Multiple-Cue Integration in Language Acquisition: A Connectionist Model of Speech Segmentation and Rule-like Behavior

Morten H. Christiansen

Cornell University

Christopher M. Conway

Cornell University

Suzanne Curtin

University of Pittsburgh

Short title: Multiple Cue Integration in Language Acquisition

Address for correspondence:

Morten H. Christiansen
Department of Psychology
Cornell University
240 Uris Hall
Ithaca, NY 14853
USA
+1 607 255-3570
mhc27@cornell.edu

1. 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 to 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 generalisations. 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 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 (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 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 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 a new set of simulations[i] that extends the coverage of the model to include recent controversial data on purported rule-learning by infants (Marcus, Vijayan, Rao & Vishton, 1999). New empirical predictions concerning the role of segmentation in rule-like behavior is derived from the model, and confirmed by artificial language learning experiments with adult participants. Finally, we discuss how multiple cue integration works and how this approach may be extended beyond speech segmentation.

2. 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 that the child is confronted with 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 sub-lexical cues which 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 byproduct 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 utilise 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 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 sub-lexical 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 utilised in early segmentation. Considerable attention has been given to lexical stress and sequential phonological regularities—two cues also utilised in the Christiansen et al. (1998) segmentation model. In particular, 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 preverbal 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 potentially relevant source of information for determining word boundaries is the phonological regularities of the target language. A recent 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 utilise such information to segment the speech stream. Work by Saffran, Newport & 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. Further research provides evidence that infants as young as 8 months show the same type of sensitivity after only three 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 where 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.

2.1. 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 hypothesised 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 (*<u>sp</u>aces* and *<u>sp</u>eech)* and at word endings (*gra<u>sp</u>*). 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 pajamas*, 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 realise that sound sequences occurring at the end of an utterance always form the end of a word, then they can utilise 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.

3. 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. For example, 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 Figure 1. This network is essentially a standard feed-forward network equipped with an extra layer of so-called context units. At a particular time 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.

[Figure 1 about here]

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, 1991) 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 assigned using a dictionary compiled from the MRC Psycholinguistic Database available from the Oxford Text Archive[ii].

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[iii] (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 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 generalise this knowledge so as to activate the boundary unit at the ends of words which occur *inside* an utterance (Aslin et al., 1996). Figure 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*).

[Figure 2 about here]

In order to provide a more quantitative measure of performance, accuracy 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 that the network postulated were actual words, whereas completeness provides a measure of how many of the actual words that 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 Figure 3, the phon-ub-stress network was significantly more accurate (42.71% vs. 38.67%: $\chi^2 = 18.27, p < .001$) and had a significantly higher completeness score (44.87% vs. 40.97%: $\chi^2 = 11.51, p < .001$) 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.

[Figure 3 about here]

To test the generalisation 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 only was able to segment 11 novel words. Thus, the phon-ub-stress network achieved a word completeness of 46% which was significantly better ($\chi^2 = 4.23, p < .05$) than the 22% completeness obtained by the phon-ub net. These results therefore support the supposition that the integration of three cues promotes better generalisation 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 word boundaries in the speech stream[iv]. This result is encouraging given that the segmentation task shares many properties with other language acquisition problems which 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, we take our approach one step further demonstrating how our model can accommodate recent evidence regarding rule-like behaviour in infancy.

4. Simulation 1: A Multiple-cue Integration Account of Rule-like Behaviour

The nature of the learning mechanisms that infants bring to the task of language acquisition is a major focus of research in cognitive science. With the rise of connectionism, much of the scientific debate surrounding this research has focused on whether rules are necessary to explain language acquisition. All parties in the debate acknowledge that statistical learning mechanisms form a necessary part of the language acquisition process (e.g., Christiansen & Curtin, 1999; Marcus et al., 1999; Pinker, 1991). However, there is much disagreement over whether a statistical learning mechanism is sufficient to account for complex rule-like behaviour, or whether additional rule-learning mechanisms are needed. In the past this debate has primarily taken place within specific areas of language acquisition, such as inflectional morphology (e.g., Pinker, 1991; Plunkett & Marchman, 1993) and visual word recognition (e.g., Coltheart, Curtis, Atkins & Haller, 1993; Seidenberg & McClelland, 1989). More recently, Marcus et al. (1999) have presented results from experiments with 7-month-olds, apparently showing that the infants acquire abstract algebraic rules after two minutes of exposure to habituation stimuli. The algebraic rules are construed as representing an open-ended relationship between variables for which one can substitute arbitrary values, "such as 'the first item X is the same as the third item Y,' or more generally, that 'item I is the same as item J'" (Marcus et al., 1999, p. 79). Marcus et al. further claim that a connectionist single-mechanism approach based on statistical learning is unable to fit their experimental data. In Simulation 1, we present a detailed connectionist model of these infant data, supporting a single-mechanism approach employing multiple-cue integration while undermining the dual-mechanism account.

Marcus et al. (1999) used an artificial language learning paradigm to test their claim that the infant has two mechanisms for learning language. The subjects were seven-month old infants randomly placed in one of two experimental conditions. In the first two experiments, the conditions were ABA or ABB. Each word in the sentence frame ABA or ABB consisted of a consonant and vowel sequence (e.g., *'li wi li'* or *'li wi wi'*). During a two-minute long familiarisation phase the infants were exposed to three repetitions of each of 16 three-word sentences. The test phase in both experiments consisted of 12 sentences made up of words the infants had not previously been exposed to. The test items were broken into 2 groups for both experiments: consistent (items constructed with the same sentence frame as the familiarisation phase) and inconsistent (constructed from the sentence frame the infants were not trained on) — see Table 1. In the second experiment the test items were altered in order to control for an overlap of phonetic features found in the first experiment. This was to prevent the infants from using this type of statistical information. The results of the first and second experiments showed that the infants preferred the inconsistent test items to the consistent ones. In the third experiment, which we focus on in this paper, the ABA grammar was replaced with an AAB grammar. The rationale was to ensure that infants could not distinguish between grammars based solely on reduplication information. Once again, the infants preferred the inconsistent items to the consistent items.

[Table 1 about here]

The conclusion drawn by Marcus et al. (1999) was that a single mechanism that relied on only statistical information could not account for the results because none of the test items appeared in the habituation part of the experiment. Instead they suggested that a dual mechanism was needed, comprising a statistical learning component and an algebraic rule learning component. In addition, they claimed that a SRN would not be able to model their data because of the lack of phonological overlap between habituation and test items. Specifically, they state,

> Such networks can simulate knowledge of grammatical rules only by being trained on all items to which they apply; consequently, such mechanisms cannot account for how humans generalise rules to new items that do not overlap with the items that appeared in training (p. 79).

We demonstrate that SRNs can indeed fit the data from Marcus et al. Other researchers have constructed neural network models specifically to simulate the Marcus et al. results (Altmann & Dienes, 1999; Elman, 1999; Shastri & Chang, 1999; Shultz, 1999). In contrast, we do *not* build a new model to accommodate the results but take the existing SRN model of speech segmentation presented above and show how this model—*without additional modification*—provides an explanation for the results.

The Christiansen et al. (1998) model acquired distributional knowledge about sequences of phonemes, the associated stress patterns, and the occurrence of utterance boundaries. This knowledge allowed it to perform well on the task of segmenting the speech stream into words. We suggest that this knowledge can be put to use in secondary tasks not directly related to speech segmentation—including artificial tasks used in psychological experiments such as Marcus et al. (1999). This suggestion resonates with similar perspectives in the word recognition literature (Seidenberg, 1995) where knowledge acquired for the primary task of learning to read can be used to perform other secondary tasks such as lexical decision.

Marcus et al. (1999) state that they conducted simulations in which SRNs were unable to fit the experimental data. As they do not provide any details of the simulations, we assume (based on other simulations reported by Marcus, 1998) that these focused on some kind of phonological output that the SRNs produced. Given our characterisation of the experimental task as a secondary task, we do not think that the basis for the infants' differentiation between consistent and inconsistent stimuli should be modelled using the phonological output of an SRN. Instead, we focus on the model's ability to integrate the phonological input with utterance boundary information in order to segment out the individual words in the test items.

4.1. Method

Networks. Corresponding to the 16 infants in the Marcus et al. study, we used 16 networks similar to the SRN used in Christiansen et al. (1998) with the exception that the original phonetic feature geometry was replaced by a new representation using 18 features (see Appendix). Each of the 24 SRNs had a different set of initial weights, randomized within the interval [0.25;-0.25]. The learning rate was set to 0.1 and the momentum to 0.95. These training parameters were identical to those used in the original Christiansen et al. model. The networks were trained using the standard back-propagation learning algorithm (Rumelhart, Hinton & Williams, 1986) to predict the next constellation of cues given the current input segment.

Materials. The materials from Experiment 3 in Marcus et al. (1999) were transformed into the phoneme representation used by Christiansen et al. (1998). Two habituation sets were created: one for AAB items and one for ABB items (see Table 1). The habituation sets used here, and in Marcus et al., consisted of three blocks of 16 sentences in random order, yielding a total of 48 sentences in each habituation condition. As in Marcus et al. there were four different test sentences: *'ba ba po'*, *'ko ko ga'* (consistent with AAB); *'ba po po'* and *'ko ga ga'* (consistent with ABB). The test set consisted of three blocks of randomly ordered test sentences, totalling 12 test items. Both the habituation and test sentences were treated as a single utterance with no explicit word boundaries marked between the individual words. The end of each utterance was marked by activating the utterance boundary unit. All habituation and test items were assigned the same level of primary stress.

Procedure. The networks were first trained on a single pass through the Korman (1984) corpus as the original Christiansen et al. model. This corresponds to the fact that the 7-month-olds in the Marcus et al. study already have had a considerable exposure to language, and have begun to develop their speech segmentation abilities (Jusczyk, 1997, 1999). Next, the networks were habituated on a *single* pass through one of the habituation corpora—one phoneme at a time—with learning parameters identical to the ones used during the pretraining on the Korman corpus.

The networks were then tested on the test set (with the weights "frozen") and the activation of the utterance boundary unit was recorded for every phoneme input in the test set for the purpose of scoring the network performance on the derived task. The boundary unit activations across the seven input tokens for each item were separated into two groups according to whether they were recorded for test sentences consistent or inconsistent with the habituation pattern.

For the purpose of measuring word segmentation performance, the mean utterance boundary activation was calculated across all the habituation items for each network. Following Christiansen et al. (1998), a network was said to have postulated a word boundary whenever the boundary unit activation in a test sentence was above its habituation mean cut-off. The word segmentation performance for consistent and inconsistent sentences was then quantified in terms of accuracy and completeness scores (Brent & Cartwright, 1996; Christiansen et al., 1998).

4.2. Results

For each of the sixteen networks, accuracy and completeness scores were computed across all test items, and submitted to the same statistical analyses as used by Marcus et al. for their infant data. The accuracy scores were submitted to a repeated measures ANOVA with condition (AAB vs. ABB) as between network factor and test pattern (consistent vs. inconsistent) as within network factor. The left-hand side of Figure 4 shows the accuracy scores for the consistent and inconsistent items pooled across conditions. There was a main effect of test pattern ($F(1,14)=4.78,\ p < .05$), indicating that the networks segmented significantly more actual words out from the inconsistent items (49.55%) compared to the consistent items (39.44%). Similarly to the infant data, neither the main effect of condition, nor the condition $\times$ test pattern interaction were significant ($F's < 1$). The completeness scores were submitted to a similar analysis, and the results are shown in the right-hand side of Figure 4. Again, there was a main effect of test pattern ($F(1,14)=5.76,\ p < .04$), indicating that the networks were significantly better at segmenting out the words in the inconsistent items (35.76%) compared to the consistent items (28.82%). Neither the main effect of condition, nor the condition $\times$ test pattern interaction were significant ($F's < 1$). The higher accuracy and completeness scores for the inconsistent items suggests that they would stand out more clearly in comparison with the consistent items, and thus explain why the infants looked longer towards the speaker playing the inconsistent items in the Marcus et al. study.

[Figure 4 about here]

Marcus et al. claim that a dual-mechanism system—involving a statistical learning mechanism and a rule-learning mechanism—is needed to account for the infant data. In contrast, Simulation 1 shows that a separate rule-learning component is not necessary to account for the data. This simulation shows how our SRN model of word segmentation can fit the data from Marcus et al. (1999) without invoking explicit rules. The pretraining allowed the SRNs to learn to integrate the regularities governing the phonological, lexical stress, and utterance boundary information in child-directed speech. We suggest that during the habituation phase, the networks then developed weak attractors specific to the habituation pattern *and* the phonology of the syllables used. These attractors will at the same time both attract a consistent item (because of pattern similarity) and repel it (because of phonological dissimilarity), causing interference with the derived task of word segmentation. The inconsistent items, on the other hand, will tend to be repelled by the habituation attractors and therefore do not suffer from the same kind of interference, making them easier for the network to process.

Multiple-cue integration learning enabled the SRN model to fit the infant data. Importantly, the model—as a statistical learning mechanism—can explain both the distinction between consistent and inconsistent items as well as the preference for the inconsistent items. Note that a rule-learning mechanism by itself only can explain how infants may distinguish between items, but not why they prefer inconsistent over consistent items. Extra machinery is needed in addition to the rule-learning mechanism to explain the preference for inconsistent items. Thus, the most parsimonious explanation is that only a statistical learning device is necessary to account for the infant data. The addition of a rule-learning device does not appear to be necessary.

5. Simulation 2: The Role of Segmentation in Rule-like Behavior

Segmentation plays a crucial role in our multiple-cue integration model of the Marcus et al. data. In contrast, the previous accounts of the infants' rule-like behavior do not couch their explanation in terms of such basic components of speech processing. Nevertheless, the previous connectionist models implicitly rely on pre-segmented input to model the infant data. All the models use syllabic input representations, and require that the input be segmented into three-syllable sentences. Sentential segmentation is accomplished outside of the models by way of marking the beginnings and endings of sentences (Altmann & Dienes, 1999; cf. Dienes et al., 1999), by resetting the network before each sentence (Dominey & Ramus, 2000), by only doing error correction after every third syllable (Elman, 1999), or by only having three nodes to encode variable position (Shastri & Chang, 1999) or syllable input (Shultz, 1999). The importance of this pre-segmentation is highlighted, if we make the pauses between words (250 ms) the same length as the pauses between sentences (1000 ms). Leaving sentential segmentation aside, an increase in the time between syllables should have little effect on the performance of the models—except perhaps for the Dominey and Ramus model in which the increased time between syllables may result in an inability to distinguish between consistent and inconsistent items (Dominey, personal communication). However, having same-length gaps between words and sentences are likely to make sentential segmentation harder. If this affects rule-like behavior then it has to be explained *outside* the models by some kind of segmentation device.

Similar considerations apply to learning mechanisms that acquire explicit symbolic rules. Marcus et al. (1999) characterized algebraic rules as representing an open-ended relationship between variables for which one can substitute arbitrary values. Their Experiment 3 was designed to demonstrate that rule-learning is independent of the physical realization of variables in terms of phonological features. The same rule, AAB, applies to—and can be learned from—*'le le we'* and *'ko ko ga'* (with '*le'* and '*ko'* filling the same A slots and *'we'* and *'ga'* the same B slot). As the abstract relationships that this rule represents only pertain to the value of the three variables, the amount of time between them should not affect the application of the rule. Thus, just as the physical realization of a variable does not matter for the learning or application of a rule, neither should the time between variables. The same rule AAB, applies to—and can be learned from—'*le* [250ms] *le* [250ms] *we*' and '*le* [1000ms] *le* [1000ms] *we*' (the '*le*'s should still fill the A slots and the '*we*'s the B slot despite the increased duration of time between the occurrence of these variables). Nevertheless, even though the rule should in principle apply, performance constraints arising *outside* the rule-learning component may prevent it from being retrieved (Marcus, personal communication). Thus, if rule-like behavior is affected by same-length gaps between words and sentences, then a separate segmentation component will be needed.

We expect, however, that this pause manipulation can be accommodated by our multiple-cue integration mechanism model—*without* any need for pre-segmentation machinery. In the model, the preference for inconsistent items is explained in terms of differential segmentation performance. Lengthening the pauses between words, as indicated above, would in effect solve the derived task for the model, and should result in a disappearance of the preference for inconsistent items. Thus, we predict that the model should show no difference between the segmentation performance on the consistent and inconsistent items when pauses between words have the same length as pauses between sentences. To test this prediction, we carried out a new set of simulations.

5.1. Method

Networks. Sixteen SRNs as in Simulation 1.

Materials. Same materials as in Simulation 1 except that utterance boundaries were inserted *between* the words in the habituation and test sentences, simulating a lengthening of pauses between words (from 250 ms to 1000 ms) such that they have the same length as the pauses between utterances.

Procedure. Same procedure as in Simulation 1.

5.2. Results

The completeness scores were submitted to the same analyses as in Simulation 2. As illustrated by Figure 5, the segmentation performance on the test items was improved considerably by the inclusion of utterance boundary-length pauses between words. As predicted, there was no difference between accuracy scores for consistent (74.43%; SE: 6.92) and inconsistent items (72.26%; SE: 7.86) ($F(1,14) = .71$). Neither was there a difference between the completeness scores for consistent (70.14%; SE: 7.622) and inconsistent items (70.49%; SE: 7.966) ($F(1,14) = .02$). As before there were no other effects or interactions (F's < 1), save for an interaction between condition and test pattern for accuracy ($F(1,14) = 5.55,\ p < .04$). This interaction was due to somewhat lower accuracy scores for the inconsistent condition in the AAB habituation pattern.

[Figure 5 about here]

Simulation 2 thus confirms the predicted effect of same-length pauses between words and sentences in the dual-task single-mechanism model. Without including an additional segmentation component, the previous connectionist models would suggest that the pause manipulation should not affect the rule-like behavior[v]. Similarly, learning mechanisms that acquire explicit symbolic rules would need to appeal to segmental performance constraints outside the rule component, in order to make the same predictions; otherwise, the pause manipulation would not be expected to affect rule-learning. To corroborate our model's predictions for the role of segmentation in rule-like behavior, we conducted an artificial language learning experiment using adult subjects.

6. Experiment 1: Replicating the Marcus et al. (1999) Results

Before investigating the role of segmentation in rule-like behavior, we need to first establish whether adults in fact exhibit the same pattern of behavior as the infants in the Marcus et al. study. The first experiment therefore seeks to replicate Experiment 3 from Marcus et al. using adult subjects.

6.1. Method

Participants. Sixteen undergraduate students were recruited from introductory Psychology classes at Southern Illinois University. The participants earned course credit for their participation.

Materials. We used the original stimuli that Marcus et al. (1999) created for their Experiment 3. Each word in a sentence was separated by 250 ms. The 16 habituation sentences for each condition were created by Marcus et al. using the Bell Labs speech synthesizer. The original habituation stimuli were limited to two predetermined sentence orders. To avoid potential order effects, we used the SoundEdit 16 version 2 software for the Macintosh to isolate each sentence as a separate sound file. This allowed us to present the habituation sentences in a random order for each subject.

The stimuli for the test phase consisted of four additional sentences that were either consistent or inconsistent with the training grammar. As mentioned earlier, these sentences contained no phonological overlap with the habituation sentences. Like the habituation stimuli, each word in a sentence was separated by a 250 ms interval. As before, we stored the test stimuli as separate SoundEdit 16 version 2 sound files to allow a random presentation order for each subject.

Procedure. The participants were seated in front of a Macintosh G3 PowerPC equipped with a New Micros button box. Participants were randomly assigned to one of two conditions, AAB or ABB. The experiment was run using the PsyScope presentation software (Cohen, MacWhinney, Flatt, and Provost, 1993) with all stimuli played over stereo loudspeakers at 75dB. The participants were instructed that they were taking part in a pattern recognition experiment. They were told that in the first part of the experiment their task was to listen carefully to sequences of sounds and that their knowledge of these sound sequences would be tested afterwards. Participants listened to three blocks of the 16 randomly presented habituation sentences corresponding either to the AAB or the ABB sentence frame. A 1000 ms interval separated each sentence as was the case in the Marcus et al. experiment.

After habituation, the participants were instructed that they would be presented with new sound patterns that they had not previously heard. They were asked to judge whether a pattern was “similar” or “dissimilar” to what they had been exposed to in the training phase by pressing an appropriately marked button. The instructions emphasized that because the sounds were novel, they should not base their decision on the sounds themselves but instead on the patterns derived from the sounds. The participants listened to three blocks of the four randomly presented test sentences. After the presentation of each test sentence, the participants were prompted for their response. Participants were allowed to take as long as they needed to respond. Each test trial was separated by a 1000 ms interval.

6.2. Results

For the purpose of our analyses, the correct response for consistent items is “similar” while the correct response for inconsistent items is “dissimilar”. The mean overall score for correct classification of test items was 8.81 (SE: 0.63) out of a perfect score of 12. A single-sample t-test showed that this classification performance was significantly better than the chance level performance of 6 ($t(15) = 4.44$, $p < .0005$). The participants' responses were then submitted to the same statistical analysis as the infant data in Marcus et al. (and Simulation 1 and 2 above). Figure 6 (left) shows the mean number of consistent and inconsistent test items that were rated as dissimilar to the habituation items. As expected, there was a main effect of test pattern ($F(1,14) = 18.98$, $p < .001$), such that significantly more inconsistent items were judged as dissimilar (4.5; SE: 0.40) than consistent items (1.69; SE: 0.40). Neither the main effect of condition, nor the condition $\times$ test pattern interaction were significant (F's < 1).

Experiment 1 shows that adults perform similarly to the infants in Marcus et al.'s Experiment 3, thus demonstrating that it is possible to replicate their findings using adult participants instead of infants. This result is perhaps not surprising given that Saffran and colleagues were able to replicate statistical learning results obtained using adults participants (Saffran, Newport & Aslin, 1996) in experiments with 8-month-olds (Saffran, Aslin, et al., 1996). More generally, their results and ours suggest that despite small differences in the experimental methodologies used in infant and adult artificial language learning studies, both methodologies appear to tap into the same learning mechanisms. More generally, one would expect that the same learning mechanisms—statistical or rule-based—would be involved in both infancy and adulthood, and that similar results should be expected in both infant and adult studies with the kind of material used here.

[Figure 6 about here]

7. Experiment 2: Segmentation and Rule-like Behavior

Having replicated the Marcus et al. (Experiment 3) infant data with adult participants, we now turn our attention to the effect of same-length pauses between words and sentences on the learning of rule-like behavior.

7.1. Method

Participants. Sixteen additional undergraduate students were recruited from introductory Psychology classes at Southern Illinois University. The participants earned course credit for their participation.

Materials. The training and test stimuli were the same as in Experiment 1 except that the 250 ms interval between words in a sentence was replaced by a *1000 ms* interval using the SoundEdit 16 version 2 software. The 1000 ms interval between sentences remained the same as before.

Procedure. The procedure and instructions were identical to that used for Experiment 1.

7.2. Results

The mean overall classification score was 5.75 (SE: 0.32) out of 12. This was not significantly different from a chance level performance of 6 ($t < 1$). The responses of the participants were submitted to the same further analysis as in Experiment 1. Figure 6 (right) shows the mean number of consistent and inconsistent items rated as dissimilar. As predicted by Simulation 3, there was *no* main effect of test pattern in this experiment ($F(1,14)=.56$), suggesting that the participants were unable to distinguish between consistent (2.75; SE: 0.17) and inconsistent (2.5; SE: 0.24) items. As in Experiment 1, both the main effect of condition and the interaction between condition and test pattern interaction were not significant ($F's = 0$).

These results show that preference for inconsistent items disappears when the pauses between words and sentences have the same length. This corroborates the prediction from the dual-task, single-mechanism model, underscoring the role of segmentation in rule-like behavior. Crucially, our approach to the Marcus et al. (1999) study as tapping into the derived task of word segmentation, allows the model to make the correct predictions without requiring additional machinery to perform sentential segmentation. The previous connectionist models, on the other hand, appear to require additional sentential segmentation components to account for the results from Experiment 2. This is also true for learning mechanisms that acquire explicit symbolic rules as suggested by Marcus et al. Without appealing to performance limitations arising from processing devices external to the rule-learning component, the lack of difference between consistent and inconsistent items in our artificial learning study cannot be explained. The combination of simulation and experimental results presented here suggest that the multiple-cue integration model provides a compelling account of rule-like behavior in infants and adults.

8. 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. We have also showed through simulations and experiments that the model provides a single mechanism for learning the statistical structure of the speech input, while the representations acquired through multiple cue integration at the same time also allow the model to exhibit rule-like behaviour, previously though to be beyond the scope of SRNs (cf. Marcus et al., 1999). Taken together, we find that the Christiansen et al. model in combination with the simulations and experiments reported here provide strong evidence in support for 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.

8.1. 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 conceptualise the effect that the cue integration process has on learning by considering the following illustration. In Figure 6, 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.

[Figure 7 about here]

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 which 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 minimise 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 disjoint (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"*. 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 provided 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 as well as 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 generalisation.

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 generalisation. As mentioned above, in neural networks this amounts to reducing the number of possible weight configurations that the learning algorithm has to choose between.[vi] 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 resulted in a better representational basis for solving this task, but also lead to better learning and generalisation. 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.

8.2. 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 to 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—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 so 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 3,000 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, and Russian (see Kelly, 1992, for a review).

Sentence prosody can also provide important probabilistic cues to the discovery of 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 one 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 (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.

9. 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 meaning, 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 (though 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 models from taking advantage of phonological cues to word class).

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.

Author Note

Morten H. Christiansen, Department of Psychology, Cornell University; Christopher M. Conway, Department of Psychology, Cornell University; Suzanne Curtin, Department of Linguistics, University of Pittsburgh. Correspondence concerning this article should be addressed to Morten H. Christiansen, Department of Psychology, Cornell University, Ithaca, NY 14853. Electronic mail may be sent via

Internet to mhc27@cornell.edu, cmc82@cornell.edu or to scurtin@pitt.edu. This work was partially supported by a grant from the Human Science Frontier program to MHC.

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.

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.

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.

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.

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.

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). 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. Special issue of *Cognitive Science*, *23 (4),* 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 *The Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 114-119). Mahwah, NJ: Lawrence Erlbaum Associates.

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.

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.

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, (1999).* Unpublished manuscript, University of California, San Diego.

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.

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, U.K.: 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.

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. (1991). *The CHILDES Project*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

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.

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 Publishers.

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 and Company.

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

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: 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, California: 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.

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.

Appendix

The Phonemes from the MRC Psycholinguistics Database and Their Feature Representations

Symbol	*IPA*	*cons.*	*son.*	*labial*	*cor.*	*dorsal*	*front*	*hi*	*low*	*mid*	*tense*	*cont.*	*nasal*	*laminal*	*strid.*	*post.*	*lateral*	*voiced*
&	Q	0	1	0	0	1	1	0	1	0	0	1	0	0	0	0	0	1
3	Œ	0	1	0	0	1	0	0	1	1	0	1	0	0	0	0	0	1
0	Å	0	1	1	0	1	0	0	1	0	0	1	0	0	0	0	0	1
@	´	0	1	0	0	1	0	0	0	1	0	1	0	0	0	0	0	1
A	A	0	1	0	0	1	0	0	1	0	1	1	0	0	0	0	0	1
I	I	0	1	0	0	1	1	1	0	1	0	1	0	0	0	0	0	1
O	ç	0	1	1	0	1	0	0	1	1	1	1	0	0	0	0	0	1
U	U	0	1	0	0	1	0	1	0	1	0	1	0	0	0	0	0	1
V	√	0	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1
a	A	0	1	0	0	1	0	0	1	0	1	1	0	0	0	0	0	1
e	e	0	1	0	0	1	1	0	0	1	1	1	0	0	0	0	0	1
i	i	0	1	0	0	1	1	1	0	0	1	1	0	0	0	0	0	1
o	o	0	1	1	0	1	0	0	0	1	1	1	0	0	0	0	0	1
u	u	0	1	1	0	1	0	1	0	0	1	1	0	0	0	0	0	1
p	p	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b	b	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
t	t	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
d	d	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
k	k	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
g	g	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
f	f	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
v	v	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1
T	T	1	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0
D	D	1	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	1
s	s	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0
z	z	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	1
S	S	1	0	0	1	0	0	0	0	0	0	1	0	0	1	1	0	0
Z	Z	1	0	0	1	0	0	0	0	0	0	1	0	0	1	1	0	1
h	h	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
m	m	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1
n	n	1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1
9	N	1	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1
l	l	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1
r	r	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1
w	w	0	1	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1
j	j	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1

Note. Cons. = consonantal; son. = sonorant; cor. = coronal; cont. = continuant; strid. = strident; post. = posterior.

	Habituation Stimuli	Test Stimuli	
		Consistent	Inconsistent
AAB Condition	*de de di, de de je, de de li, de de we*	*ba ba po*	*ba po po*
	ji ji di, ji ji je, ji ji li, ji ji we	*ko ko ga*	*ko ga ga*
	le le di, le le je, le le li, le le we		
	wi wi di, wi wi je, wi wi li, wi wi we		
ABB Condition	*de di di, de je je, de li li, de we we*	*ba po po*	*ba ba po*
	ji di di, ji je je, ji li li, ji we we	*ko ga ga*	*ko ko ga*
	le di di, le je je, le li li, le we we		
	wi di di, wi je je, wi li li, wi we we		

Table 1. The Habituation and Test Stimuli for the Two Conditions in Marcus et al. (1999).

Figure Captions

Figure 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).

Figure 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).

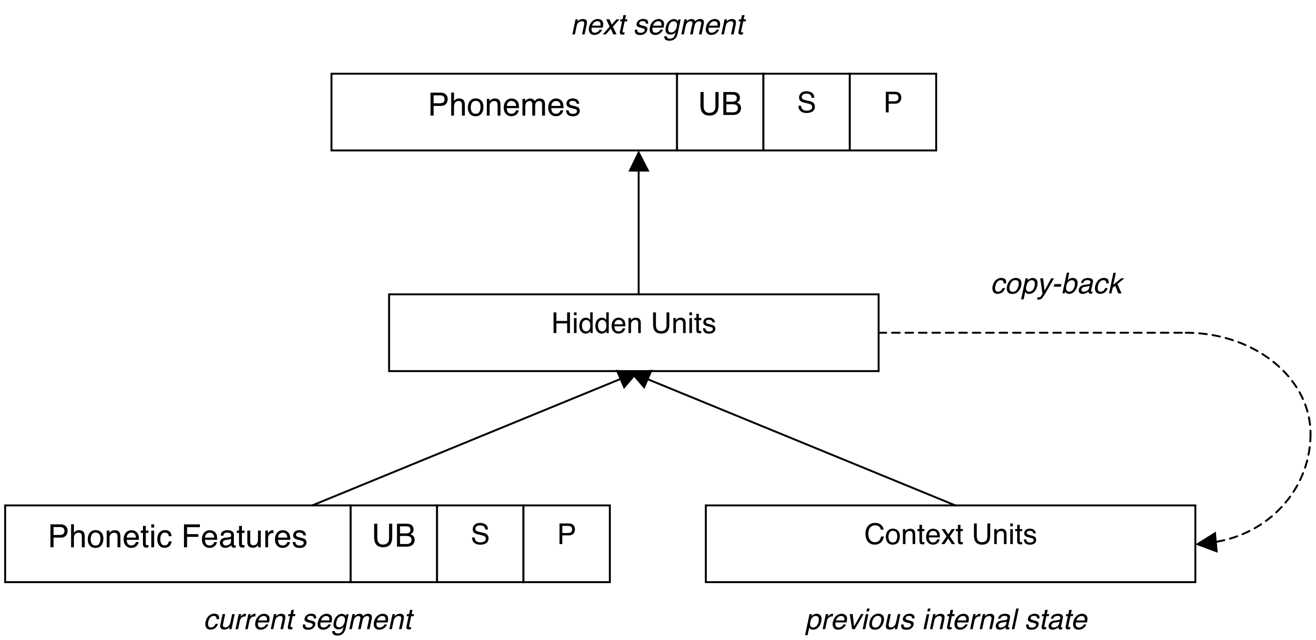
Figure 3. Word accuracy (left) and completeness (right) scores for the net trained with three cues (phon-ub-stress—white bars) and the net trained with two cues (phon-ub—grey bars).

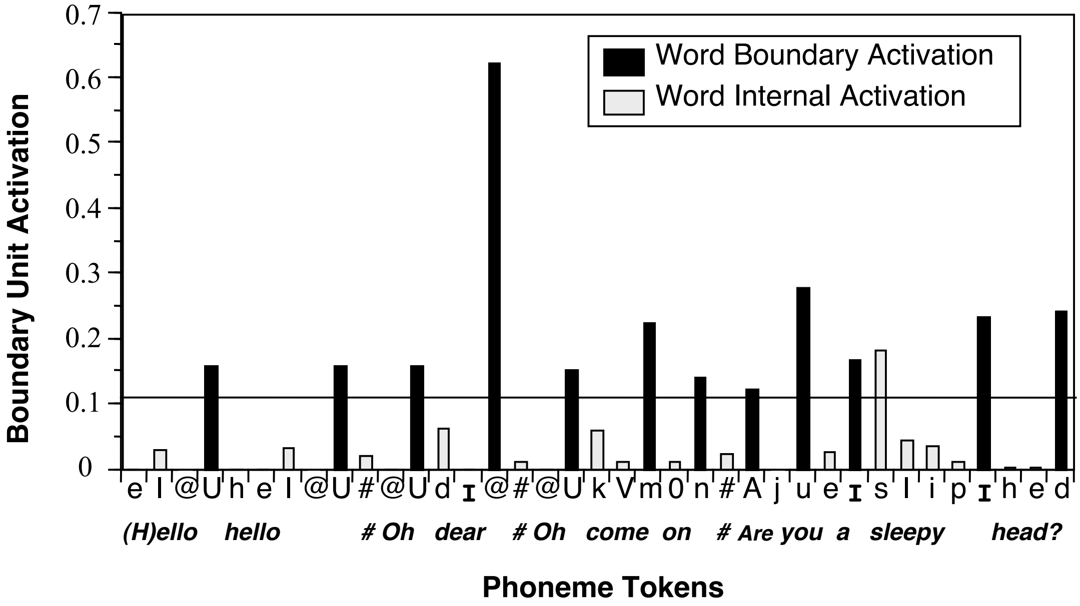
Figure 4. Word accuracy (left) and completeness (right) scores in Simulation 1 for the consistent (white bars) and the inconsistent test items (grey bars).

Figure 5. Word accuracy (left) and completeness (right) scores in Simulation 2 for the consistent (white bars) and the inconsistent test items (grey bars).

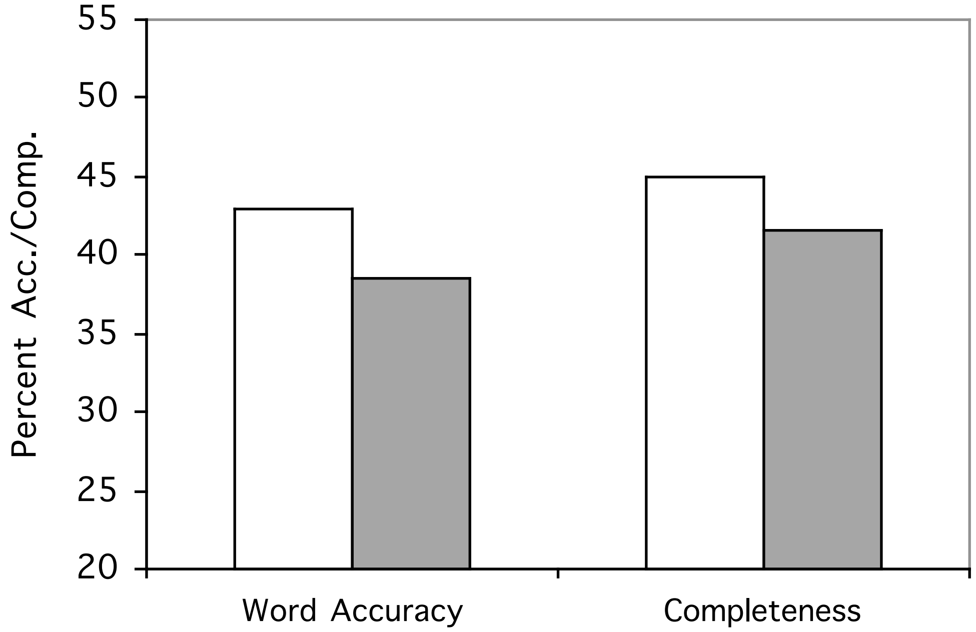
Figure 6. The mean proportion of consistent (white bars) and inconsistent (grey bars) test items rated as dissimilar to the habituation pattern in Experiments 1 (left) and 2 (right).

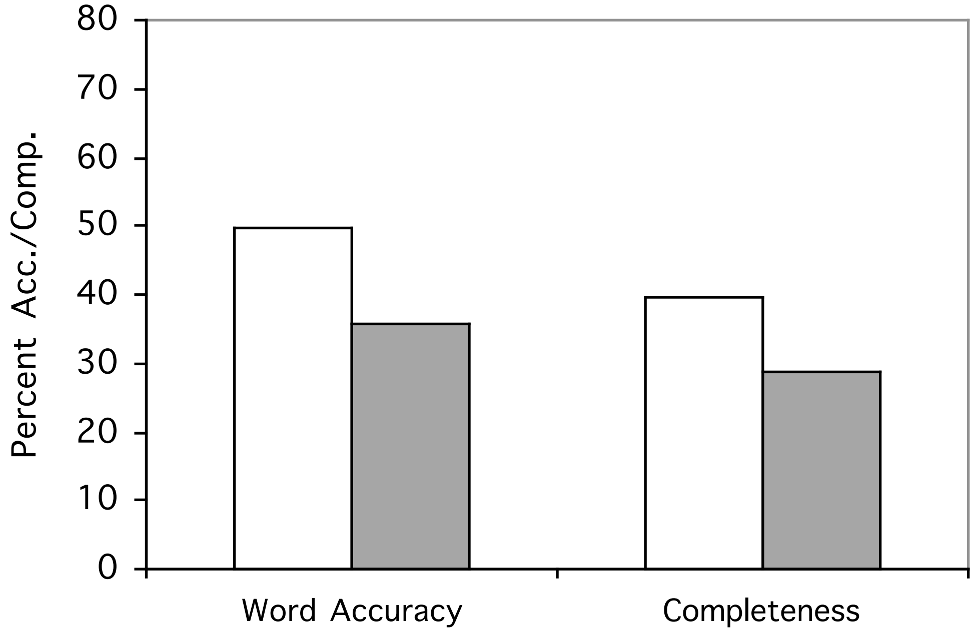
Figure 7. An abstract illustration of the reduction in weight configuration space that follows as a consequence of accommodating several partially overlapping cues within the same representational substrate. (Adapted from Christiansen et al., 1998).


next segment
Phonemes
UB
S
P
copy-back
Hidden Units
Phonetic Features
UB
S
P
Context Units
current segment
previous internal state

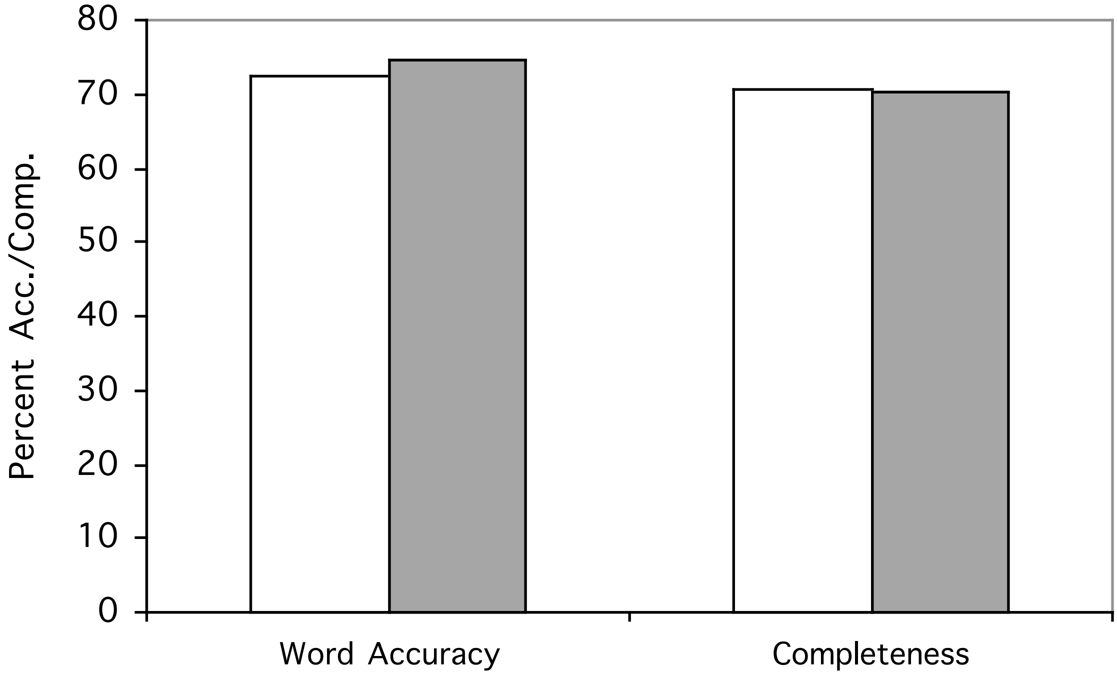


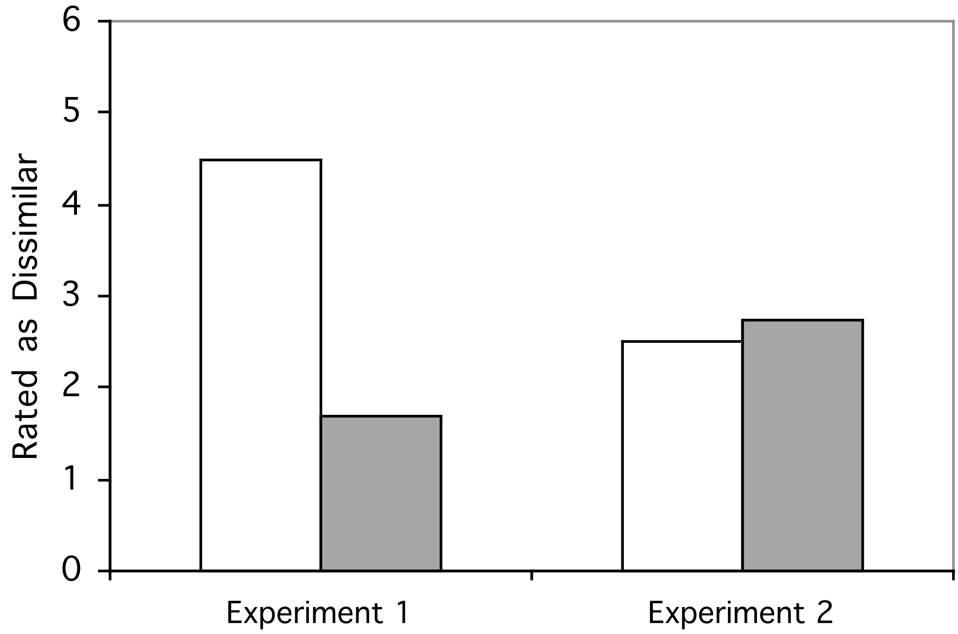
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Boundary Unit Activation
Word Boundary Activation
Word Internal Activation
e l @ U h e l @ U # @ U d ɪ @ # @ U k V m 0 n # A j u e ɪ s l i p ɪ h e d
(H)ello hello # Oh dear # Oh come on # Are you a sleepy head?
Phoneme Tokens

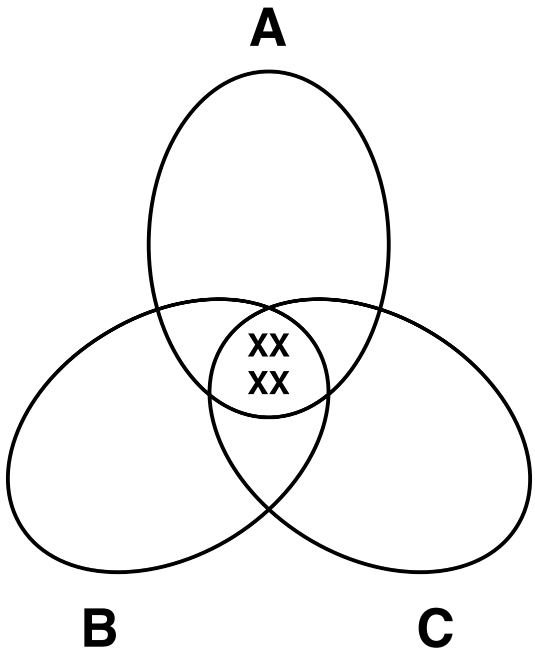

55
50
45
40
35
30
25
20
Percent Acc./Comp.
Word Accuracy
Completeness



80
70
60
50
40
30
20
10
0
Percent Acc./Comp.
Word Accuracy
Completeness


80
70
60
50
40
30
20
10
0
Percent Acc./Comp.
Word Accuracy
Completeness


6
5
4
3
2
1
0
Rated as Dissimilar
Experiment 1
Experiment 2


A
XX
XX
B
C

Footnotes

---

[i] Parts of the simulation results have previously been reported in conference proceedings: Christiansen, Conway & Curtin (2000).

[ii] 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.

[iii] Phonemes were used as output in order to facilitate subsequent analyses of how much knowledge of phonotactics the net had acquired.

[iv] These results were replicated across different initial weight configurations and with different input/output representations.

[v] Even though the Dominey and Ramus (2000) model is predicted to display similar behavior to our dual-task model (Dominey, personal communication), it is nevertheless still vulnerable to this problem because it requires pre-segmented input (i.e., resetting of internal states at the start of each sentence) to account for the original Marcus et al. (1999) results

[vi] 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 the simulations presented here.