An Artificial Grammar Investigation into the Mental Encoding of Syntactic Structure

Pyeong Whan Cho (pyeong.cho@uconn.edu)[1,2]
Emily Szkudlarek (emily.szkudlarek@uconn.edu)[1]
Anuenue Kukona (anue.kukona@uconn.edu)[1,2]
Whitney Tabor (whitney.tabor@uconn.edu)[1,2]
[1] Department of Psychology and Cognitive Science Program, University of Connecticut,
406 Babbidge Road U-1020, Storrs, CT 06269 USA
[2] Haskins Laboratories, 300 George St., Suite 900, New Haven, CT 06511 USA

Abstract

We explore neural network learning and parallel human learning on an artificial language task. The task generates rich data on human interaction with syntactic systems, including recursive ones. Studying the network's properties, we argue for a "Structured Manifold" view of syntactic representation. The "Structured Manifold" lies in the parameter space (weight space) of the network. It exhibits (1) loci of high order, corresponding to complex rule systems, (2) continuity, which explains how one rule system can morph into another one, and (3) "recursion approximation", a concept related to symbolic recursion, which addresses some of the puzzles about embedding patterns in human behavior.

Keywords: artificial grammar learning; artificial neural networks; recurrent networks; simple recurrent networks; sequence learning; recursion; center embedding; rules.

Introduction

What kind of structural system underlies human syntactic processing ability? Much work in linguistics addresses this question by exploring syntactic behaviors in natural languages. Work on artificial grammars offers a chance to obtain detailed information about human interaction with formal syntactic systems in the absence of semantic content or task-independent pragmatic function. Here, we introduce a variant on existing artificial grammar learning tasks that supports careful comparison between human and artificial neural network models. The results help clarify the difference between standard, rule-based conceptions of grammatical knowledge and the claims of the neural net perspective, providing some evidence that, at least in the artificial grammar task, humans resemble the networks. We focus, in particular, on the status of center embedding recursion, which many authors view as an important feature of natural language systems, but whose status in the theory of representation has been much debated (e.g., Chomsky, 1957; Christiansen & Chater, 1999; Friederici, 2002).

Center-embedding recursive patterning can be generated by context free grammars. Context free grammars are rule systems like Grammar G (Table 1) in which rules take the form (A → X_1 X_2 ... X_N, for N a finite number), and there are designated starting rules. The grammar is said to generate a finite sequence of symbols, called a "sentence", if it is possible to make successive substitutions for symbols on the right hand side of a starting rule until no more substitutions can be made; the resulting right hand side is the generated sentence. Grammar G generates the sentences "1 2 3 4" (a Level 1 sentence), "1 1 2 3 4 2 3 4" (Level 2), "1 1 1 2 3 4 2 3 4 2 3 4" (Level 3), etc. In formal language terminology, a case where the system shifts to a deeper level of embedding (here, 1 after 1)—is called a "push" and a case where it shifts back (2 after 4) is called a "pop". Keeping track of the syntactic dependencies requires correlating the pops with the pushes. The term "recursion" refers to the situation in which a rule can be invoked an unbounded number of times. "Center embedding recursion" is the case in which the symbol for such a repeatedly used rule occurs in the middle of one of the rules with symbols on either side of it (e.g., in G, "S" occurs with "1" to its left and "2" to its right in the first rule). Center embedding context free grammars are of particular interest because a system for generating or recognizing all and only the sentences produced by a center embedding grammar needs an unbounded memory.

It is generally recognized that some degree of center embedding is present in natural languages, for there are many situations where natural languages employ patterns within patterns of the same type—e.g., in relative clauses. This suggests that minds have recursive rule systems at their disposal for keeping track of these patterns. The recursive rule system is appealing as an explanation because it permits efficient description of many cases and predicts the way people exercise their language knowledge in many new combinations of words and phrases (Pinker, 1994).

Yet humans have great difficulty processing more than a few levels of embedding in natural language (see Lewis, 1996). Similar findings characterize artificial grammar work on recursion (de Vries, Monaghan, Knecht, & Zwitserlood, 2008; Poletiek, 2002). If a symbol processing system must only handle a few levels of embedding, then it is not strictly necessary to employ a recursive process—a weaker, finite-state device, which has a limited memory capacity, can do the job. Proponents of recursive rules have suggested that memory limitations obscure a fundamentally infinite

Table 1: Grammar *G*. Both rules are starting rules.

S → 1 S 2 3 4
S → 1 2 3 4

mechanism. But even if humans had such a mechanism, it would be impossible for finite life-span researchers to observe its infinite behavior. Thus, the argument for human employment of recursive systems seems to founder on a shoal of infinity: true center-embedding recursion is distinguished by its employment of infinite memory, but we cannot observe infinite memory, so it is hard to justify recursive rules.

Relevant to this discussion, artificial neural networks have been used to model many aspects of natural language behavior and they are often claimed to do so without recourse to "explicit" rules (e.g., McClelland & Patterson, 2002; see also Pinker & Ullman 2002). Elman's Simple Recurrent Network or "SRN" (Elman 1991) is a model of this sort that processes structured sequences. The SRN and its relatives have learned some elaborate patterns of center-embedded recursion and have successfully generalized from training on less deeply embedded cases to prediction of more deeply embedded cases (Rodriguez, 2001; Wiles & Elman, 1995). However, they also do not typically extend the patterns very far beyond their training (Christiansen & Chater, 1999). In light of the difficulty that humans have with processing deep center embeddings, Christiansen and Chater argue that the networks' behavior provides an appealing alternative account to the recursive rule approach.

However, the network representations are not well understood. In particular, if the networks do not employ rules, it is not clear what kinds of order they predict should occur; nor is it clear why observed behaviors can often be given a parsimonious description with systems of rules. We suggest that it will help to look closely at the nature of the network representations, in conjunction with detailed measurement of human behavior on a task that both networks and humans can perform well. Through such an approach, we can acquire some insight to the conundrums of human recursive patterning.

In particular, we suggest that the network view is well described as a "Structured Manifold" account. We use the term *manifold* to draw attention to the fact that the network parameters are real-valued so they can change continuously, and continuous change of parameter values is associated with continuous change in the network's behavior (see Spivey, 2007). This property is useful for explaining the learning phenomena—it makes it so the networks can be sensibly described as "getting closer" to a particular structural behavior before the behavior actually appears. On the other hand, the *structured* part of "Structured Manifold" refers to the fact that the network behaviors in the context of a particular environment tend to concentrate around a few types. These types correspond to qualitatively distinct lawful patterns in the network's relationship to its environment. They are closely related to rule-systems, for they correspond to systematic insights about the patterns in the world. In particular, the Structured Manifold approach suggests a way of understanding "recursion" that avoids the "shoal of infinity" mentioned above. We say that a pattern of behavior *approximates a recursive mechanism* if knowledge of one structural feature of the environment transfers to another structural feature which is recursively related to the first.[1] The knowledge need not transfer perfectly and thus the system may not follow the recursive rule to arbitrary levels, but to the degree that the system's knowledge is iteratively effective, it will be said to form a "good" approximation of the recursion. Thus the definition clarifies the sense in which a network can be "close to" a recursive behavior without embodying it. The definition also allows perfect recursion to be present at a locus in parameter space, in keeping with formal analyses of some recurrent networks (Tabor, 2000; 2009). There is also a way of gleaning empirical evidence for recursion approximation: statistical evidence that a system bases its behavior with a more embedded case on its knowledge about a less embedded one counts as such.

The remainder of the paper is organized as follows: in "Task" we introduce the grammar learning task. In "Simple Recurrent Network Model", we describe the outcome of training many SRNs on the task and testing three hypotheses—Grouping, Continuous Interpolation, and Recursion Approximation—generated by the Structured Manifold view. In "Human Grammar Learning Experiment", we report on a parallel study with human grammar learners. "General Discussion" concludes.

Task

We employed a grammar learning task called the *Box Prediction Task* that is a variant of sequence learning tasks (Clegg, Di Girolamo, & Keele, 1998). In sequence learning, a popular task is the Serial Reaction-Time task (Nissen & Bullemer, 1987) where stimuli are presented sequentially and participants respond to each stimulus (e.g., by clicking on the place where the stimulus appeared). Participants' responses trigger the presentation of the next stimulus. In patterned sequential data, reaction times often reflect the predictability of the sequence, suggesting that participants develop a structured encoding of the data. However, it is difficult to tell from the data in such a task when a participant has reliably detected complex dependencies like those that occur in center-embedding.

In the Box Prediction Task, stimuli are presented sequentially but participants are asked to predict the next stimulus instead of simply reacting to the current stimulus. Human participants predict by clicking a box on a screen. They immediately get feedback because the correct box changes color (from black to green or blue). The networks predict by activating output nodes corresponding to boxes. They also get immediate feedback in the form of a vector indicating which symbol the grammar produced next.

[1] We assume, for analysis purposes, that the environment contains patterns which are describable by recursive rules. This assumption does not commit us to claiming that actual environments have infinite patterning. Instead, one can think of this assumption as a tool for understanding the structure of human and network behaviors.

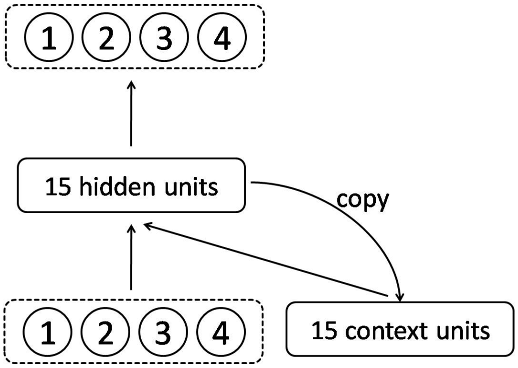

Figure 1: Network architecture.

Many artificial grammar learning tasks have tested languages with high nondeterminism (e.g., Reber, 1967). Performing the Box Prediction Task with such data would be very frustrating because only a few responses are likely to be correct. We therefore employed a grammar (G) with very little nondeterminism and we color-coded the pushes (blue), which are the only nondeterministic transitions, telling the participants that they need not predict them.

Sentences generated by the grammar were concatenated to form long training sequences that were presented sequentially to networks and human participants.

Simple Recurrent Network Simulations

Method. 22 Simple Recurrent Networks (SRNs) with the same architecture (Figure 1) were constructed and the initial weights were randomly set (uniform distribution on [-0.1, 0.1]). Each network was trained twice from the same initial weights on a sequence of Level 1 and Level 2 sentences. In the first sequence, the average frequency of Level 2 sentences increased over the course of 8000 trials (Table 2). In the second sequence, it decreased. We expected Sequence 1, which emphasized Level 1 before Level 2, to produce better recursion approximation because, in recursive generalization, (Level) 2 to 3 parallels 1 to 2, not 2 to 1.

Results. We asked three questions about the ensemble of networks: (1 "Grouping") Can the networks be grouped into a few, qualitatively distinct behaviors which correspond to rational responses to the task environment? (2 "Continuous Interpolation") Do the networks favor intermediate states, in which they blend the qualitative behaviors just mentioned? (3 "Recursion") Is there evidence that the more successful individuals approximate a recursive mechanism?

Table 2: Distribution of sentences in the 2 sequence types.

Sequence Type	No. of sentences per phase				
	Phase1	Phase2	Phase3	Phase4	Total
Sequence 1					
L1 sentence	2000	2000	2000	2000	8000
L2 sentence	200	400	600	800	2000
Sequence 2					
L1 sentence	2000	2000	2000	2000	8000
L2 sentence	800	600	400	200	2000

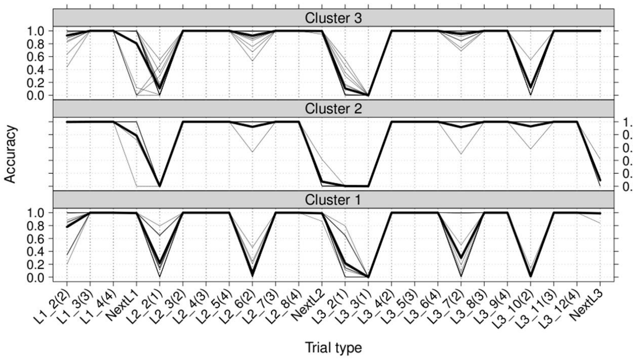

Figure 2: Network Clusters. Li_j(k) means Level i, j'th word, symbol k.

For (1 "Grouping") we used a cluster analysis. After training, we fixed the weights of each network and tested it on a Level-1, a Level-2, and a Level-3 sentence, thus examining a total of 24 word-to-word transitions. Level-1 and Level-2 sentences occurred in training, but Level-3 did not. We interpreted the network's output nodes as probabilities by using the Luce Choice Rule with base e^{10} and computed the expected accuracy of each network at each transition from these probabilities. We then applied *K*-means clustering to the 24 accuracy values. A standard method of choosing the number of clusters, selecting the "knee" in the plot of within-group sum of squares vs. number of clusters, suggested 3, 4, or 5 clusters.[2] For simplicity, and for alignment with the analysis of human data reported below, we focused on the 3-cluster case.The accuracies of the three clusters are shown in Figure 2 (means shown in bold). The means of Cluster 1 indicate that Cluster 1 networks tend to employ a "Simple Markov" strategy: 1→2, 2→3, 3→4, 4→1 (these numbers refer to grammar symbols). Each prediction by the network is conditioned strictly on the input, even if a push or a pop produces a violation of expectation. The means of Cluster 2 indicate that Cluster 2 networks also use the rules, 1→2, 2→3, and 3→4, but the networks switch between two modes of responding to input 4: if the previous successor of 4 was 1, then the next response to 4 is 1. If the previous successor of 4 was 2, then the next response to 4 is 2. This "2-Mode Perseverater" has some memory for the past, but cannot keep track of the correlation between pushes and pops. The Cluster 3 means indicate that Cluster 3 networks expect a Level 1 sentence if the sentence begins 1-2, and they expect a Level 2 sentence if the sentence begins 1-1-2. However, they don't, on average, generalize the dependency to level 3; instead, most of them tend to treat 1-1-1-2 the same as 1-1-2, thus failing on the second pop of Level 3 sentences. Nevertheless, these "Fragile 1-Counters" approximate the behavior of the unbounded recursion generating process better than the other two types (Mean

[2] We also sought a maximum of the Calinski-Harabasz pseudo-*F* statistic (Calinski & Harabasz, 1974), another standard method, but there was no clear maximum.

Table 3: Networks per cluster for each condition.

	Cluster 1	Cluster 2	Cluster 3
Sequence 1	11	1	10
Sequence 2	6	5	11

Accuracies: Simple Markov 77%, 2-Mode Perseverator 77%, Fragile 1-Counters 83%). In sum, the clustering analyses reveal that the networks fall into distinct qualitative categories which are associated with distinct systematic responses to the task. Although one might expect Sequence 1 to encourage 1-counting and Sequence 2 to discourage it, a likelihood ratio test showed no effect of training condition on the distribution of clusters, even with clusters 2 and 3 treated as one ($\chi^2(1) = 2.42$, $p = .120$) (Table 3).

To investigate (2 "Continuous Interpolation"), we contrasted two hypotheses: (a) network behaviors are distributed as cluster prototypes + noise (equal distortion in all directions); (b) the networks approximate blends of behaviors associated with the various cluster prototypes. Under both (a) and (b), a network could be proximal to a pure complex behavior (e.g., a recursive grammar) without precisely embodying it. But in (a) deviations have low likelihood of leading to purer recursion because they can occur in any direction; in (b) deviations are more likely to lead to purer behavior because the models are restricted to a low-dimensional manifold. In this sense, proximity of a network in case (b) to an ideal complex behavior is a more reliable indication that the network will robustly exhibit complexity, than in (a). How can we tell (b) apart from (a) empirically? If (a) holds, then the variation of each network's behavior on each transition is expected to be equal. If (b) holds, then the individual networks are expected to show greater variation on transitions in which networks tend to contrast than on transitions in which all networks tend to agree. We tested this hypothesis by comparing the variances of individual networks' behaviors on different types of transitions during the last 99 trials to the global variances on the same types of transitions (global variance on a transition T is the total variance across all network behaviors on T). Examining all 44 trained networks, we considered the 24 transition types associated with 1-level, 2-level and 3-level sentences. We hypothesized that each 24-element vector of individual variances would be more aligned with the 24-element vector of global variances than with a 24-element vector of uniform variances. Indeed, a paired t-test on our 44 network sample showed that the cosine of the angle between the individual variance vector and the global variance vector was significantly bigger than the cosine of the angle between the individual variance vector and any positive uniform variance vector ($p < .001$).

For (3 "Recursion"), we examined individual network response patterns to see if any of the networks generalized to correct performance on Level 3, provided they had learned correct performance on Level 2. We counted a network as having correct performance on a sentence if its accuracy was above 0.5 on all deterministic transitions in the sentence. Indeed, by this criterion, three of the networks from Cluster 3 exhibited correct performance on Level 3 sentences. Moreover, a regression analysis showed that even when these three networks were removed from the data set, better performance on Level 1 and 2 sentences predicted better performance on Level 3 sentences ($r = 0.32$, $p < .05$). These observations suggest that the networks that do well on Level 3 sentences do so in virtue of their ability to do well on Level 2 sentences, even if they do not precisely embody the recursive generating process. Under the definition given in the Introduction, this observation suggests that the networks approximate a recursive mechanism.

Human Grammar Learning Experiment

Method

Participants. 44 college students from the University of Connecticut participated for course credit.

Materials. Two sequences of 400 trials each were created. In both sequences 1 and 2 there were 38 Level-1 sentences, 25 Level-2 sentences, and 4 Level-3 sentences. The last 99 trials of both sequences were identical. The first instance of a level-3 sentence occurred at trial 302 so trials 1-301 served as an analog of network training and trials 302-400 served as an analog of network testing. In Sequence 1 the density of Level 2 sentences changed from low to high over the course of trials 1-301. Sequence 2 had the reverse progression. A windows PC with speakers on the monitor and a standard mouse were used for the display and input. The experiment was run in E-Prime.

Procedure. Participants saw 4 black boxes on a screen. The boxes were positioned in a circle with grammar G numbers associated counterclockwise, but not indicated on the screen. Each participant ran only one sequence. When a participant clicked a box, one of the 4 boxes would turn a different color indicating that it was the next box in the sequence. The participant was instructed to try to predict which box would next change color and click on it. It was emphasized that the goal was prediction, and not to simply click the box which had previously changed color. If a participant predicted the wrong box on non-push trial, a short beep sounded. No sound was played if the participant predicted the correct box. The correct next box generally changed from black to green, except on push trials, where the second 1 box changed to blue, and the third to cyan (in a Level 3 sentence). Participants were instructed that they need not predict the blue/cyan boxes. The computer recorded the accuracy of the participants' predictions.

Results

Mean accuracies over the course of the task are shown in Figure 3, separated into 3 classes: sequential transitions pops, and pushes. A logistic regression analysis supported staged learning—e.g., final pops learned before intermediate pops in Level 2 sentences for Sequence 1 (p < .001). We

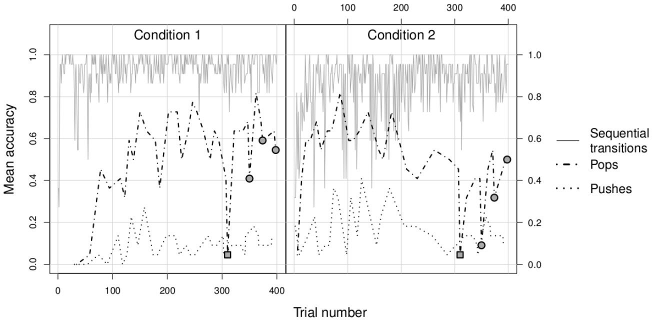

Figure 3: Mean accuracy change during the task. Level 3 2nd pop trials (L3_10(2)) are circled.

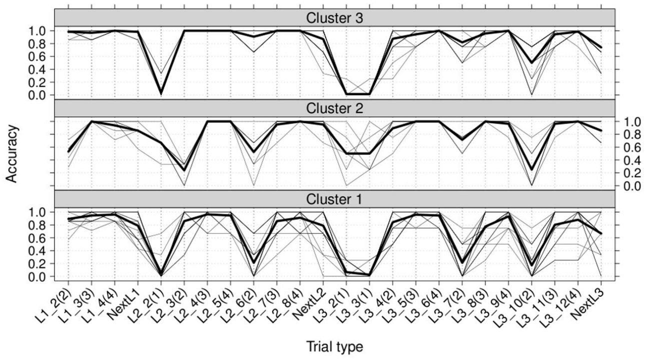

Figure 4: Human clusters.

focused our analysis on the last 99 trials, when both sequence types experienced the same sequence of boxes.

(1 “Grouping”) In the human case, the Calinski-Harabasz pseudo-*F* statistic had a clear maximum at 3 clusters so we examined this case (Figure 4). Figure 4 suggests that the participants in Human Cluster 1, like the networks in Network Cluster 1, employ the Simple Markov system. Cluster 2 is much more sensitive to the temporal structure of pops and pushes, for these participants perform more accurately on the Level 2 pop and the first Level 3 pop, than Cluster 1 participants. Cluster 2 participants tend to employ the rule, 1 → 1, so they have relatively high accuracies on pushes (even though the instructions said that the blue boxes need not be predicted). Although these “Push Predictors” did numerically better than Cluster 1 on the second pop of Level 3, their mean performance on this transition was less than 0.5, suggesting that they are not robustly sensitive to the correlation between pushes and pops. Cluster 3 uses a different strategy with pushes—they generally predict 2 after 1, thus failing to predict the pushes and successfully predicting the finite state transitions from 1 to 2. These “Push Blindsiders” are even better than Cluster 2 at predicting the correlation between pushes and pops. In fact, the Cluster 3 mean accuracy on the second Level 3 pop is above 0.5. These results indicate that the human behaviors, like the network behaviors, can be grouped into several different coherent responses to the task, though the human cluster prototypes are associated with somewhat different strategies than the network cluster prototypes. Table 4 gives the number of participants in each cluster as a function of training sequences. A likelihood ratio test of showed no effect of training condition on the distribution of clusters, even with Clusters 2 and 3 treated as one ($\chi^2(1) = 2.34$, $p = .126$).

Table 4: Participants per cluster for each condition.

	Cluster 1	Cluster 2	Cluster 3
Sequence 1	7	3	12
Sequence 2	12	4	6

Regarding (2 “Continuous Interpolation”), comparison of variance vectors confirmed that, for humans, like networks, most of the individual variation was on dimensions on which there was high global variance (p < .001). As with the nets, this result suggests that, when the humans diverge from the coherent behaviors associated with the clusters, they tend to diverge in the direction of other coherent behaviors. Interestingly, when we performed the global variance test on a cluster by cluster basis, Clusters 2 (N = 7) and 3 (N = 18) showed significant correlation, but Cluster 1 (N = 19) did not, even though Cluster 1 had the largest sample size. These results provide suggestive evidence that the Push Predicters and the Push Blindsiders are hamstrung between the pull of their cluster prototypes and the impulse to be like Simple Markov processes, or like each other, while the Simple Markov processors are, on average, insensitive to the non Markovian structure in the data.

Regarding (3 “Recursion”), there were five participants, all Push Blindsiders, whose mean accuracy on all deterministic Level 1, Level 2, and Level 3 transitions never strayed more than 0.5 away from the predictions of the generating process over the last 99 trials. These people can be said to have mastered the push-pop correlation across the three levels, providing suggestive evidence that they employ a recursive mechanism. Furthermore, a regression analysis showed that mean accuracy on Level 3 sentences was positively correlated with Level 1/2 accuracy in the last 99 trials ($b = 0.633$, $t = 3.38$, $p < .01$). This result is consistent with the recursion approximation hypothesis: the correlation between Level 1/2 and Level 3 suggests that the structural insight about Level 1/2 is being used to solve Level 3. However, the humans, unlike the networks, can keep on learning during the “test” trials, so the correlation might stem from a greater learning facility in some humans than others: those who have greater learning facility will learn Level 1 and 2 sentences better during trials 1-301 and they will also learn Level 3 sentences better during trials 302-400, but they might not use any of their knowledge of Level 1 and 2 sentences to solve Level 3. However, in a separate analysis, Sequence Type predicted Level 1/2 accuracy ($b = -0.052$, $t = -2.53$, $p < .05$). These results are unexpected on the Learning Facility account for two reasons: the density manipulation does not change the total amount of exposure to Level 1 and Level 2 sentences, so if these were simply learned on the basis of exposure, there would be no reason for a Sequence Type effect on Level 1/2. Second if, as claimed by the Learning Facility account, Level 3 sentences are learned independently of structural insight gleaned from Levels 1 and 2, then there would be no reason to expect Level 3 variation to be related to anything except participant identity. Instead, the data suggest that Sequence type influences the learning of Levels 1 and 2, and the nature of this learning, in turn, influences performance on Level 3s, consistent with the recursion approximation hypothesis.

General Discussion

The similarities between the network and human results provide some evidence that the Structured Manifold is a good framework for understanding human syntactic encoding, at least in artificial grammar learning.

The network analysis helps clarify the notions of "Grouping", "Continuous Interpolation" and "Recursion Approximation". In particular, the Grouping results provide evidence that network interaction with the environment focuses on a small finite number of coherent behaviors. Even the Simple Markov system, though it is not optimal for the task, detects a level of regularity which is inherent in the task structure---the so-called "second-order" statistical approximation. In dynamical systems terms, it seems likely that these structures are attractors of some sort. It may be helpful to ask what the nature of their stability is within the panoply of dynamical stabilities (see Tabor, 2009).

The Continuous Interpolation results are related to parameter-setting models of syntax (e.g. Chomsky, 1981) in the sense that they provide a reduce-dimension description of the range of expected behavior. An important difference between the current model and linguistic parameter setting models, is that the structure of the "parameters" was derived from the interaction of a very general-purpose learning mechanism with the environmental data. Thus, this appears to be a less nativist kind of parameter setting.

Finally, the Recursion Approximation analysis suggests a way of reconciling the desirable properties of recursive rules with the facts that human behavior is imperfect and cannot be infinitely observed. Combined with the Continuous Interpolation observation, the results suggest understanding states of a system as being related not just to one but to many ideal forms. This suggests shifting away from a view of organisms as "having a knowledge system" and toward a view in which they can be "in the sphere" of multiple systems. Their actual behavior is not static, and may be understood as a structured trajectory through these spheres.

Acknowledgments

We gratefully acknowledge support from NICHD grant HD40353 to Haskins Laboratories and NICHD Predoctoral NRSA HD060414 to AK.

References

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*, 1-27.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton and Co.

Chomsky, N. (1981). *Lectures on Government and Binding*. Mouton de Gruyter.

Christiansen, M. & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science, 23*, 157-205.

Clegg, B. A., DiGirolamo, G. J., & Keele, S. W. (1998). Sequence learning. *Trends in Cognitive Sciences*, *2*, 275-281.

De Vries, M. H. Monaghan, P., Knecht, S., & Zwitserlood, P. (2008). Syntactic structure and artificial grammar learning: The learnability of embedded hierarchical structures. *Cognition, 107*, 763-774.

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

Friederici, A. (2002). Towards a Neural Basis of Auditory Sentence Processing. *Trends in Cognitive Sciences*, *6*, 78-84.

Lewis, R. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, *25*, 93-115.

Magnuson, J., Tanenhaus, M., Aslin, R., Dahan, D. (2003). The time course of spoken word learning and recognition: Studies with artificial lexicons. *Journal of Experimental Psychology: General, 132*, 202-227.

McClelland, J. & Patterson, K. (2002). Rules or connections in past-tense inflections: what does the evidence rule out? *Trends in Cognitive Sciences*, *6*, 465-472.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: evidence from performance measures. *Cognitive Psychology*, *19*, 1-32.

Pinker, S. (1994). *The Language Instinct: How the Mind Creates Language*. New York: Perennial Classics.

Pinker, S. & Ullman, T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*, 456-463.

Poletiek, F. H. (2002). Implicit learning of a recursive rule in an artificial grammar. *Acta Psychologica, 111*, 323-335.

Reber, A.S. (1967) Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*, 855–863.

Rodriguez, P. (2001). Simple recurrent networks learn context-free and context-sensitive languages by counting. *Neural Computation*, *13*, 2093-2118.

Spivey, M. (2007). *The Continuity of Mind*. New York: Oxford University Press.

Tabor, W. (2000). Fractal encoding of context-free grammars in connectionist networks. *Expert Systems*, *17*, 41-56.

Tabor, W. (2009). A dynamical systems perspective on the relationship between symbolic and non-symbolic computation. *Cognitive Neurodynamics, 3*, 415-427.

Wiles, J. & Elman J. (1995). Learning to count without a counter: a case study of dynamics and activation landscapes in recurrent networks. *Proceedings of the 17th Annual Cognitive Science Conference* (pp. 482-487). Mahwah, NJ: Lawrence Erlbaum Associates.