional related functions, encoded over extra output units. These units are often ignored after they have served their purpose during training (hence the name "catalyst" hint). The learning process is facilitated by catalyst hints because fewer weight configurations can accommodate both the original target function and the additional catalyst function(s). As a consequence of reducing the weight space, hints have been shown to constrain the problem of finding a suitable set of weights, promoting faster learning and better generalization.

Mathematical analyses in terms of the Vapnik-Chervonenkis (VC) dimension (Abu-Mostafa, 1993) and vector field analysis (Suddarth & Kergosien, 1991) have shown that learning with hints may reduce the number of hypotheses a learning system has to entertain. The VC dimension establishes an upper bound for the number of examples needed by a learning process that starts with a set of hypotheses about the task solution. A hint may lead to a reduction in the VC dimension by weeding out bad hypotheses, and reduce the number of examples needed to learn the solution. Vector field analysis uses a measure of "functional" entropy to estimate the overall probability for correct rule extraction from a trained network. The introduction of a hint may reduce the functional entropy, improving the probability of rule extraction. The results from this approach demonstrate that hints may constrain the number of possible hypotheses to entertain, and thus lead to faster convergence.

In sum, these mathematical analyses have revealed that the potential advantage of using multiple-cue integration in neural network training is twofold: First, the integration of multiple cues may reduce learning time by reducing the number of steps necessary to find an appropriate implementation of the target function. Second, multiple-cue integration may reduce the number of candidate functions for the target function being learned, thus potentially ensuring better generalization. As mentioned above, in neural networks this amounts to reducing the number of possible weight configurations that the learning algorithm has to choose between.[10] Thus, because the phonology, utterance boundary, and lexical stress cues designate functions that correlate with respect to the derived task of word segmentation in our simulations, the reduction in weight space not only results in a better representational basis for solving this task, but also leads to better learning and generalization. However, the mathematical analyses provide no guarantee that multiple-cue integration will necessarily improve performance. Nevertheless, this is unlikely to be a problem with respect to language acquisition because, as we shall see next, the input to children acquiring their first

language is filled with cues that reflect important and informative aspects of linguistic structure.

Multiple-cue integration beyond word segmentation

Recent research in developmental psycholinguistics have shown that there is a variety of probabilistic cues available for language acquisition (for a review, see contributions in Morgan & Demuth, 1996). These cues range from cues relevant to speech segmentation (as discussed above) to the learning of word meanings and the acquisition of syntactic structure. We briefly discuss the two latter types of cues here.

Golinkoff, Hirsh-Pasek, & Hollich (1999) studied word learning in children of 12, 19, and 24 months of age. They found that perceptual salience and social information in the form of eye gaze are important cues for learning the meaning of words. The study also provided some insights into the developmental dynamics of multiple-cue integration. In particular, individual cues are weighted differently at different stages in development, changing the dynamics of the multiple-cue integration process across time. At 12 months, perceptual salience dominates—only names for interesting objects are learned, while other cues need to correlate considerably for successful learning. Seven months later, eye gaze cues come into play, but the children have problems when eye gaze and perceptual salience conflict with each other (e.g. when the experimenter is naming and looking at a perceptually uninteresting object). Only at 24 months has the child's lexical acquisition system developed sufficiently that it can deal with conflicting cues. From the viewpoint of multiple-cue integration, this study thus demonstrates how correlated cues are needed early in acquisition to build a basis for later performance based on individual cues.

There are a variety of cues available for the acquisition of syntactic structure. Phonology not only provides information helpful for word segmentation, but also includes important probabilistic cues to the grammatical classes of words. Lexical stress, for example, can be used to distinguish between nouns and verbs. In a 3000-word sample, Kelly & Bock (1988) found that 90% of the bisyllabic trochaic words were nouns, whereas 85% of the bisyllabic iambic words were verbs (e.g. the homograph *record* has stress on the first syllable when used as a noun, and stress on the second syllable when used as a verb). They furthermore demonstrated that people are sensitive to this cue. More recent evidence shows that people are faster and more accurate at classifying words as nouns or verbs if the words have the prototypical stress patterns for their grammatical class (Davis & Kelly, 1997). The number of syllables that a word contains also provides information about its grammatical class. Cassidy & Kelly (1991) showed that 3 year-olds are sensitive to the

probabilistic cue that English nouns tend to have more syllables than verbs (e.g. *gorp* tended to be used as a verb, whereas *gorpinlak* tended to be used as noun). Other important cues to noun hood and verb hood in English include differences in word duration, consonant voicing, and vowel types—and many of these cues have also been found in other languages, such as Hebrew, German, French, Russian (see Kelly, 1992, for a review).

Sentence prosody can also provide important probabilistic cues to the discovery grammatical word class. Morgan, Shi, & Allopenna (1996) demonstrated using a multivariate procedure that content and function words can be differentiated with 80% accuracy by integrating distributional, phonetic, and acoustic cues. More recently, Shi, Werker, & Morgan (1999) found that infants are sensitive to such cue differences. Sentence prosody also provides cues to the acquisition of syntactic structure. Fisher & Tokura (1994) used multivariate analyses to integrate information about pauses, segmental variation and pitch and obtained 88% correct identification of clause boundaries. Other studies have shown that infants are sensitive to such cues (see Jusczyk, 1997, for a review). Additional cues to syntactic structure can be derived through distributional analyses of word combinations in everyday language (e.g. Redington, Chater, & Finch, 1998), and from semantics (e.g. Pinker, 1989).

As should be clear from this short review, there are many types of probabilistic information readily available to the language learner. We suggest that integrating these different types of information, similarly to how the segmentation model was able to integrate phonology, utterance boundary, and lexical stress information, is also likely to provide a solid basis for learning aspects of language beyond speech segmentation. Indeed, a recent set of simulations inspired by the modelling described here have demonstrated that the learning of syntactic structure by an SRN is facilitated when it is allowed to integrate phonological and prosodic information in addition to distributional information in a small artificial language (Christiansen & Dale, 2001). Specifically, an analysis of network performance revealed that learning with multiple-cue integration resulted in faster, better, and more uniform learning. The SRNs were also able to distinguish between relevant cues and distracting cues, and performance did not differ from networks that received only reliable cues. Overall, these simulations offer additional support for the multiple-cue integration hypothesis in language acquisition. They demonstrate that learners can benefit from multiple cues, and are not distracted by irrelevant information. Moreover, this work has recently been scaled up to deal with actual child-directed speech (Reali, Christiansen, & Monaghan, in press).

CONCLUSION

In this chapter, we have presented a number of simulation results that demonstrate how multiple-cue integration in a connectionist network, such as the SRN, can provide a solid basis for solving the speech segmentation problem. We have also discussed how the process of integrating multiple cues may facilitate learning, and have reviewed evidence for the existence of a plethora of probabilistic cues for the learning of word meanings, grammatical class, and syntactic structure. We conclude by drawing attention to the kind of learning mechanism needed for multiple-cue integration.

It seems clear that connectionist networks are well suited for accommodating multiple-cue integration. First, our model of the integration of multiple cues in speech segmentation was implemented as an SRN. Second, and perhaps more importantly, the mathematical results regarding the advantages of multiple-cue integration were couched in terms of neural networks (although they may also hold for certain other, non-connectionist, statistical learning devices). Third, in the service of immediate tasks, such as encoding phonological information, connectionist networks can develop representations that can then form the basis for solving derived tasks, such as word segmentation. Symbolic, rule-based models, on the other hand, would appear to be ill-equipped for accommodating the integration of multiple cues. First, the probabilistic nature of the various cues is not readily captured by rules. Second, the tendency for symbolic models to separate statistical and rule-based knowledge in dual-mechanism models is likely to hinder integration of information across the two types of knowledge. Third, the inherent modular nature of the symbolic approach to language acquisition further blocks the integration of multiple cues across different representational levels (e.g. preventing symbolic syntax models from taking advantage of phonological cues to word class).

As attested by the other chapters in this volume, connectionist networks have provided important insights into many aspects of cognitive psychology. In particular, connectionism has shown itself to be a very fruitful, albeit controversial, paradigm for research on language (see e.g. Christiansen & Chater, 2001b, for a review; or contributions in Christiansen, Chater, & Seidenberg, 1999; Christiansen & Chater, 2001a). Based on our work reported here, we further argue that connectionist networks may also hold the key to a better and more complete understanding of language acquisition, because they allow for the integration of multiple probabilistic cues.

Notes

1. Parts of the simulation results have previously been reported in conference proceedings (simulation 2, Christiansen & Curtin, 1999; and simulation 1, Christiansen & Allen, 1997).

2. Note that these phonological *citation forms* were unreduced (i.e. they do not include the reduced vowel *schwa*). The stress cue therefore provides additional information not available in the phonological input.
3. Phonemes were used as output in order to facilitate subsequent analyses of how much knowledge of phonotactics the net had acquired.
4. These results were replicated across different initial weight configurations and with different input/output representations.
5. Christiansen et al. (1998) represented function words as having primary stress, based on early evidence suggesting that there is little stress differentiation of content and function words in child-directed speech (Bernstein-Ratner, 1987). More recently, Shi, Werker, & Morgan (1999) have found evidence in support of such differentiation. However, for simplicity we have retained the original representation of function words as having stress.
6. According to the Oxford Text Archive, the following words were coded as having two equally stressed syllables: *upstairs, inside, outside, downstairs, hello*, and *seaside*.
7. It would, of course, have been desirable to use child-directed speech as in Christiansen et al. (1998), but it was not possible to find a corpus of phonetically transcribed child-directed speech.
8. This idealization is reasonable, because most monosyllabic words are stressed and because most of the weak syllables in the multisyllabic words from the corpus involved a *schwa*. Further support for this idealization comes from the fact that the addition of vowel stress implemented in this manner significantly improved performance, compared to a training condition in which no stress information was provided.
9. Note that the random insertion of utterance boundaries may lead to the occurrence of utterance boundaries where they often do not occur normally (not even as pauses), e.g. after determiners. Because the presence of pauses in the input is what leads the network to postulate boundaries between words, this random approach is more likely to improve rather than impair overall performance, and thus will not bias the results in the direction of the co-articulation training condition.
10. It should be noted that the results of the mathematical analyses apply independently of whether the extra catalyst units are discarded after training (as is typical in the engineering literature) or remain a part of the network, as in the simulations presented here.

REFERENCES

Abu-Mostafa, Y. S. (1990). Learning from hints in neural networks. *Journal of Complexity, 6,* 192–198.

Abu-Mostafa, Y. S. (1993). Hints and the VC Dimension. *Neural Computation, 5,* 278–288.

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

Altmann, G. T. M., & Dienes, Z. (1999). Rule learning by seven-month-old infants and neural networks. *Science, 284,* 875.

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.), *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). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning, 34,* 71–106.

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

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

Carterette, E., & Jones, M. (1974). *Informal speech: alphabetic and phonemic texts with statistical analyses and tables*. Berkely, CA: University of California Press.

Cassidy, K. W., & Kelly, M. H. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language, 30,* 348–369.

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

Chomsky, N. (1986). *Knowledge of Language*. New York: Praeger.

Christiansen, M. H. (1998). Improving learning and generalization in neural networks through the acquisition of multiple related functions. In J. A. Bullinaria, D. G. Glasspool, & G. Houghton (Eds.), *Proceedings of the Fourth Neural Computation and Psychology Workshop: Connectionist representations* (pp. 58–70). London: Springer-Verlag.

Christiansen, M. H., & Allen, J. (1997). Coping with variation in speech segmentation. In A. Sorace, C. Heycock, & R. Shillcock (Eds.), *Proceedings of GALA 1997: Language acquisition: Knowledge representation and processing* (pp. 327–332). Edinburgh: University of Edinburgh Press.

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

Christiansen, M. H., & Chater, N. (Eds.) (2001a). *Connectionist psycholinguistics*. Westport, CT: Ablex.

Christiansen, M. H., & Chater, N. (2001b). Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences, 5,* 82–88.

Christiansen, M. H., Chater, N., & Seidenberg, M. S. (Eds.) (1999). Connectionist models of human language processing: Progress and prospects. *Cognitive Science, 23* (4, Special Issue), 415–634.

Christiansen, M. H., Conway, C. M., & Curtin, S. (2000). *A connectionist single-mechanism account of rule-like behavior in infancy*. Submitted for presentation at the 22nd Annual Conference of the Cognitive Science Society, Philadelphia, PA.

Christiansen, M. H., & Curtin, S. (1999). The power of statistical learning: No need for algebraic rules. In *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 114–119). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Christiansen, M. H., & Dale, R. A. C. (2001). Integrating distributional, prosodic and phonological information in a connectionist model of language acquisition. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 220–225). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review, 100,* 589–608.

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

Cottrell, G. W. (1989). *A connectionist approach to word sense disambiguation*. London: Pitman.

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

Cutler, A. (1996). Prosody and the word boundary problem. In J. L. 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.

Davis, S. M., & Kelly, M. H. (1997). Knowledge of the English noun–verb stress difference by native and nonnative speakers. *Journal of Memory and Language, 36,* 445–460.

Demuth, K., & Fee, E. J. (1995). *Minimal words in early phonological development*. Unpublished manuscript, Brown University and Dalhousie University.

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

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

Elman, J. (1999). *Generalization, rules, and neural networks: A simulation of Marcus et al.* Unpublished manuscript, University of California at San Diego, CA.

Fikkert, P. (1994). *On the acquisition of prosodic structure* ••••: Holland Institute of Generative Linguistics.

Fischer, C., & Tokura, H. (1996). Prosody in speech to infants: Direct and indirect acoustic cues to syntactic structure. In J. L. Morgan, & K. Demuth (Eds.), *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.

Gold, E. M. (1969). Language identification in the limit. *Information and Control, 10,* 447–474.

Golinkoff, •., Hirsh-Pasek, •., & Hollich, •. (1999). In J. L. Morgan, & K. Demuth (Eds.), *Signal to syntax* (pp. 305–329). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Hochberg, J. A. (1988). Learning Spanish stress. *Language, 64,* 683–706.

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. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

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 & Language, 32,* 402–420.

Jusczyk, P. W., & Thompson, E. (1978). Perception of a phonetic contrast in multisyllabic utterances by two-month-old infants. *Perception & Psychophysics, 23,* 105–109.

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

Kelly, M. H., & Bock, J. K. (1988). Stress in time. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 389–403.

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

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

MacWhinney, B. (2000). *The CHILDES Project* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning in seven month-old infants. *Science, 283,* 77–80.

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

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38,* 465–494.

Morgan, J. L., & Demuth, K. (Eds.) (1996). *From signal to syntax*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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. L. Morgan, & K. Demuth (Eds.), *From signal to syntax* (pp. 263—281). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Towards an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 1–11.

Omlin, C., & Giles, C. (1992). Training second-order recurrent neural networks using hints. In D. Sleeman, & P. Edwards (Eds.), *Proceedings of the Ninth International Conference on Machine Learning* (pp. 363–368). San Mateo, CA: Morgan Kaufmann.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal and Memory and Language, 39,* 246–263.

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

Pinker, S. (1991). Rules of language. *Science, 253,* 530–535.

Pinker, S. (1994). The language instinct: How the mind creates language. New York: William Morrow.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building. *Cognition, 48,* 21–69.

Reali, F., Christiansen, M. H., & Monaghan, P. (in press). Phonological and distributional cues in syntax acquisition: Scaling up the connectionist approach to multiple-cue integration. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22,* 425–469.

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

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barruego, S. (1997). Incidental language learning—listening (and learning) out of the corner of your ear. *Psychological Science, 8,* 101–105.

Seidenberg, M. S. (1995). Visual word recognition: An overview. In P. D. Eimas, & J. L. Miller (Eds.), *Speech, language, and communication. Handbook of perception and cognition* (2nd ed., Vol. 11). San Diego, CA: Academic Press.

Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science, 275,* 1599–1603.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523–568.

Shastri, L., & Chang, S. (1999). *A spatiotemporal connectionist model of algebraic rule-learning* (TR-99–011). Berkeley, CA: International Computer Science Institute.

Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition, 72,* B11–B21.

Shultz, T. (1999). Rule learning by habituation can be simulated by neural networks. In *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 665–670). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Suddarth, S. C., & Holden, A. D. C. (1991). Symbolic-neural systems and the use of hints for developing complex systems. *International Journal of Man–Machine Studies, 35,* 291–311.

Suddarth, S. C., & Kergosien, Y. L. (1991). Rule-injection hints as a means of improving network performance and learning time. In L. B. Almeida, & C. J. Wellekens (Eds.), *Proceedings of the Networks/EURIP Workshop 1990* (Lecture Notes in Computer Science, Vol. 412, pp. 120–129). Berlin: Springer-Verlag.

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.