Colorless Green Ideas Sleep Furiously Revisited: A Statistical Perspective

Florencia Reali (fr34@cornell.edu)
Rick Dale (rad28@cornell.edu)
Morten H. Christiansen (mhc27@cornell.edu)

Department of Psychology; Cornell University; Ithaca, NY 14853 USA

Abstract

In the present study we provide empirical evidence that human learners succeed in an artificial-grammar learning task that involves recognizing grammatical sequences whose bigram frequencies from the training corpus are zero. This result begs explanation: Whatever strategy is being used to perform the task, it cannot rely on the simple co-occurrence of elements in the training corpus. While rule-based mechanisms may offer an account, we propose that a statistical learning mechanism is able to capture these behavioral results. A simple recurrent network is shown to learn sequences that contain null-probability bigram information by simply relying on distributional information in a training corpus. The present results offer a simple but stark challenge to previous objections to statistical learning approaches to language acquisition that are based on sparseness of the primary linguistic data.

Introduction

The importance of statistical structure in language learning and processing has been a matter of intense debate. Initial data-driven empirical approaches embraced the idea that word co-occurrences are important sources of information in language processes (e.g., Harris, 1951). This approach fell out of favor in the 1950's, in part due to the influential work of Noam Chomsky (1957) who believed that language behavior should be analyzed at a much deeper level than its surface statistics. In one of his most famous examples, he pointed out that it is reasonable to assume that neither the sentence (1) *Colorless green ideas sleep furiously* nor (2) *Furiously sleep ideas green colorless* has ever occurred, and yet (1), though nonsensical, is grammatical, while (2) is not. Therefore, a common argument against statistical approaches to language is that there are sentences containing low or zero probability sequences of words that can nonetheless be judged as grammatical. As Chomsky remarked, "… we are forced to conclude that ... probabilistic models give no particular insight into some of the basic problems of syntactic structure" (Chomsky, 1957, p. 17). Most theoretical linguists have accepted this argument, developing little interest in the role of statistical approaches to language.

Recently there has been a reappraisal of statistical approaches, partly motivated by research indicating that distributional regularities may provide an important source of information for bootstrapping syntax (e.g., Redington, Chater & Finch, 1998; Mintz, 2002)—especially when integrated with prosodic or phonological information (e.g., Morgan, Meier & Newport, 1987; Monaghan, Chater & Christiansen, in press). Moreover, statistical approaches have been supported by recent research demonstrating that young infants are sensitive to statistical information inherent in bigram transitional probabilities (e.g., Saffran, Aslin & Newport, 1996; –for a review, see Gómez & Gerken, 2000). These studies demonstrate that at least some learning mechanisms employed by infants are statistical in nature. However, as suggested by the perceived grammaticality of sentences like (1), human learning capacities certainly need to go beyond the information conveyed by item co-occurrences. In the present study we explore the extent to which humans are capable of learning the regularities of an artificial grammar, and generalizing them to new sentences in which transitional probabilities are completely uninformative. The task involves "discovering" the underlying regularities and using them to recognize sequences in which the bigram transitions are completely novel. We find that humans perform well in this task.

Two possible explanations could account for these results. First, as previously suggested (Marcus, Vijayan, Bandi Rao & Vishton, 1999), it could be that humans possess at least two learning mechanisms, one for learning statistical information and another for learning "algebraic" rules. Thus, regardless of available statistics, we could rely on open-ended abstract relationships into which we substitute arbitrary items. In an artificial-grammar learning scenario, we could know the structure or rules underlying a grammar and substitute variables with specific examples by mechanisms independent of the surface statistical information. This rule-based mechanism could therefore account for our ability to successfully generalize to sequences with uninformative bigram probabilities. Alternatively, we suggest that there is a second and equally plausible account. In this paper we demonstrate that this generalization can be accounted for on the basis of distributional learning. In the second part of this paper, we show that a simple connectionist model, trained purely on distributional information, is capable of simulating correct grammaticality judgments of test sentences that comprise bigram transitions absent in the training corpus. These

results build on previous work showing that lexical categories can emerge naturally from learning processes inherent to the SRN's distributionally driven internal representations (Elman, 1990). They also demonstrate that the distributed nature of SRNs' storage allows generalization that goes beyond traditional computational models (such as simple n-gram models) whose limitations motivated a historical shift away from statistical approaches. While these models are sensitive only to the information in the co-occurrence of word sequences, SRNs go beyond co-occurrence information, being capable of forming useful representations of lexical classes. This study is therefore important in demonstrating the need to look deeper at learning properties of more sophisticated distributional models, such as connectionist networks, in order to reassess the claims of weakness many cast onto a statistical approach to language learning and processing.

Experiment 1: Learning Null-Probability Sequences

In this experiment, we explore whether learners are capable of generalizing to novel sequences after being exposed to examples from a constrained subset of all possible grammatical sequences. Crucially, participants will be asked to recognize sequences whose bigram transitions did not occur the training corpus.

Method

Subjects Forty-nine undergraduate participants were recruited at Cornell University in exchange for extra credit in psychology classes.

Materials The stimuli were sequences of capital letters generated from a simple artificial phrase-structure grammar defined as follows:

S	→	Adj N V Adv
Adj	→	{adj_1 adj_2 adj_3}
N	→	{n_1 n_2 n_3}
V	→	{v_1 v_2 v_3}
Adv	→	{adv_1 adv_2 adv_3}

Note that the vocabulary of the grammar consists of 12 words, 3 in each of lexical categories of adjective (adj_n), noun (n_n), verb (v_n), and adverb (adv_n). The stimuli we used consisted of twelve consonants, C, Q, M, P, X, S, W, Z, K, H, T and L, which represented each of the twelve words of the vocabulary respectively (adj_1, adj_2, and adj_3 = C, Q, and M; n_1, n_2 and n_3 = P, X, and S; v_1, v_2, and v_3 = W, Z, and K; adv_1, adv_2, and adv_3 = H, T and L). Grammatical sequences consisted of a four letters string where the first one is an adjective, followed by noun, a verb and an adverb in that order.

Participants were presented with sixty grammatical sequences in a training phase. The test session comprised a set of nine grammatical and nine ungrammatical sequences. Both grammatical and ungrammatical sequences in the test set contained *at least* one bigram transition (co-occurrence of two letters) that had never been presented in the training set. To accomplish that, strings containing the following bigram transitions were excluded from the training set: adj_1 n_1 (C P), n_1 v_1 (P W), v_1 adv_1 (W H).

Grammatical sequences in the test set fell under one of three categories: The first category included sentences with just one null-probability transition ([adj_1 n_1] **X** adv_1; **X** [n_1 v_1] **X**; adj_1 **X** [v_1 adv_1], where "**X**" represents some arbitrary grammatical word); the second set contained two null-probability transitions ([adj_1 n_1 v_1] **X**; **X** [n_1 v_1 adv_1]) and the third category sentences had sentences containing three null-probability transitions ([adj_1 n_1 v_1 adv_1]). The latter category represents the artificial version of "colorless green ideas sleep furiously". The test set itself contained six grammatical sequences of the first category, two sequences of the second category and one sequence of the third category. However, each sequence was presented twice in random order, thus, participants saw a total of eighteen grammatical sequences in the test session.

Ungrammatical sequences fell under one of two categories: In the first category two words were interchanged (**n_1 adj_1** v_1 adv_1; adj_1 **v_1 n_1** adv_1; adj_1 n_1 **adv_1 v_1**; **v_1** n_1 **adj_1** adv_1; **adv_1** n_1 v_1 **adj_1**; adj_1 **adv_1** v_1 n_1); and in the second category all words were interspersed (**n_1 adj_1 adv_1 v_1**; **v_1 adv_1 adj_1 n_1**; **adv_1 v_1 n_1 adj_1**). Each sequence was presented twice in random order, thus, a total of eighteen ungrammatical sequences comprised the test session.

Procedure The experiment was conducted using the Psyscope experimental software package (Cohen, MacWhinney, Flatt, & Provost, 1993) with stimuli presented on a computer monitor. Participants were instructed that they were participating in a memory experiment. They were told that in the first part of the experiment they would see sequences of letters displayed on the screen and had to type the sequence they just saw. Each sequence was presented individually for a period of 4 seconds. The 60 sequences of the training set were presented twice, for a total of 120 input exposures, presented in random order. Immediately after seeing each sequence, participants typed it using the computer keyboard, before going to the next one.

After the training phase, participants were instructed that they would be exposed to a new set of sequences some of which were "similar" to the ones they saw in the first part of the experiment and some "dissimilar." They were instructed to press a button marked "YES" or "NO" according to whether they thought a presented string was similar to the ones they saw in the previous phase. The participants were instructed that they would probably find the task difficult and therefore they should follow their first impression without spending too much time thinking about each sequence. Each of the 18 sequences comprising the test set (9 grammatical and 9

ungrammatical) was presented twice, and all of them were randomly interspersed.

Results and discussion

The mean number of correct endorsements on the 36 test items was 25.30 (70.28%). A one-way t-test indicates that this performance is significantly above chance ($t(48) = 10.32$, $p < .0001$). We explored the percentage of correct endorsements across grammatical and ungrammatical categories. As illustrated in Fig. 1, the grammatical sequences containing one, two and three novel bigram transitions were correctly recognized as grammatical 72.2%, 64.9% and 59.2 % of the time respectively, whereas ungrammatical sequences with two or all letters interchanged, were *incorrectly* labeled as grammatical 28.8% and 27.9% of the time respectively. We also computed planned comparisons between the number of *yes*-responses (grammatical-labeling) elicited by each of the three grammatical categories vs. the number of yes-responses elicited by the ungrammatical sentences. We found that each of the three different types of grammatical sentences elicited significantly more *yes*-responses than the ungrammatical sentences (all p's < .001). These results indicate that grammatical sequences with one, two or three null-probability bigram transitions are successfully distinguished from ungrammatical sequences. Importantly, subjects are capable of learning the pattern after being exposed to only a small number of examples.

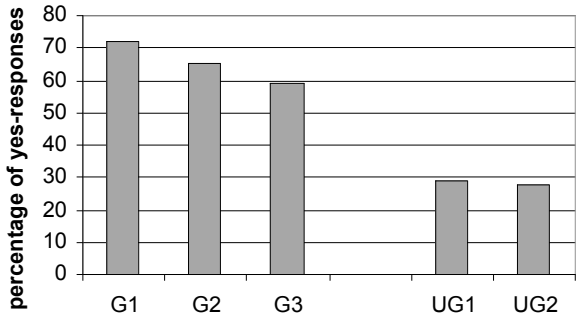


Fig. 1: Percentage of elicited yes-responses across subjects. G3, G2, G1 = Grammatical sentences comprising three, two and one probability-cero transitions respectively; UG1, UG2 = Ungrammatical sentences comprising two words interchanged and all words interspersed respectively.

These results are not necessarily surprising: We know humans are good at making grammaticality judgments of sentences they have not previously encountered. Thus, the crucial question is what kind of learning mechanism underlies success in this task. In particular, it is not clear whether these results reflect the manifestation of rule-based learning mechanism of the kind proposed by Marcus et al. (1999), or alternatively, whether these results might reflect emergent learning resulting from acquisition processes that rely only on statistical information.

In order to address this question we performed a series of computational simulations in which SRNs were trained using purely distributional cues, and without any labeling of lexical categories. After the training of the network, we tested sentences in which bigram frequencies were zero in the training corpus.

Experiment 2: Connectionist Learning of Null-Probability Sequences.

In this simulation, simple recurrent networks (SRN; Elman, 1990) are trained to predict the next word in a sentence given a corpus of sentences generated by an artificial grammar. Each word was assigned a unique vector consisting of 0s and a single 1 in a so-called *localist representation*. The representation deliberately deprives the network of any information about grammatical category, such as its syntactic distribution or semantics, etc. This type of input and output representation is the same as the one originally used by Elman (1990), and is often employed in connectionist simulations. It is important to note that the only type of information the network can rely on to learn the grammar is the distribution of these localist representations presented sequentially. As a new word is input, the network's task is to predict the next word in the sentence. As in the experiment, we prevented the network from being exposed to certain sequences of words during training. This constraint allowed us to create grammatical test sentences in which all transitions had null probability, that is, sentences in which consecutive words never co-occur in the training set.

While we are not postulating SRNs as *exact* emulators of human learning mechanisms here, we argue that they can be viewed as a model of what can be acquired by a system that is not dependent on rule-based mechanisms. Indeed, the SRN is well suited for such simulations, and has been successfully applied to a wide range of language learning and processing phenomena (e.g., Elman, 1990; Cleeremans, 1993; Christiansen & Chater, 1999). Importantly, neural networks are not simply lookup tables; instead, they are statistically-driven function approximators capable of complex generalization in a human-like fashion (Elman, 1993).

Additionally, although the task performed in Experiment 1 is not identical to the SRN's prediction task, they share the fact that both involve learning an artificial grammar and generalizing to new sentences in which transitional probabilities are uninformative.

Method

Networks The SRNs were used with initial weight randomizations in the interval [-0.1; 0.1]. Learning rate was set to 0.1, and momentum to 0.9. Each input to the network contained a localist representation of the incoming word. With a total of 36 different words and a pause marking boundaries between utterances, the network had 37 input units. The network was trained to predict the next word in a sequence, and thus the number of output units was 37. Each

network additionally had 40 hidden units and 40 context units.

Materials We trained and tested the network on an artificial grammar, containing a vocabulary composed of 8 adjectives, 12 nouns, 10 verbs, 6 adverbs. While we have equal numbers of category members in Experiment 1, we chose this distribution to meet loosely the distribution of such classes in a natural language such as English. The training corpus contained 500 sentences. Sentences were generated from a simple artificial grammar defined as follows:

S	→	Adj Adj N V Adv
S	→	Adj N V Adv
S	→	Adj N V
S	→	N V Adv
S	→	N V
Adj	→	$\{adj_1, \ldots, adj_8\}$
N	→	$\{n_1, \ldots, n_{12}\}$
V	→	$\{v_1, \ldots, v_{10}\}$
Adv	→	$\{adv_1, \ldots, adv_6\}$

Ten different training sets were generated using a random algorithm to create sentences. Importantly, all sentences were created according to the following restriction: Some of the words from each lexical category were prevented from occurring next to other ones. Specifically, the following sequences were not allowed to co-occur in the training set: Adj_2 never occurred after to Adj_1, N_1 never occurred after Adj_2, V_1 never occurred after N_1, and Adv_1 never occurred after V_1. This generation constraint allowed us to produce the following grammatical test sentence in which all transitional probabilities (bigram frequencies) had null probability in the training corpus: Adj_1 Adj_2 N_1 V_1 Adv_1. This test sentence represents is a toy-model version of the famous "Colorless green ideas sleep furiously".

The test set consisted in three target sentences, all of which had probability-zero transitions but varied in degree of grammaticality:

1) Grammatical:	Adj_1 Adj_2 N_1 V_1 Adv_1
2) Ungrammatical type I:	*Adj_1 $\mathbf{N_1}$ $\mathbf{Adj_2}$ V_1 Adv_1
3) Ungrammatical type II:	***$\mathbf{Adv_1}$ $\mathbf{V_1}$ $\mathbf{N_1}$ $\mathbf{Adj_2}$ $\mathbf{Adj_1}$**

Note that in 2) the ungrammatical sentence consists in only a single interchange of words with respect to Sentence 1, while in 3) the ungrammatical sentence consists in the complete reversal of Sentence 1. Thus, 3) corresponds to a toy-model version of the famous "Furiously sleep ideas green colorless". We want to explore whether the network is sensitive to distances between the grammatical sentence 1) and the two ungrammatical versions 2) and 3). We therefore expect that sentence 3) elicits a higher error than sentence 2), and conversely, we expect that sentence 2) elicits a higher error than sentence 1).

Procedure An SRN was trained on a single training set and tested. The training consisted of 5 passes through the training corpus. Performance was assessed based on the networks' ability to predict the next word given the prior context. In order to compute statistical comparisons we repeated the procedure with the ten different training corpora using different initial connection weights.

Results and discussion

Each word was represented by the activation of a single unit in the output layer. After training, SRNs trained with localist output representations will produce a distributional pattern of activation closely corresponding to a probability distribution of possible next items. In order to assess the overall performance of the SRNs, we computed the average mean square error (MSE) in predicting the next word across each test sentence.

Results are displayed in Fig. 2: The average MSEs were 0.75, 0.82, and 0.95 for grammatical, ungrammatical type I, and ungrammatical type II respectively. We found that the difference between the MSE elicited by grammatical sentences was significantly lower than the MSE elicited by ungrammatical type I ($t(9) = 4.66$, $p < 0.005$) and ungrammatical type II ($t(9) = 13.15$, $p < 0.001$). To establish a baseline, we also computed the average MSE elicited across all the sentences contained in the training set after the training stage. Interestingly the difference of MSE between the grammatical test sentence comprising null-probability bigram transitions and the MSE elicited by grammatical sentences contained in the training set was not statistically significant ($t(9) = 0.13$; $p = 0.84$), suggesting that the network recognized the novel sentence as one of the training set.

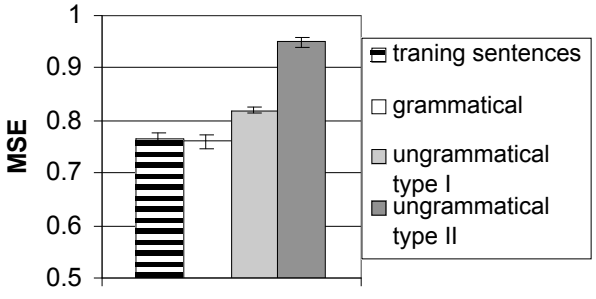


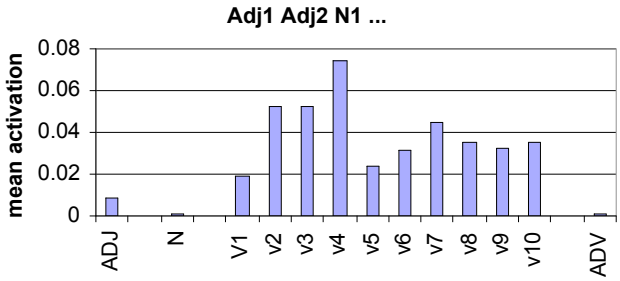
Fig. 2: Mean square error across words in four type of sentences: Striped pattern: average across all words in the training sent; White: grammatical test sentence comprising null-probability transitions; Light gray: ungrammatical type I test sentence; Dark gray: ungrammatical type II test sentence. Displayed values result from the average across the five simulations using different training sets.

But how do the SRNs stack up against more traditional statistical models whose weaknesses compelled Chomsky (1957) to abandon probabilistic methods? N-gram models are standard statistical models used in psycholinguistics that

are based on co-occurrences of words in natural language corpora. Traditional n-gram models trained on the same corpus here would therefore assign equal probability to test sentences (1), (2), and (3) above. The results obtained here demonstrate that SRNs are capable of going beyond n-gram models in generalizing to new input.

As an illustration, Figure 3 shows the mean activation of the output units at different points in the sequence. The graph in Fig. 3A shows the averaged mean activation of the SRNs after being presented with the test sequences of words: "Adj_1 Adj_2 N_1 …". The figure shows the averaged mean activation of the units corresponding to adjectives (ADJ), noun (N), and adverbs (ADV), while the activations for each of the individual verb units (V_1 through V_{10}) are shown in detail. Even though V_1 never occurred after N_1 in the training set, the activation of V_1 elicited by the string "Adj_1 Adj_2 N_1 …" is comparable to the activation of other verbal verbs, such as V_5. Fig. 3B shows the averaged mean activation of the SRNs after being presented with the test sequences of words: "Adj_1 Adj_2 N_1 V_1 …". The activation of adjective, noun, and verbs units are shown averaged, while the activations of each of the adverb-units (Adv_1 through Adv_6) are shown in detail. The activation of Adv1 elicited by the string "Adj_1 Adj_2 N_1 V_1…" is comparable to (and in some cases higher than) the activation of other adverbial units, despite the fact that Adv_1 never occurred after V_1.

A



B

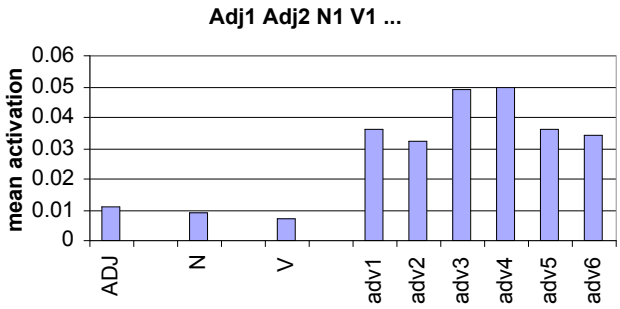


Fig. 3: Mean activation across different units elicited by previous context. A) Activation elicited by the word substring: "Adj1 Adj2 N1 …". B) Activation elicited by the word substring: "Adj1 Adj2 N1 V1…". ADJ, N, V, ADV: mean activation across adjective, noun, verb, and adverb units respectively. v1, v2,…,v10, adv1, adv2,…,adv6: Individual activation of verbal and adverb units respectively.

The activation values displayed in Fig. 3 illustrate that the networks are successfully learning to predict the next lexical class. These results demonstrate that SRNs trained purely on distributional information are sensitive to grammaticality differences between different sentences in which bigram transitions have null probabilities.

General Discussion

Even among the billions of words in available databases, innumerable reasonable sentences remain absent. This so-called *sparse data problem* continues to be a serious challenge not only for the study of human language acquisition and processing, but also in the area of artificial intelligence devoted to natural language processing (see Lee, 2004). The results of Experiment 1 reveal that humans become sufficiently sensitive to the regularities of training examples to recognize novel sequences whose bigram transitions are absent in training. Therefore, subjects must be relying on something other than co-occurrence of consecutive elements to generalize from our experimentally induced sparse sentence samples. The remaining question concerns what type of cognitive mechanism can accomplish this task. One such mechanism might be the rule-based learning mechanism recommended by Marcus et al. (1990) and others (e.g., Peña, Bonatti, Nespor, & Mehler, 2002), which does not rely on statistical learning. Alternatively, the implicit knowledge of the underlying regularities needed to succeed in the task could be acquired by distributional learning through training exemplars. Our connectionist simulations in Experiment 2 provide some evidence that the latter alternative should be considered.

It has been previously argued by Elman (1990, see also Elman, 2004) that SRNs are capable of forming internal representations of grammatical classes from distributional information. The present findings build on that idea, showing that SRNs are capable of good performance in the prediction task even in sentences having null transitional probabilities relative to the training corpus. Previous studies have demonstrated that the network uses distributional information to induce categories. These categories are reflected in the analysis of hidden unit activations evoked in response to each word (e.g., Christiansen & Chater, 1999; Reali et al. 2003; see also Elman, 2004). These analyses involve measures in terms of Euclidean distance in the hidden unit space, representing the similarity of words' hidden unit activations, and cluster according to lexical categories. Interestingly, the present results show that SRNs' prediction of the next word seems to be at least in part determined by lexical category membership, rather than being determined by specific word co-occurrences in the training corpora. This is an important achievement for a distributional learning mechanism, seeing as it was not provided with information about grammatical classes. Traditional n-gram models of language are not capable of representing lexical classes in the same way. Standard bigram or trigram models trained on our artificial grammar, would assign exactly the same probability to all our test

sentences. SRNs seem to be effective in learning something that goes beyond surface properties of language, suggesting they could be understood as "regularity discoverers" rather than mere statistical learners resembling n-gram models. The present results are consistent with previous arguments about connectionist models' generalization properties (Christiansen & Chater, 1994). Recently, Allen & Seidenberg (1999), used connectionist simulations to show that low probability sentences like (1) could be statistically learned when other information such as word types or semantics are used in its comprehension. Simulations in Experiment 2 build on these previous studies by demonstrating that *pure* distributional information can provide a basis in the process of learning low probability sentences.

In order to dismiss statistical approaches to language, particularly through limitations imposed by sparse data issues, it is necessary to thoroughly understand the learning capabilities of systems such as connectionist models. The present results challenge one of the most well-established objections to statistical approaches, which might be based on an underestimation of the ability of connectionist models to deal with sparse input. One of the principal arguments for innateness of grammar, often referred to as "Poverty of the Stimulus" logic (e.g., Crain & Pietroski, 2001), is based on precisely that property of the linguistic data: sparseness. It is therefore crucial to determine the extent to which connectionist models and statistical approaches in general can overcome some of the difficulties related to the sparseness of the linguistic input. The present study constitutes a step in that direction.

References

Allen, J., & Seidenberg, M.S. (1999). The emergence of grammaticality in connectionist networks. In B. MacWhinney (Ed.) *The Emergence of Language*. Mahwah, NJ: Lawrence Erlbaum.

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*. Mouton and co.: The Hague.

Christiansen, M.H. & Chater, N. (1994). Generalization and connectionist language learning. *Mind and Language*, *9*, 273-287.

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

Cleeremans, A.. (1993). Attention and awareness in sequence learning. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*. (pp. 330-335). Mahwah, NJ: Lawrence Erlbaum.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1992). Psyscope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers, 25*, 257-271.

Crain, S., & Pietroski, P. (2001). Nature, nurture and universal grammar. *Linguistics and Philosophy, 24*, 139-186.

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

Elman, J.L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48*, 71-99.

Elman, J.L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Science, 7*, 301-306.

Gómez, R.L., & Gerken, L.A. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4*, 178-186.

Harris, Z. (1951). *Methods in Structural Linguistics*. University of Chicago Press. Reprinted by Phoenix Books in 1960 under the title *Structural Linguistics.*

Lee, L. (2004). I'm sorry Dave, I'm afraid I can't do that: Linguistics, statistics, and natural language processing circa 2001. In *Computer Science: Reflections on the Field, Reflections from the Field.* (pp. 111-118). Washington, DC: National Academies Press.

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

Mintz, T.H.(2002) Category induction from distributional cues in an artificial language. *Memory & Cognition, 30*, 678-686.

Monaghan, P., Chater, N. & Christiansen, M.H. (in press). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition.*

Morgan, J. L., Meier, R.P., & Newport, E.L. (1987). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology, 19*, 498–550.

Peña, M., Bonatti, L.L., Nespor, M., & Mehler J. (2002). Signal-driven computations in speech processing. *Science*, *298,* 604-607.

Reali, F., Christiansen, M.H. & Monaghan, P. (2003). Phonological and Distributional Cues in Syntax Acquisition: Scaling up the Connectionist Approach to Multiple-Cue In *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 970-975). Mahwah, NJ: Lawrence Erlbaum.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*, 425-469.

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