Learning nonadjacent dependencies in thought, language, and action: Not so hard after all…

Jon A. Willits (jwillits@indiana.edu)
Indiana University, Department of Psychological and Brain Sciences, 1101 E. 10th St.
Bloomington, IN 47405 USA

Abstract

Learning to represent hierarchical structure and its nonadjacent dependencies (NDs) is thought to be difficult. I present three simulations of ND learning using a simple recurrent network (SRN). In Simulation 1, I show that the model can learn distance-invariant representations of nonadjacent dependencies. In Simulation 2, I show that purely localist SRNs can learn abstract rule-like relationships. In Simulation 3, I show that SRNs exhibit facilitated learning when there are correlated perceptual and semantic cues to the structure (just as people do). Together, these simulations show that (contrary to previous claims) SRNs *are* capable of learning abstract and rule-like nonadjacent dependencies, and show critical perceptual- and semantics-syntax interactions during learning. The studies refute the claim that neural networks and other associative models are fundamentally incapable of representing hierarchical structure, and show how recurrent networks can provide insight about principles underlying human learning and the representation of hierarchical structure.

Keywords: hierarchical structure; recurrent connectionist networks; nonadjacent dependencies

Background

Human concepts, languages, goals, and patterns of action are all describable in terms of complex hierarchical structures, but our experience of them as inputs, and our production of them as outputs, is often arranged in linear strings that unfold over time. A necessary consequence of this transformation of complex structure into linear strings is that most human knowledge involves many nonadjacent dependencies, where one element predicts another element, but at a distance. These nonadjacent dependencies, whether in thought, language, or action, enormously expand the computational complexity of representing the structure of the world.

In several subfields of cognitive science, difficulty learning and representing nonadjacent dependencies has generated considerable theoretical controversy. In linguistics, the limitation of simple associative structures has been a cornerstone of arguments for abstract syntactic structures (Chomsky, 1957). In cognitive psychology, researchers argued that associative mechanisms cannot learn the vast range of nonadjacent dependencies in the world, and thus rule-based representations are necessary for human cognition (Bever et al., 1968). In early artificial intelligence, arguments about the limitations of associative systems led to a focus on symbolic, rule-based systems (Newell & Simon, 1961).

However, recent research has questioned the need for rule-based representations of nonadjacent structure. A number of studies have demonstrated or modeled simple learning of nonadjacent structure in memory (Cleeremans & McClelland, 1991), goals and event structure (Botvinick & Plaut, 2004; visual sequences (Fiser & Aslin, 2002), and artificial grammars using linguistic stimuli (Gomez, 2002; Newport & Aslin, 2004). These results have changed the nature of the debate concerning the extent to which knowledge of nonadjacent dependencies requires a rule-based or an association-based explanation. Although there are many specific examples of learning or failing to learn in particular situations, what is lacking is a general account of nonadjacent dependency learning. As a result, the many subfields of cognitive science (such as linguistics, cognitive psychology, and artificial intelligence) continue working on the problem separately, without a clear theory or explanation for some of the most foundational human behaviors.

The current work aims to make progress toward a general account by examining whether a fairly simple neural network model, the simple recurrent network (SRN; Elman, 1990) can provide a general model of nonadjacent dependency learning. An SRN was used because previous research (Botvinick & Plaut, 2006; Cleeremans & McClelland, 1991; Elman, 1991) suggests that SRNs and other recurrent networks are capable of learning nonadjacent structure. However, there is controversy about whether they can serve as general solution for all cases, especially those involving abstract, rule-like relationships (Marcus, 2000) or complex interactions between structure and meaning (Fodor & Pylyshyn, 1988).

In the service of testing the viability of SRNs, the current work had two distinct sub-goals. First, to be a general model of nonadjacent dependency learning, SRNs ought to be able to learn nonadjacent dependencies of the types that exist in the natural

world. This includes abstract, rule-like nonadjacent dependencies, such as learning "distance-invariant" representations (for example, learning the link between *the* and a noun, independent of how many adjectives come between them). Second, SRNs ought to capture behavioral phenomena observed in laboratory experiments, such as facilitated learning in the presence of perceptual (Newport & Aslin, 2004) and semantic (Willits, Lany, & Saffran, 2013) cues. Close analysis of model behavior can then shed light on the bases of the empirical effects. The following three studies test SRNs' abilities to satisfy these criteria.

General Methodology

The three studies shared three core features common in connectionist-modeling approaches (Rumelhart & McClelland, 1986). First, all simulations used sets of interconnected units and weights specifying how strongly each unit was connected to each other unit. The units in the model were divided into an input group, used to specify the input stimulus in each sequence; an output group, used to specify the output response (which also served as a prediction about the next item in the sequence); and a hidden group that mediated between the input and output groups. Second, the models featured recurrent connectivity, allowing the model to feed back information about its own previous internal state in ways critical to forming internal representations of sequential structure. Third, the models all made use of weight-based encoding, where the network's knowledge was encoded in the weighted connections between units.

The goal of the network was to learn a set of weights such that, for any given input, the model's weights led to activation in the output layer that was a correct prediction of the next item in the sequence. During training, a model was given an input, its output activation was treated as a prediction of what the next input would be. This prediction was compared to the target output, and divergence error was calculated across each unit and was used to adjust the weights of the model, using a version of recurrent backpropagation through time. For each simulation, 30 different randomly initialized models were trained. Each model was trained until it reached a predetermined level of overall error, corresponding to optimal prediction performance in the task. The critical test in each simulation was the relative rate of learning across the different conditions in that study.

Study 1: Distance Invariance

In experiments on nonadjacent dependencies using artificial grammars, the distance between dependent items is usually fixed, with one intervening item separating dependent items. However, in many real-world cases (such as the distance between nonadjacently related events in the world, or words in language) the distance between dependent items varies. In fact, learning a "distance-invariant" representation of a nonadjacent dependency has been considered a critical phenomenon, proving the need for a rule-based mechanism.

In Simulation 1, I attempted to train an SRN to learn distance-invariant representations of nonadjacent dependencies by exposing them to the same nonadjacent dependency at multiple spans of distance between the related items. A second issue of interest was whether SRNs would show facilitation in learning longer-distance dependencies if they also had experience with the dependency at a shorter distance, a learning effect that has been demonstrated in both infants and adults (Lany & Gomez, 2008).

Stimuli and Design

The models in Study 1 were trained on sequences where the first element (hereafter the *A* item) perfectly predicted the last element in each sequence (hereafter the *B* item), with the sequences having a number of items (hereafter the *X* items) intervening between them. The sequences were of lengths 2 to 5, resulting in distances between the *A* and *B* items spanning from zero (adjacent dependencies) to three. There were two *AB* pairs (A_1 & B_1, A_2 & B_2) and six possible intervening X-items ($X_1 \ldots X_6$). The x-items were distributed across trials such that they provided zero predictive value for which *B* would occur. The only way to predict the correct *B* (B_1 or B_2) was to have stored which *A* (A_1 or A_2) initiated the sequence. The full set of stimuli used in Study 1 is shown in Table 1.

Thirty different networks (starting from different randomly initialized weights) were trained in each of six different training conditions: (1) only Span 0 trials; (2) only Span 1 trials; (3) only Span 2 trials; (4) only Span 3 trials; (5) a mixture of all Span trials; (6) a mixture of all Span trials *except* Span 3.

Over the course of training, networks from all six conditions were tested on stimuli from all Span conditions (without updating the network weights during those test trials), to assess the network's performance on strings of various spans. Networks were compared at points where they had experienced the same number of trials, controlling for the amount of experience the networks had with each *AB* pair.

Network Architecture

The network had 10 input and output units (one for

each *A*, *B*, and *X*) and 25 hidden units. A simplification of the network architecture is shown in Figure 1.

Table 1. Stimulus inputs used in Study 1.

Span 0	Span 1	
A_1 B_1	A_1 x_1 B_1	A_2 x_1 B_2
A_2 B_2	A_1 x_2 B_1	A_2 x_2 B_2
Span 2	**Span 3**	
A_1 x_1 x_3 B_1	A_1 x_1 x_3 x_5 B_1	A_2 x_1 x_3 x_5 B_2
A_1 x_1 x_4 B_1	A_1 x_1 x_4 x_5 B_1	A_2 x_1 x_4 x_5 B_2
A_1 x_2 x_3 B_1	A_1 x_2 x_3 x_5 B_1	A_2 x_2 x_3 x_5 B_2
A_1 x_2 x_4 B_1	A_1 x_2 x_4 x_5 B_1	A_2 x_2 x_4 x_5 B_2
A_2 x_1 x_3 B_2	A_1 x_1 x_3 x_6 B_1	A_2 x_1 x_3 x_6 B_2
A_2 x_1 x_4 B_2	A_1 x_1 x_4 x_6 B_1	A_2 x_1 x_4 x_6 B_2
A_2 x_2 x_3 B_2	A_1 x_2 x_3 x_6 B_1	A_2 x_2 x_3 x_6 B_2
A_2 x_2 x_4 B_2	A_1 x_2 x_4 x_6 B_1	A_2 x_2 x_4 x_6 B_2

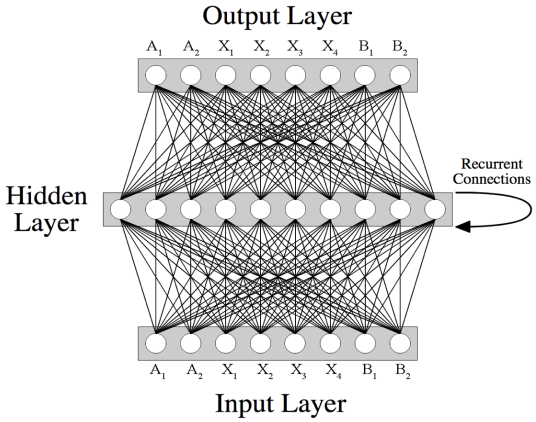

Figure 1. A simplified depiction of the network architecture used in Study 1. The actual model had 8 *X*-units (X_1…X_8) and 25 units in the hidden layer.

Hypotheses

Three main hypotheses were under investigation. First, do networks trained on longer-distance dependencies (bigger Spans) take longer to learn the dependency, as people do? Second, do networks trained in more variable conditions (Conditions 5 & 6) learn more slowly due to increased variability and noise? Or do they, like people (e.g. Lany & Gomez, 2008) show facilitated learning of more distant dependencies due to experience with shorter dependencies? Third, are SRNs capable of learning a distance-invariant representation? Specifically, do the networks that are trained only on Spans of 0, 1, and 2, predict the correct B item on Span 3 trials, even though they have never before experienced the dependency at that distance?

Results & Discussion

Figure 2 shows the average SRN performance predicting the correct B (the network's activation level for the correct B output, on X trials) for networks trained on only a single Span, when tested on the same Span. Networks showed strong effect of taking longer to learn, as the distance between the dependent items increased. Figure 3 shows the average performance on items of Span3 distance, for networks (1) trained on Span3, compared to (2) networks trained on a mixture of all the spans (SpanX) and (3) to networks trained on all the spans except Span3 (SpanX-3). At the earliest stages of training (trials 0-1000), the networks that experienced more variability showed slight decrements in performance on Span3 test items, relative to networks trained on Span3 alone. However, at later stages of training, both SpanX and SpanX-3 networks outperformed the Span3 network on Span3 items.

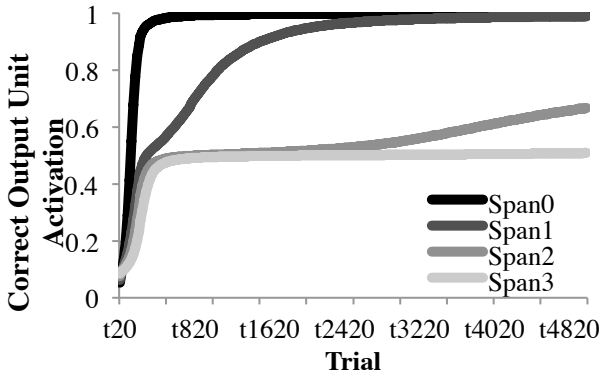

Figure 2. Average SRN performance for networks trained on a single span between nonadjacently dependent items, when tested on items of the same span. The y-axis is the network's softmax activation level of the correct B unit, when the network was presented with the preceding X item.

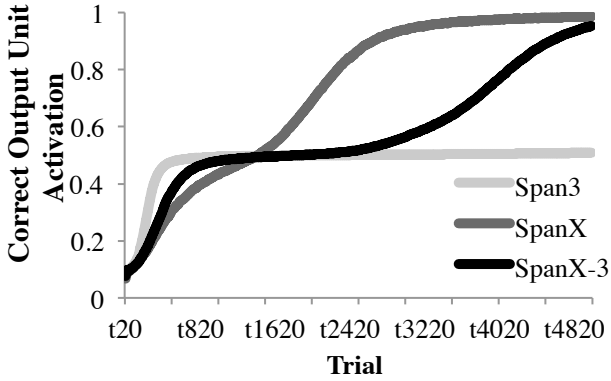

Figure 3. Average SRN performance for networks trained on Span 3, a mixture of all Span conditions (SpanX), or all Span conditions except Span3 (SpanX-3).

Thus, in Study 1 I show that SRNs display three critical features of human learning: (1) they show increased difficulty with longer dependencies; (2) they show facilitated learning when they have had experience with shorter-distance variations of that dependency; (3) they learn distance-invariant representations of nonadjacent dependencies, making

the correct prediction for Span3 items even in the SpanX-3 condition, where they had no training with dependencies of that span. This evidence that SRNs can learn a distance-invariant representations of nonadjacent dependencies is a critical finding, as it undercuts one of the fundamental arguments against association-based representations of knowledge, and in favor of rule-based explanations of cognition.

Study 2: Abstract Rules

Marcus et al. (1999) performed a learning study with infants, where the infants where played sequences of syllables following either an ABB repetition pattern (e.g. "*go-la-la*") or an ABA alternation pattern (e.g. "*go-la-go*"). After hearing many examples repeated multiple times, infants then heard novel test sequences that either followed or violated that rule, and showed evidence of discriminating the legal and illegal sequences. Marcus argued that because no items were co-present at training and test, associative accounts were inadequate and only rule-based models could explain behavior. Marcus (2000) further argued that SRNs (like in Figure 4), could not *in principle* account for this finding. A number of researchers (Altmann & Dienes, 1999; Christiansen & Curtin, 1999) presented distributed SRN models of this phenomenon, where microfeatures (but not items) were co-present at training and test. Marcus, however, argued that resorting to such microfeatures was proof that SRNs and other network models are fundamentally incapable of learning abstract, algebraic rules, which some believe to be fundamental to human cognition.

In Study 2, I show that a simple, localist SRN without any distributed microfeature information learns to represent abstract, rule-like structure. Marcus's (2000) characterization of SRNs was correct; a localist SRN trained in the manner he described cannot show transfer of the rule-like knowledge. That is because the network learns (during the initial training) that the elements in the test items *never occur*, and thus their weights are set to zero, making them unable to make use of any information about the previous items' sequential structure that may have been learned and stored in the network's recurrent or output connections. However, there is no reason to restrict training in this way; one could instead allow the model to *continue learning* during the test phase, and again determine whether the model learns about the rule consistent test strings more quickly than the rule-violating ones.

Stimuli and Design

The models in Study 2 (using the architecture in Figure 4) were trained on the exact design from Marcus et al., shown in Table 2. During the first training phase, models were trained in one of two conditions: (1) an ABA condition, where the first item perfectly predicted the last item, and predicted that it would be repetition of itself; (2) an ABB condition, where the middle item perfectly predicted the last item, again a repetition of itself. These ABA and ABB strings were composed of six possible A's and B's, which all occurred in all possible combinations, thus making all transition probabilities uninformative, and leaving the item-independent ABA or ABB rule as the only way to correctly predict whether the final element should be an A or B. The models were then given a second training phase, where they were trained on a new ABA or ABB sequences using new A and B items, and tested to see if they learned these sequences more quickly if the new rule was consistent with the rule on which they had been trained in phase 1.

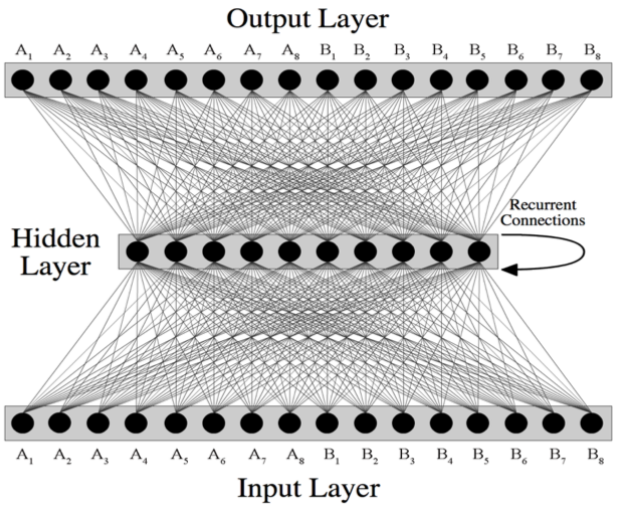

Figure 4. A depiction of the architecture in Study 2. The actual model had 12 A- and B-units and 25 hidden units.

Table 2. Stimulus inputs used in Study 2.

ABA1	ABB1	ABA2	ABB2
$A_1B_1A_1$	$A_1B_1B_2$	$A_7B_7A_7$	$A_7B_7B_7$
$A_1B_2A_1$	$A_1B_2B_2$	$A_7B_8A_7$	$A_7B_8B_8$
$A_1B_3A_1$	$A_1B_3B_3$	$A_7B_9A_7$	$A_7B_9B_9$
…	…	…	…
$A_1B_6A_1$	$A_1B_6B_6$	$A_7B_{12}A_7$	$A_7B_{12}B_{12}$
…	…	…	…
$A_6B_6A_6$	$A_6B_6B_6$	$A_{12}B_{12}A_{12}$	$A_{12}B_{12}B_{12}$

Results & Discussion

The results from Study 2 are shown in Figure 5. When the model was allowed to continue learning during the second training phase, it shows facilitated learning if the new items follow the same structural sequence as the items in the first phase. Follow-up

analyses of the network's weight configurations show this is because the network's recurrent and output weights are effectively learning the abstract structural order of the sequence. Because of this, if the new set of items are following the same structural rule, all the network needs to do is learn to adjust the input weights for the new items so that they work well with the already-learned recurrent and output weights.

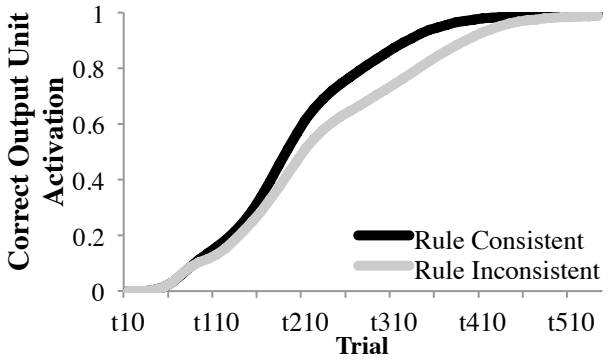

Figure 5. Average SRN performance during the second stage of learning in Study 2.

These findings have very significant implications, as they (along with the findings in Simulation 2), refute claims that associative models are not capable of learning the abstract and rule-like knowledge that seems fundamental to human cognition.

Study 3: Perceptual/Semantic Bootstrapping

Previous research on nonadjacent dependencies has mainly focused on learning to represent sequences of events, actions, or words independent of other cues about those entities, such as perceptual or semantic features or similarity. Learning structure in such a purely symbolic way *would* be hard. However, there is no reason to limit attention to this type of impoverished input, which is uncharacteristic of naturalistic conditions.. Studies that have examined the use of correlated perceptual cues (Newport & Aslin) or semantic cues Willits et al.), have found that under these circumstances nonadjacent dependencies are significantly easier to learn. For example, Willits et al. found that when the items to be learned are from the same category (e.g. nonadjacently related items both *foods*), both infants and adults learn the dependency more easily. Learners even learn the nonadjacent dependency if the two words form a consistent mapping *between* categories (e.g. across set of nonadjacent pairs, *foods* are always paired with an *animals*). These findings are critical, because many of the nonadjacent dependencies people need to learn have these kinds of correlated perceptual and semantic attributes.

The question, then, is whether SRNs also exhibit facilitated learning from correlated cues, thus broadening their appeal as a general model of dependency learning, and whether they provide any insights as to why learning might be easier under these circumstances. This was investigated in Study 3.

Stimuli and Design

The models in Study 3 were trained using the architecture in Figure 6. This architecture allowed for tests of whether correlated similarity structure affected learning by allowing each input to activate two units: (1) one item-specific unit (either an A_N, X_N, or B_N), where the letter refers to which category the item is from); (2) a category-specific unit (either $Category_A$, $Category_X$, or $Category_B$), where the category unit turned on for all inputs that came from that category.

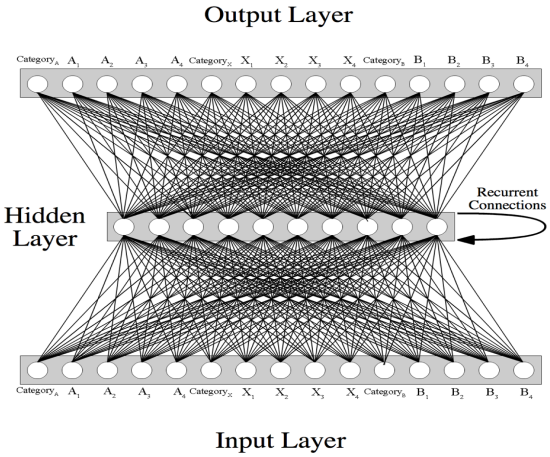

Figure 6. A depiction of the architecture used in Study 3. The actual model had 25 hidden units.

The models were trained in one of three conditions (shown in Table 3). In the *Consistently Same* condition, the nonadjacently dependent items were always from the same category (e.g. the first item in each sequence would activate the A_1 unit and the $Category_A$ unit, and third item would activate the A_3 unit and the $Category_A$ unit). In the *Consistently Different* condition, the nonadjacently dependent items were consistently from opposite A & B categories. In the *Inconsistent* condition, the dependent items'

Table 3. Stimulus inputs used in Study 3

Consistently Same Category	Consistently Different Categories	Inconsistent Categories
$A_1\ X_n\ A_3$	$A_1\ X_n\ B_3$	$A_1\ X_n\ A_3$
$A_2\ X_n\ A_4$	$A_2\ X_n\ B_4$	$A_2\ X_n\ B_4$
$B_1\ X_n\ B_3$	$B_1\ X_n\ A_3$	$B_1\ X_n\ B_3$
$B_2\ X_n\ B_4$	$B_2\ X_n\ A_4$	$B_2\ X_n\ A_4$

categories were not predictable in terms of the other unit in the dependency. Across training trials, the

models were compared to see if any of the conditions showed facilitated learning.

Results & Discussion

The results for Study 3 are shown in Figure 7. SRNs showed facilitated learning in both consistent conditions, but not the inconsistent conditions, results similar to behavioral experiments with infants and adults. Follow-up analyses of network behavior show this is because the network has an easier time learning the category sequences, an intriguing hypothesis to test in future work with human learners.

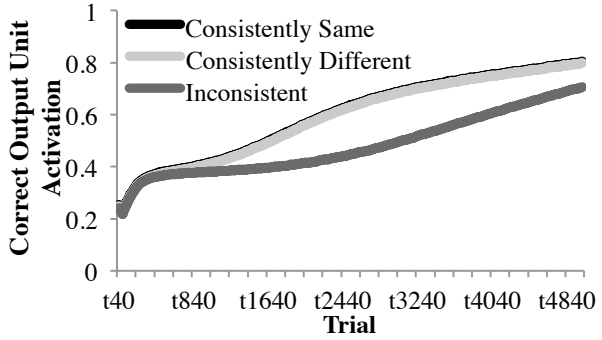

Figure 7. Average SRN performance for the three training conditions in Study 3.

Conclusions

Nonadjacent dependencies are a necessary consequence of experiencing a hierarchically structured world though a linear sequence of inputs and actions. The current studies support the notion that SRNs and other recurrent networks are viable models of the representation of hierarchical knowledge. They are capable of learning to represent abstract, rule-like structure (Study 1 & 2), and they show critical learning effects that people exhibit, such as the interaction between structure and similarity (Study 3).

In addition, these simulations also provide evidence for the hypothesis that many learning situations that are thought to be especially difficult (of which the learning of nonadjacent dependencies is but one example) are only difficult because the problem has been underrepresented. Many cues learners might use are stripped away in overly controlled experiments, making the problem harder than it is in the real world. Complexity is not the same thing as noise, if that complexity provides learners with useful cues to the structure of the world.

Acknowledgments

This work was funded by NICDC F31-DC00936-02 to Jon Willits. The work received much useful input from Mark Seidenberg, Jenny Saffran, Maryellen MacDonald, and Timothy Rogers, and Jessica Montag.

References

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

Bever, T. G., Fodor, J., A., & Garrett, M. (1968). A formal limitation of associationism. . In T. R. Dixon & D. L. Horton's *Verbal Behavior and General Behavior Theory*. Englewood Cliffs: Prentice Hall.

Botvinick, M., & Plaut, D.C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, *111*, 395– 429.

Chomsky, N. (1957). *Syntactic Structures*. The Hague.

Christiansen, M. H., & Curtin, S. L. (1999). The power of statistical learning. In *Proceedings of the twenty-first conference of the Cognitive Science Society*.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*, 235-253

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

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

Fiser, J., & Aslin, R.N. (2002). Statistical learning of higher-order temporal structure from visual shape-sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 458-467.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3-71.

Gómez R.L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431-36.

Lany, J., & Gómez, R. (2008). Twelve-month-old infants benefit from prior experience in statistical learning. *Psychological Science*, *19*, 1247-1252.

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

Marcus, G. F. (2000). Rethinking eliminative connectionism. *Cognition*, *37*, 243-282.

Newell, A. & Simon, H. A. (1972). *Human Problem Solving*. Prentice Hall.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance: I. Statistical learning of nonadjacent dependencies. *Cognitive Psychology*, *48*, 127-162.

Rumelhart, D. E., McClelland, J. L., & the PDP research group. (1986). *Parallel distributed processing: Volume I*. Cambridge, MA: MIT Press.

Willits, J. A., Lany, J., & Saffran, J. R. (in review). Semantic cues facilitate nonadjacent dependency learning in infancy.