Improving Learning and Generalization in Neural Networks through the Acquisition of Multiple Related Functions

Morten H. Christiansen
Program in Neural, Informational and Behavioral Sciences
University of Southern California
Los Angeles, CA 90089-2520, U.S.A.

Abstract

This paper presents evidence from connectionist simulations providing support for the idea that forcing neural networks to learn several related functions together results in both improved learning and better generalization. More specifically, if a neural network employing gradient descent learning is forced to capture the regularities of many semi-correlated sources of information within the same representational substrate, it then becomes necessary for it to only represent hypotheses that are consistent with all the cues provided. When the different sources of information are sufficiently correlated the number of candidate solutions will be reduced through the development of more efficient representations. To illustrate this, the paper draws briefly on research in the neural network engineering literature, while focusing on recent work on the segmentation of speech using connectionist networks. Finally, some implications for language acquisition of the present approach are discussed.

1 Introduction

Systems that learn from examples are likely to run into the problem of induction —that is, given any finite set of examples, there will always be a considerable number of different hypotheses consistent with the example set. However, many of these hypotheses may not lead to correct generalization. The problem of induction is pervasive in the domain of cognitive behavior—especially within the field of language acquisition where it has promoted the influential idea that a child must bring a substantial amount of innate linguistic knowledge to the acquisition process in order to avoid false generalizations (e.g., [7]). However, this conclusion may be premature because it is based on a simplistic view of computational mechanisms. Recent developments within connectionist modeling have revealed that neural networks embody a number of computational properties that may help constrain learning processes in appropriate ways.

This paper focuses on one such property, presenting evidence from connectionist simulations that provides support for the idea that forcing neural networks to learn several related functions together results in better learning and generalization. First, learning with hints as applied in the neural network

engineering literature will be discussed. The following section addresses the problem of learning multiple related functions within cognitive domains, using word segmentation as an example. Next, an analysis of how learning multiple functions may help constrain the hypothesis space that a learning system has to negotiate. The conclusion suggests that the integration of multiple partially informative cues may help develop the kind of representations necessary to account for acquisition data which have previously formed the basis for poverty of stimulus arguments against connectionist and other learning-based models of language acquisition.

2 Learning using hints

One way in which the problem of induction may be reduced for a system learning from examples is if it is possible to furnish the learning mechanism with additional information which can constrain the learning process. In the neural network engineering literature, this has come to be known as learning with hints. Hints are ways in which additional information not present in the example set may be incorporated into the learning process [1, 21], thus potentially helping the learning mechanism overcome the problem of induction.

There are numerous ways in which hints may be implemented, two of which are relevant for the purposes of the present paper: (a) The *insertion* of explicit rules into networks via the pre-setting of weights [16]; and (b) the addition of extra *"catalyst"* units encoding additional related functions [20, 21]. The idea behind providing hints in the form of rule insertion is to place the network in a certain part of weight space deemed by prior analysis to be the locus of the most optimal solutions to the training task. The rules used for this purpose typically encode information estimated by prior analysis to capture important aspects of the target function. If the right rules are inserted, it will reduce the number of possible weight configurations that the network has to search through during learning. Catalyst hints are also introduced to reduce the overall weight configuration space that a network has to negotiate, but 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 will be explained in more detail below). As a consequence of reducing the weight space, both types of hints have been shown to constrain the induction problem, promoting faster learning and better generalization.

Mathematical analyses in terms of the Vapnik-Chervonenkis (VC) dimension [2] and vector field analysis [21] 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 solu-

tion. 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 hints in neural network training is twofold: First, hints may reduce learning time by reducing the number of steps necessary to find an appropriate implementation of the target function. Second, hints may reduce the number of candidate functions for the target function being learned, thus potentially ensuring better generalization. As mentioned above, in neural networks this amounts to reducing the number of possible weight configurations that the learning algorithm has to choose between[1]. However, it should be noted that there is no guarantee that a particular hint will improve performance. Nevertheless, in practice this does not appear to pose a major problem because hints are typically carefully chosen to reflect important and informative aspects of the original target function.

From the perspective of language acquisition we can construe rule-insertion hints as analogous to the kind of innate knowledge prescribed by theories of Universal Grammar (e.g., [7]). Although this way of implementing a Universal Grammar is an interesting topic in itself (see [17] for a discussion) and may potentially provide insights into whether this approach could be implemented in the brain, the remainder of this paper will focus on learning with catalyst hints because this approach may provide learning-based solutions to certain language acquisition puzzles. In particular, this conception of learning allows for the possibility that the simultaneous learning of related functions may pose significant constraints on the acquisition process by reducing the number of possible candidate solutions.

Having thus established the potential advantages of learning with hints in neural networks, we can now apply the idea of learning using catalyst units to the domain of language acquisition—exemplified by the task of learning to segment the speech stream.

3 Learning multiple related functions in language acquisition

The input to the language acquisition process—often referred to as motherese—comprises a complex combination of multiple sources of information. Clusters of such information sources appear to inform the learning of various linguistic

[1]It should be noted that the results of the mathematical analyses apply independently of whether the extra catalyst units are discarded after training (as is typical in the engineering literature) or remain a part of the network as in the simulations presented below.

tasks (see contributions in [15]). Individually, each source of information, which will be referred to as a *cue*, is only partially reliable with respect to the task in question. Consider the task of locating words in fluent speech.

Speech segmentation is a difficult problem because there are no direct cues to word boundaries comparable to the white spaces between words in written text. Instead the speech input contains numerous sources of information, each of which is probabilistic in nature. Here I discuss three such cues which have been hypothesized to provide useful information with respect to locating word boundaries: (a) phonotactics in the form of phonological regularities [18], (b) utterance boundary information [4, 5], and (c) lexical stress [11]. As an example consider the two unsegmented utterances:

Therearenospacesbetweenwordsinfluentspeech#
Yeteachchildseemstograspthebasicsquickly#

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

(b) The pauses at the end of utterances (indicated above by #) also provide useful information for the segmentation task. If children realize that sound sequences occurring at the end of an utterance must also be the end of a word, then they can use information about utterance final phonological sequences to postulate word boundaries whenever these sequences occur inside an utterance. Thus, in the example above knowledge of the rhyme *eech#* from the first utterance can be used to postulate a word boundary after the similar sounding sequence *each* in the second utterance. As with phonology, utterance boundary information cannot be used as the only source of information about word boundaries because some words, such as the determiner *the*, rarely, if ever, occur at the end of an utterance.

(c) Lexical stress is another useful cue to word boundaries. Among the disyllabic words in English, most take 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, but, once again, this source of segmentation information is only partially reliable because in the above example there is also a disyllabic word with the opposite iambic stress pattern: *between*.

Returning to the notion of learning with hints, we can usefully construe word segmentation in terms of two simultaneous learning tasks [9]. For children acquiring their native language, the goal is presumably to comprehend the utterances to which they are exposed for the purpose of achieving specific outcomes. In the service of this goal the child pays attention to the linguistic input. Recent studies [18, 19] have shown that adults, children and 9-month old

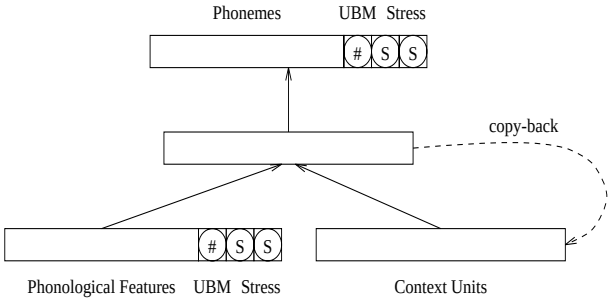

Figure 1: Illustration of the SRN used in [9]. Arrows with solid lines indicate trainable weights, whereas the arrow with the dashed line denotes the copy-back weights (which are always 1). The SRN had 14 input units, 36 output units and 80 hidden/context units.

infants cannot help but incidentally encode the statistical regularities in the input. This task of encoding statistical regularities governing the individual cues will be referred to as the *immediate* task. In the case of word segmentation, phonology, utterance boundary information, and lexical stress would be some of the more obvious cues to attend to. On the basis of the acquired representations of these regularities the learning system may derive knowledge about aspects of the language for which there is no single reliable cue in the input. This means that the individual cues may be integrated and serve as hints towards the *derived* task of detecting word boundaries in the input. In other words, the hints represent a set of related functions which together may help solve the derived task.

This is illustrated by the account of early word segmentation developed in [9]. A Simple Recurrent Network [12] was trained on a single pass through a corpus consisting of 8181 utterances of child directed speech. These utterances were extracted from the Korman corpus [13] (a part of the CHILDES database [14]) consisting of speech directed at pre-verbal infants aged 6–16 weeks. 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 [9] 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) ut-

[2]Note that these phonological *citation forms* are unreduced (i.e., they do not include the reduced vowel *schwa*). The stress cue therefore provides additional information not available in the phonological input.

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

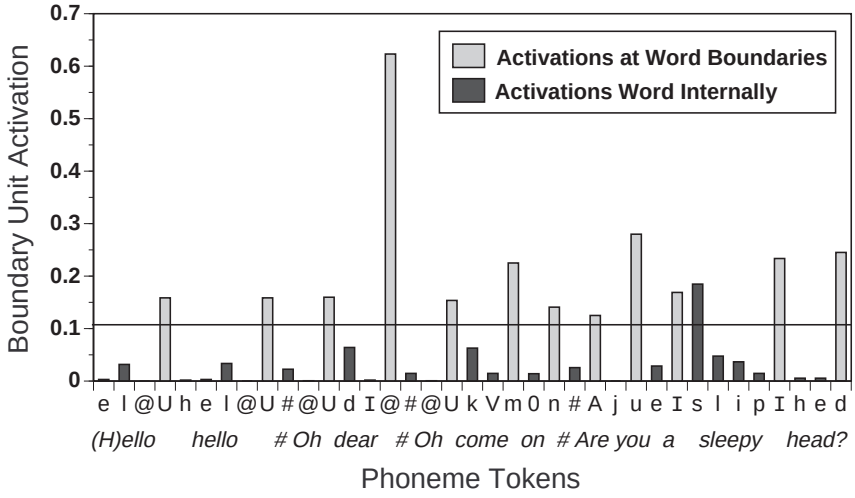

Figure 2: The activation of the boundary unit during the processing of the first 37 phoneme tokens in the training corpus. A gloss of the input utterances is found beneath the input phoneme tokens.

terance boundary information represented as an extra feature (UBM) 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 *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. On the reasonable assumption that phonology is the basic cue to word segmentation, the utterance boundary and lexical stress cues can then be considered as extra *catalyst units*, providing hints towards the derived task.

With respect to the network, the logic behind the derived task is that the end of an utterance is also the end of a word. If the network is able to integrate the provided cues in order to activate the boundary unit at the ends of words occurring at the end of an utterance, it should also be able to generalize this knowledge so as to activate the boundary unit at the ends of words which occur *inside* an utterance [4]. 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. Grey bars indicate the activation at lexical boundaries, whereas the black bars correspond to activation at word internal positions. Activations above the mean (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*).

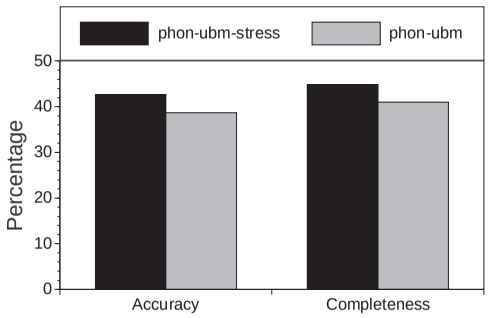

Figure 3: Word accuracy and completeness for the net trained with three cues (phon-ubm-stress – black bars) and the net trained with two cues (phon-ubm – grey bars).

In order to provide a more quantitative measure of performance, accuracy and completeness scores [5] were calculated for the separate test corpus consisting of utterances not seen during training:

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

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

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 hint. Figure 3 shows the accuracy and completeness scores for the networks forced to integrate two or three cues during training. The phon-ubm-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-ubm network. These results thus demonstrate that having to integrate the additional stress cue with the phonology and utterance boundary cues during learning provides for better performance.

To test the generalization abilities of the networks, segmentation performance was recorded on the task of correctly segmenting novel words. Figure 4 shows the performance of the two networks on this task. The three cue net was

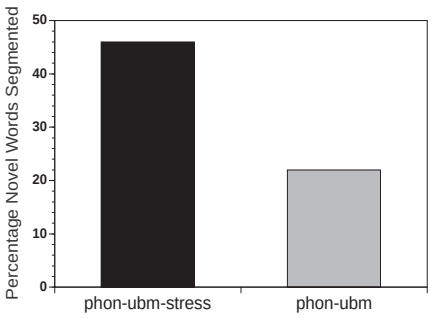

Figure 4: Percentage of novel words correctly segmented (word completeness) for the net trained with three cues (phon-ubm-stress – black bar) and the net trained with two cues (phon-ubm – grey bar).

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-ubm-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-ubm net. These results therefore supports the supposition that the integration of three cues promotes better generalization than the integration of two cues.

Overall, these simulation results from [9] 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[4]. The comparisons between the nets provided with one and two additional related cues in the form of catalyst units, demonstrate that the availability of the extra cue results in the better learning and generalization. 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.

4 Constraining the hypothesis space

The integration of the additional cues provided by the catalyst units significantly improved network performance on the derived task of word segmentation. We can get insight into why such hints may help the SRN by considering one of its basic architectural limitations, originally discovered in [10]; namely that SRNs tend only to encode information about previous subsequences if this information is locally relevant for making subsequent predictions. This means that the SRN has problems learning sequences in which the local dependencies are essentially arbitrary. For example, results in [6] show that the SRN performs poorly on the task of learning to be a delay-line; that is, outputting the

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

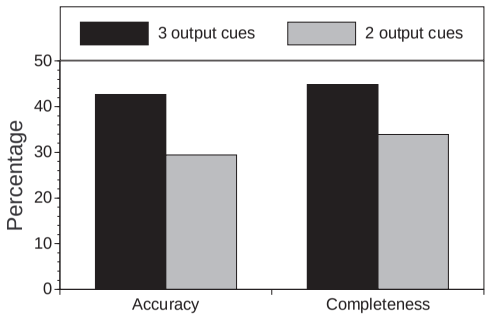

Figure 5: Word accuracy and completeness for the net trained with three output cues (phon-ubm-stress – black bars) and the net trained with two output cues (phon-ubm – grey bars). Both nets received three cues as input.

current input after a delay of N time-steps.

However, this architectural limitation can be alleviated to some degree if the set of training items has a nonuniform probability distribution. This forces the SRN to encode sequences further back in time in order to minimize the error on subsequent predictions. Interestingly, many aspects of natural language are characterized by nonuniform probability distributions; for example, approximately 70-80% of the disyllabic words in English speech directed at infants have a trochaic stress pattern (e.g., 77.3% of the disyllabic words in the training corpus used in [9] had a strong-weak stress pattern).

What the integration of cues buys the network is that it forces it to encode more previous information than it would otherwise. For example, analyses of the simplified model of word segmentation in [3] showed that if an SRN only had to predict the next phoneme, then it could get away with encoding only relatively short sequences. However, the addition of another cue in the form of a catalyst unit representing utterance boundary information forced the net to represent longer sequences of previous input tokens. Encoding longer sequences is necessary in order to reduce the error on the task of predicting both the next phoneme and the on-off status of the utterance boundary unit. The network can thus reduce its error by keeping track of the range of previous sequences which are likely to lead to the utterance boundary unit being activated. A similar story appears to hold with respect to the stress cue in [9].

These analyses suggest how cue integration may force the SRN to acquire more efficient internal representations in order to make correct predictions, focusing on the benefit of having extra catalyst units in the output layer. However, given that the above phon-ubm-stress SRN received three cues both as target output and as input, it is conceivable that it is the *extra input* that is causing the improved performance over the two cue net, rather than the *extra output* cue. In other words, perhaps it is the availability of the extra information on the input which underlies the performance improvement. To investigate this possibility, additional simulations were run. In these simulations, an SRN received three cues as input (i.e., phonology, utterance boundary, and stress

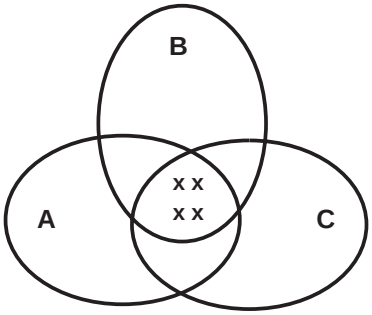

Figure 6: An abstract illustration of the reduction in weight configuration space which follows as a product of accommodating several partially overlapping cues within the same representational substrate.

information), but was only required to make predictions for two of these cues; that is, for the phonology and utterance boundary cues. All other simulation details were identical to [9].

Figure 5 provides a comparison between the network provided with three input/two output cues and the earlier presented phon-ubm-stress network which received three input/output cues. The latter network was both significantly more accurate (42.71% vs. 29.44%: $\chi^2 = 118.81, p < .001$) and had a significantly higher completeness score (44.87% vs. 33.95%: $\chi^2 = 70.46, p < .001$). These additional results demonstrate that it is indeed the integration of the extra stress cue with respect to the prediction task, rather than the availability of this cue in the input, which is driving the process of successful integration of cues. Cue integration via catalyst units thus seems to be able to constrain the set of hypotheses that the SRN can successfully entertain.

4.1 Reducing weight space search

We can conceptualize 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 which will enable a network to learn the function denoted by that cue. For example, the ellipse marked A designates the set of weight configurations which allow for the learning of the function A described by the A cue. With respect to the simulation reported above, A, B and C can be construed as the phonology, utterance boundary, and lexical stress cues, respectively.

If a gradient descent network 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 minimize its error. This constrains the overall set of weight configurations that

the net has to choose between—unless the cues are entirely overlapping (in which case there would not be any added benefit from learning this 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.

Thus, the introduction of cues via catalyst units may reduce the size of the weight space that a network has to search for an appropriate set of weights. And since the cues designate functions which correlate with respect to the derived task, the reduction in weight space is also likely to provide a better representational basis for solving this task and lead to better learning and generalization.

5 Conclusion

This paper has presented evidence in support of the idea that the integration of multiple sufficiently correlated, partially informative cues may constrain learning and over-generalization. In this connection, results from an SRN model of word segmentation was presented which was able to achieve a high level of performance on a derived task for which there is no single reliable cue. This SRN model has also recently been shown to be able to successfully deal with variations in the speech input in terms of coarticulation and high degrees of segmental variation [8].

The approach presented here may have ramifications outside the domain of speech segmentation insofar as children readily learn aspects of their language for which traditional theories suggest that there is insufficient evidence (e.g., [7]). The traditional answer to this poverty of the stimulus problem is that knowledge of such aspects of language is specified by an innate Universal Grammar. A more compelling solution may lie in the integration of cues as exemplified in the word segmentation model. Since recent research has revealed that higher level language phenomena also appear to involve a variety of probabilistic cues [15], the integration of such cues may provide a sufficient representational basis for the acquisition of other kinds of linguistic structure through derived tasks.

Acknowledgments

Many thanks to Joe Allen, Jim Hoeffner, Mark Seidenberg, and two anonymous reviewers for their helpful comments on an earlier version of this paper.

References

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

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

[3] J. Allen & M.H. Christiansen, 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, 1996.

[4] R.N. Aslin, J.Z. Woodward, N.P. LaMendola & T.G. Bever, 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, 1996.

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

[6] N. Chater & P. Conkey, 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, 1992.

[7] N. Chomsky, *Knowledge of Language*, New York: Praeger, 1986.

[8] M.H. Christiansen & J. Allen, Coping with variation in speech segmentation, in submission.

[9] M.H. Christiansen, J. Allen & M.S. Seidenberg, Learning to segment speech using multiple cues: A connectionist model, *Language and Cognitive Processes*, in press.

[10] A. Cleeremans, *Mechanisms of implicit learning: Connectionist models of sequence processing*, Cambridge, Mass: MIT Press, 1993.

[11] A. Cutler & J. Mehler, The periodicity bias, *Journal of Phonetics*, 21, 103–108, 1993.

[12] J.L. Elman, Finding structure in time. *Cognitive Science*, 14, 179–211, 1990.

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

[14] B. MacWhinney, *The CHILDES Project*, Hillsdale, NJ: Lawrence Erlbaum Associates, 1991.

[15] J. Morgan & K. Demuth (Eds), *From Signal to Syntax*, Mahwah, NJ: Lawrence Erlbaum Associates, 1996.

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

[17] W. Ramsey & S. Stich, Connectionism and three levels of nativism, in W. Ramsey, S. Stich & D. Rumelhart (Eds.), *Philosophy and Connectionist Theory*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 287–310, 1991.

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

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

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

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