The power of statistical learning: No need for algebraic rules

Morten H. Christiansen (MORTEN@SIU.EDU)
Department of Psychology; Southern Illinois University
Carbondale, IL 62901-6502 USA

Suzanne L. Curtin (CURTIN@GIZMO.USC.EDU)
Department of Linguistics; University of Southern California
Los Angeles, CA 90089-1693 USA

Abstract

Traditionally, it has been assumed that rules are necessary to explain language acquisition. Recently, Marcus, Vijayan, Rao, & Vishton (1999) have provided behavioral evidence which they claim can only be explained by invoking algebraic rules. In the first part of this paper, we show that contrary to these claims an existing simple recurrent network model of word segmentation can fit the relevant data without invoking any rules. Importantly, the model closely replicates the experimental conditions, and no changes are made to the model to accommodate the data. The second part provides a corpus analysis inspired by this model, demonstrating that lexical stress changes the basic representational landscape over which statistical learning takes place. This change makes the task of word segmentation easier for statistical learning models, and further obviates the need for lexical stress rules to explain the bias towards trochaic stress patterns in English. Together the connectionist simulations and the corpus analysis show that statistical learning devices are sufficiently powerful to eliminate the need for rules in an important part of language acquisition.

Introduction

One of the basic questions in cognitive science pertains to whether or not explicit rules are necessary to account for complex behavior. Nowhere has the debate over rules been more heated than within the study of language acquisition. Traditionally, generative grammarians have postulated the need for rules in order to account for the patterns found in natural languages (Chomsky & Halle, 1968). In addition, much of the acquisition literature within this framework requires the child to map underlying representations to a surface realization via rules (Smith, 1973; Macken, 1980). On this account, statistical learning is assumed to play little or no role in the acquisition process; instead, abstract rules have been claimed to constitute the fundamental basis of language acquisition and processing. Recently, an alternative approach has emerged emphasizing the role of statistical learning in both the acquisition and processing of language. A growing body of research have explored the power of statistical learning in infancy from both behavioral (e.g., Saffran, Aslin, Newport, 1996) and computational perspectives (e.g., Brent & Cartwright, 1996; Christiansen, Allen & Seidenberg, 1998). This line of research has demonstrated the viability of statistical learning; including cases that were previously thought to require the acquisition of rules and cases for which the input was thought to be too impoverished for learning to take place. In this paper, we extend this research within the area of early infant speech segmentation, providing further evidence against the need for algebraic rules in language acquisition.

Within the traditional rule-based approach Marcus, Vijayan, Rao, & Vishton (1999) have recently presented results from experiments with 7-month-old infants apparently showing that they acquire abstract algebraic rules after two minutes of exposure to habituation stimuli. Marcus et al. further claim that statistical learning models—including the simple recurrent network (SRN; Elman, 1990)—are unable to fit their experimental data. In the first part of this paper we show that knowledge acquired in the service of learning to segment the speech stream can be recruited to carry out the kind of classification task used in the experiment by Marcus et al. For this purpose we took an existing model of early infant speech segmentation (Christiansen et al., 1998) and used it to simulate the results obtained by Marcus et al. Crucially, our simulations do not focus on the phonological output of the network, but rather seek to determine whether the network develops on-line internal representations—that is, transient hidden unit patterns—which can form the basis for reliable classification of input patterns. Stimulus classification then becomes a signal detection problem based on the internal representation, and the preference for one type of stimuli over another is explained in terms of differential segmentation performance. Thus, no rules are needed to account for the data; rather, statistical knowledge related to word segmentation can explain the rule-like behavior of the infants in the Marcus et al. study.

In the second part of the paper we turn our attention to another claim about the necessity of rules in language acquisition. Within the area of the acquisition of lexical stress researchers have debated whether children learn stress by rule or lexically (Hochberg 1988; Klein, 1984). The evidence so far appears to support the claim that children learn stress by rule (Hochberg, 1988) or by setting a parameter in an abstract rule-based system (Fikkert, 1994). In contrast, the segmentation model of Christiansen et al. (1998) acquires lexical stress through statistical learning. The superior performance of the model when provided with lexical stress information, suggests that lexical stress may change the basic representational landscape from which the SRN acquires the statistical regularities relevant for the word segmentation task. We investigate this suggestion through the means of a corpus analysis. The results demonstrate that representational changes caused by lexical stress facilitate learning and obviate the need for rules to explain lexical stress acquisition. Together the results from the corpus analysis and the connectionist simulations suggest that statistical learning is sufficiently powerful to avoid the postulation of abstract rules—at least within the area of speech segmentation.

Rule-Like Behavior without Rules

Marcus et al. (1999) used an artificial language learning paradigm to test their claim that the infant has two mechanisms for learning language, one that uses statistical information and another which uses algebraic rules. They conducted three experiments which tested infants' ability to generalize to items not presented in the familiarization phase of the experiment. They claim that because none of the test items appeared in the habituation part of the experiment the infants would not be able to use statistical information.

The subjects in Marcus et al. (1999) were seven-month old infants randomly placed in an experimental condition. 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 the two-minute long familiarization 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 grammar as the familiarization phase) and inconsistent (constructed from the grammar the infants were not trained on). 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 over 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 over the consistent items.

The conclusion drawn by Marcus et al. (1999) was that a system which relied on statistical information alone could not account for the results. 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 generalize 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. Crucially, we do *not* build a new model to accommodate the results (see Elman, 1999, for a simulation of experiment 2[1]), but take an existing SRN model of speech segmentation (Christiansen et al., 1998) and show how this model—*without additional modification*—provides an explanation for the results.

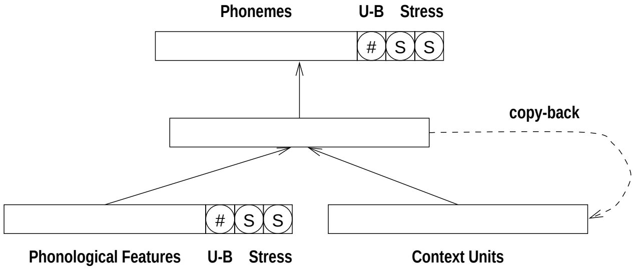

Figure 1: Illustration of the SRN used in Christiansen et al. (1998). Solid lines indicate trainable weights, whereas the dashed line denotes the copy-back weights (which are always 1). U-B refers to the unit coding for the presence of an utterance boundary.

Simulations

The model by Christiansen et al. (1998) was developed as an account of early word segmentation. An SRN 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 of British English speech directed at pre-verbal infants aged 6-16 weeks (a part of the CHILDES database, MacWhinney, 1991). The training corpus consisted of 24,648 words distributed over 814 types (type-token ratio = .03) 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[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 marking utterance endings, and (c) lexical stress coded over two units as either no stress, secondary or primary stress. Figure 1 provides an illustration of the network.

The network was trained on the 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 task of segmenting the input into words.

With respect to the network, the logic behind the segmentation 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

[1] It is not clear that these simulation results can be extended to Experiment 3 because this SRN was trained to activate a unit when reduplication occurred. In Experiment 3, however, both conditions, and therefore both types of test items, contain reduplication and hence cannot be distinguished on the basis of reduplication alone.

[2] Note that these phonological *citation forms* were unreduced (i.e., they did not include the reduced vowel *schwa*). The stress cue therefore provided 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.

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, Woodward, LaMendola & Bever, 1996).

Classification as a Secondary Signal Detection Task

The Christiansen et al. (1998) model acquired distributional knowledge about sequences of phonemes and the associated stress patterns. 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 characterization 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 modeled using the phonological output of an SRN. Instead, it should primarily be based on the internal representations generated during the processing of a sentence. On our account, the differentiation of the two stimulus types becomes a signal detection task involving the internal representation of the SRN (though we shall see below that a part of the *non*-phonological output can explain why the inconsistent items elicited longer looking-times).

Method *Network.* We used the SRN from Christiansen et al. (1998) trained on all three cues.

Materials. The materials from Experiment 3 in Marcus et al. (1999) were transformed into the phoneme representation used by Christiansen et al. Two habituation sets were created in this manner: one for AAB items and one for ABB items. The habituation sets used here, and in Marcus et al., consisted of 3 blocks of 16 sentences in random order, yielding a total of 48 sentences. Each sentence contained 3 monosyllabic nonsense words. As in Marcus et al. there were four different test trials: "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 trials, totaling 12 test sentences. 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 the utterance was marked by activating the utterance boundary unit.

Procedure. The network was habituated by providing it with a *single* pass through the habituation corpus—one phoneme at a time—with learning parameters identical to the ones used originally in Christiansen et al. (1998) (i.e., learning rate = .1 and momentum = .95). The test set was presented to the network (with the weights "frozen") and the hidden unit activation for the final input phoneme in each test sentence was recorded. Given the processing architecture of the SRN, the activation pattern over the hidden units at this point provides a representation of the sentence as a whole; that is, a compressed version of the sequence of hidden unit states that the SRN has gone through during the processing of the sentence. Each hidden unit representation constitutes an 80-dimensional vector.

Result and Discussion We used discriminant analysis (Cliff, 1987; see Christiansen & Chater, in press, for an earlier application to SRNs) to determine whether the hidden unit representations contained sufficient information to distinguish between the consistent and inconsistent items for a given habituation condition. The 12 vectors were divided into two groups depending on whether they were recorded for an AAB or ABB test item. The vectors were entered into a discriminant analysis to determine whether they contained sufficient information to be linearly separated into the relevant two groups. As a control, we randomly re-assigned three vectors from each group to the other group such that our random controls cut across the two original groupings (i.e., both random groups contained three AAB and three ABB vectors).

The results from both the AAB and ABB habituation conditions showed significant separation of the correct vectors ($df = 5, p < .001; df = 6, p < .001$), but not for the random controls ($df = 6, p = .3589; df = 6, p = .4611$). Consequently, it was possible on the basis of the hidden unit representation derived from the model to correctly predict the appropriate group membership of the test items at 100% accuracy in both conditions. However, for the random control items in both conditions the accuracy (83.3%) was not significantly different from chance.

The superficially high classification of the random vectors is due to the high number of hidden units (80) and the low number of test items (6) in each group. This increases the probability that a random variable may provide information that can distinguish between the two random groups by chance. Nonetheless, the significance statistics suggest that only the original correct grouping of hidden unit patterns contain sufficient information for the reliable categorization of the items. This information can be used by the network to distinguish between the consistent and inconsistent test items. Similarly, we argue that infants may have access to same type of information on which they can classify the test items presented to them in the Marcus et al. (1999) study.

Explaining the Preference for Inconsistent Items

The results from the discriminant analyses demonstrate that no algebraic rules are necessary to account for the differential classification of consistent and inconsistent items in Experiment 3 of Marcus et al. (1999). However, the question remains as to why the infants looked longer at the inconsistent items compared to the consistent items. To address this question we looked at the activation of the non-phonological output unit coding for utterance boundaries. Christiansen et al. (1998) used the activation of this unit as an indication of predicted word boundaries. Our prediction for the current simulations was that the SRN should show a differential ability to predict word boundaries for the words in the two test conditions. As in Christiansen et al., we used accuracy and completeness scores (Brent & Cartwright, 1996) as a quantitative measure of segmentation performance.

$$\text{Accuracy} = \frac{\text{Hits}}{\text{Hits + False Alarms}} \quad (1)$$

$$\text{Completeness} = \frac{\text{Hits}}{\text{Hits + Misses}} \quad (2)$$

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 in the test sets that the net discovered. Consider the following hypothetical utterance 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 2/(2+4) = 33.3% and a completeness of 2/(2+3) = 40.0%.

Given these performance measures, Christiansen et al. (1998) found that the network trained with all three cues (phonology, stress and utterance boundary information) achieved an accuracy of 42.71% and a completeness of 44.87%. So, nearly 43% of the words the network segmented out were actual words and it segmented out nearly 45% of the words in the test corpus. We used the same method to compare how well the network segmented the words in the test sentences from Marcus et al. (1999).

Method *Network and Materials.* Same as in the previous simulation.

Procedure. The network habituated in the previous simulation was retested on the test set (with the weights "frozen") and the output for the utterance boundary unit was recorded for every phoneme input. For each habituation condition, the output was divided into two groups dependent on whether the trials were consistent or inconsistent with the habituation. For each habituation condition, the activation of the boundary unit was recorded across all items and the mean activation was calculated. For a given habituation condition, the network was said to have postulated a word boundary whenever the boundary unit activation was above the mean.

Results and Discussion Word boundaries were posited more accurately for the inconsistent items across both conditions (80.00% and 75.00%) than for the consistent items. The scores for word completeness were also higher for the inconsistent items (see Table 1). The results indicate that overall there was better segmentation of the inconsistent items. This suggests that the inconsistent items would stand out more clearly and thus may explain why the infants looked longer towards the speaker playing the inconsistent items in the Marcus et al. (1999) study.

There was a clear effect of habituation on the segmentation performance of the model in the present study compared to the model's performance in Christiansen et al. (1998) where scores were generally lower on both measures. However, in Christiansen et al. the average number of phonemes per word was three, whereas the average number in the current study was only two phonemes per word, thus making the present task easier.

Table 1: Word completeness and accuracy for consistent and inconsistent items in the two habituation conditions.

	AAB Condition		ABB Condition	
	Con.	Incon.	Con.	Incon.
Accuracy	75.00%	80.00%	66.67%	75.00%
Completeness	50.00%	66.67%	44.44%	50.00%

Note. Con. = Consistent items; Incon. = Inconsistent items.

The simulations show how an existing SRN model of word segmentation can fit the data from Marcus et al. (1999) without invoking explicit rules. The SRN had learned to integrate the regularities governing the phonological, lexical stress, and utterance boundary information in child-directed speech. This form of statistical learning enabled it to fit the infant data. In this context, the positive impact of lexical stress information on network performance (as reported in Christiansen et al. 1998) suggests that lexical stress changes the representational landscape over which statistical learning takes place. As we shall see next, this removes the need for lexical stress rules to explain the strong/weak (trochaic) bias in English over weak/strong (iambic).

Taking Advantage of Lexical Stress without Rules

Evidence from infant research has shown that infants between one and four months are sensitive to changes in stress patterns (Jusczyk & Thompson, 1978). Additionally, researchers have found that English infants have a trochaic bias at nine-months of age yet this preference does not appear to exist at six-months (Jusczyk, Cutler & Redanz, 1993). This suggests that at some time between 6 and 9 months of age, infants begin to orient to the predominant stress pattern of the language. One might then assume that if the infant does not have a rule-like representation of stress that assigns a trochaic pattern to syllables, then he/she cannot take advantage of lexical stress information in the segmenting of speech.

The arguments put forth in the literature for rules are based on the production data of children, and based on these productions, it has been shown that word-level (lexical) stress is acquired through systematic stages of development across languages and children (Fikkert, 1994; Demuth & Fee, 1995). If children are learning stress without the use of rules, then systematic stages would not be expected. In other words, due to the consistent patterns of children's productions, a rule must be postulated in order to account for the data (Hochberg, 1988). However, we believe that this conclusion is premature. Drawing on research on the perceptual and distributional learning abilities of infants, 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.

Stress Changes the Representational Landscape: A Corpus Analysis

Infants are sensitive to the distributional (Saffran et al., 1996) and stress related (Jusczyk & Thompson, 1978) properties of language. We suggest that infants' perceptual differentiation of stressed and unstressed syllables result in a *representational* differentiation of the two types of syllables. The same syllable is represented differently depending on whether it is stressed or unstressed. This changes the representational landscape, and we employ a corpus analysis to demonstrate how this facilitates the task of speech segmentation.

Method *Materials.* For the corpus analysis 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. We note that in child-directed speech there appears to be little differentiation in lexical stress between function and content words (at least at the level of abstraction we are representing here; Bernstein-Ratner, 1987; see Christiansen et al. for a discussion). Function words were therefore encoded as having primary stress. We further used a whole syllable representation to simplify our analysis, whereas Christiansen et al. used single phoneme representations.

Procedure. All 258 bisyllabic words were extracted from the corpus. For each bisyllabic word we recorded 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 syllable 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 only occurred 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 chose 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:

$$\mathrm{MI} = \log\left(\frac{P(X,Y)}{P(X)P(Y)}\right) \tag{3}$$

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.

Table 2: 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 3: 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

Results and Discussion 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 2. Although the lack of stress in the no-stress condition resulted in a lower MI mean for the no-stress condition than for 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 away 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 3 provides the MI means for words and nonwords for 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[4] 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 like English. In other words, the results indicate that no prior bias is needed

[4] According to the Oxford Text Archive, the following words were coded as having two equally stressed syllables: *upstairs, inside, outside, downstairs, hello,* and *seaside*.

toward a trochaic stress patterns 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) developed a bias towards trochaic patterns, such that, when segmenting test corpora with either iambic or trochaic syllable groupings, the model was better at segmenting out words that followed a trochaic pattern. Thus, the SRN acquired the trochaic bias given the change in the distributional landscape that stress provides.

Conclusion

In this paper, we have demonstrated the power of statistical learning in two areas of language acquisition in which abstract rules have been deemed necessary for the explanation of the data. Using an existing model of infant speech segmentation (Christiansen et al., 1998), we first presented simulation results fitting the behavioral data from Marcus et al. (1999). The SRN's internal representations incorporated sufficient information for a correct classification of the test items; and the differential segmentation performance on the stimuli words in the consistent and inconsistent conditions provided an explanation for the inconsistent item preference: They are more salient. No rules are needed to explain these data. We then used a corpus analysis to test predictions from the same model concerning the way lexical stress changes the representational landscape over which statistical analyses are done. These changes result in more information being available to a statistical learner, and provide the basis for the trochaic stress bias in English. Again, no rules are needed to explain these data.

There are, of course, other aspects of language for which we have not shown that rules are not needed. Future research will have to determine whether rules may be needed outside the domain of speech segmentation. Some of our other work (Christiansen & Chater, in press) suggests that rules may not be needed to account for one of the supposedly basic rule-based properties of language: Recursion. But why is statistical learning often dismissed as a plausible explanation of language phenomena? We suggest that this may stem from an impoverished view of statistical learning. For example, Pinker (1999) in his commentary on Marcus et al. (1999) forces statistical learning, and connectionist models in particular, into a behavioristic mold: Only input-output relations are said to matter. However, connectionists have also taken part in the cognitive revolution and therefore posit internal representations mediating between input and output. As we demonstrated in the first part of the paper, hidden unit representations provide an important source of information for the modeling of rule-like behavior. Another oversight relates to the significance of combining several kinds of information within a single statistical learning device. The second part of the paper showed how the addition of lexical stress information to the phonological representations resulted in more information being available for the learner. Thus, a more sophisticated approach to statistical learning is likely to reveal its true power, and may obviate the need for algebraic rules.

References

Aslin, R.N., Woodard, 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*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Brent, M.R. & Cartwright, T.A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61*, 93–120.

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

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

Chomsky, N. & Halle, M. (1968). *The Sound Pattern of English*. New York: Harper and Row.

Cliff, N. (1987). *Analyzing Multivariate Data*. Orlando, FL: Harcourt Brace Jovanovich.

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

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

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

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

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

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

Klein, H. (1984). Learning to stress: A case study. *Journal of Child Language, 11*, 375–390.

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

Macken, M.A. (1980). The child's lexical representation: The "puzzle-puddle-pickle" evidence. *Journal of Linguistics, 16*, 1–17.

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

Marcus, G.F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology, 37*, 243–282.

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

Pinker, S. (1999). Out of the minds of babes. *Science, 283*, 40–41.

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

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

Smith, N.V. (1973). *The Acquisition of Phonology: A case study*. Cambridge: Cambridge University Press.