Modeling Implicit and Explicit Processes in Recursive Sequence Structure Learning

Jamie D. Alexandre (jdalexan@ucsd.edu)
Department of Cognitive Science, 9500 Gilman Drive
La Jolla, CA 92093 USA

Abstract

Recursive structure is viewed as a central property of human language, yet the mechanisms that underlie the acquisition and processing of this structure are subject to intense debate. The artificial grammar learning paradigm has shed light onto syntax acquisition, but has rarely been applied to the more complex, context-free grammars that are needed to represent recursive structure. We adapt the artificial grammar serial reaction time task to study the online acquisition of recursion, and compare human performance to the predictions made by a number of computational language models, chosen to reflect multiple levels and types of syntactic complexity (n-grams, hidden markov models, simple recurrent networks, and Bayesian-induced probabilistic context-free grammars). Evidence is found for a dissociation between explicit and implicit mechanisms of sequence processing, with the SRN more highly correlated with implicit performance, and the PCFG more correlated with explicit awareness of the sequential structure.

Keywords: artificial grammar learning; syntax; recursion; serial reaction time task; simple recurrent network; context-free grammars; implicit/explicit processes.

Introduction

The nature of linguistic structure, and the computational mechanisms by which humans comprehend it, have long been subject to heated debate. Recursion – the ability to hierarchically embed elements within instances of themselves – has been a central point of contention. Although the recursive structure of language was not a new idea at the time, Chomsky formalized the notion of syntactic recursion, touting it as *the* fundamental property that allows for human linguistic ability, a thesis he continues to popularize today (Chomsky, 1956; Hauser, Chomsky, & Fitch, 2002).

In the Chomskyan tradition, the human syntactic system implements a set of rules that allow for theoretically unbounded levels of recursive embedding ("competence"), but this system is then subject to processing constraints, such as working memory limitations, that explain our limited ability to process recursive structures beyond a few levels of embedding ("performance"). Other theorists, particularly from the connectionist camp, have attempted to explain the (limited) human ability to process recursive structure without hypothesizing unbounded competence, by modeling syntactic processing in systems that do not make use of rules or explicit representations (e.g. Elman, 1990; Pollack, 1990; Christiansen & Chater, 1999).

The artificial grammar learning paradigm (initiated by Reber, 1967) has been used to examine processes of syntactic acquisition, but this has been largely restricted to the class of regular grammars, which doesn't shed light onto the acquisition or processing of context-free or recursive structure. The goal of the present study is to obtain estimates of a subject's online string continuation expectancies while responding to sequences generated by a context-free grammar (palindromes), so that these may be compared with the predictions made by a variety of language models trained on the same input history as the subject. The traditional measures of successful acquisition in artificial grammar experiments – such as grammaticality judgments or recall error rates – are not able to provide the incremental (symbol-by-symbol) expectancy data that we require. We adapt a paradigm first employed by Cleeremans & McClelland (1991), known as a serial reaction time task, in which subjects respond to a sequences of stimuli (with a button mapped onto each stimulus class) by pressing the corresponding button as quickly as possible after perceiving stimulus onset. The resulting reaction times are then correlated with the probabilities generated by the competing computational models.

Surprisal

Surprisal, or self-information, is a notion from information theory that quantifies the amount of novel information that a particular event carries with it. An event's surprisal is defined as its negative log probability:

$$-\log(P(x \mid \text{context}))$$

The concept of surprisal has been used in psycholinguistics as a potential measure of incremental processing difficulty, and is thus expected to correlate with behavioral measures such as reading times in eye-tracking studies, and response times in self-paced reading studies (Hale, 2001; Levy, 2008).

The surprisal model requires that we adopt some measure of the probability of a word's occurrence given the preceding sentential context. Hale (2001) uses a probabilistic Earley parsing algorithm to generate incremental word probabilities, using the resulting surprisal values to explain the garden path effect. Levy (2008) uses a similar model to explain a wide-range of effects found in the psycholinguistic literature, such as predictability (e.g. effect of Cloze probability), locality effects (e.g. preference for local dependencies), competition/dynamical models (e.g. greater ease in highly constrained contexts), the tuning hypothesis (e.g. effect of structural frequency), and connectionist models (e.g. predictions made by an SRN). The case of the SRN is particularly interesting, because there are significant divergences between the predictions made by an SRN and a

PCFG-based surprisal model, particularly for constructions such as recursive center-embeddings, which PCFGs process flawlessly, and SRNs – much like humans – have difficulty processing beyond a few levels of embedding (Christiansen & Chater, 1999).

Frank (2009) tested a surprisal model against human eye-tracking data from the Dundee corpus, comparing PCFG-with SRN-generated probabilities, and found that the PCFG produced more accurate objective probabilities, but that the SRN produced probabilities that better matched the human data. He concludes from this, firstly, that subjective probabilities diverge from the objective probabilities, and secondly, that the SRN may in fact be a better model of human performance. Other surprisal studies have used n-gram statistics, such as a trigram model with Kneser-Ney smoothing (Smith & Levy, 2008), and also shown close correspondences with human eye-tracking data.

Language Models

A probabilistic language model is a distribution over the strings (sentences) in a language. The models considered in this paper also all support incremental prediction; that is, given a sentence prefix, they assign a distribution over the symbols that might come next.

To allow for comparison with the human data, each of the models is trained on the precise input that a subject has been exposed to at every point in the experiment (rather than training on a larger corpus, or simply using the probabilities assigned by the model that generated the stimuli). This allows us to observe how a subject's predictions change over the course of learning, to gain insight into the rate at which a system is acquired, as well as possible shifts in strategy, rather than simply comparing fully trained systems.

It is also important to note that none of the model parameters are fit to the human data; a model is trained to predict a sequence's continuation based on the set of sequences it has seen up to that point in the experiment, making use of the algorithmic and representational resources at its disposal, but agnostic to human performance.

The models were chosen from amongst those most commonly used within computational linguistics to model sequential structure, at various levels of complexity (some corresponding roughly to levels in the Chomsky hierarchy).

N-grams (bigrams/trigrams)

One of the simplest but most commonly used language models, n-grams calculate the probability of a symbol in terms of the frequency with which it occurs in its immediately preceding context. Here, we will consider bigrams (which take into account the preceding symbol of context) and trigrams (which take into account the 2 preceding symbols). The predictions made by the n-grams at every step were based on training on all preceding sequences (excluding the sequences that had not yet been seen).

Hidden Markov Model (HMM)

Whereas n-gram transition probabilities are defined between sets of adjacent words, the transitions in a hidden markov model (HMM) are defined over a set of "hidden" states, and these states, in turn, generate the individual words. The idea is that there is an underlying "hidden markov process" that we cannot access directly, and all we can observe is the final sequence of words that is produced by this underlying state sequence. Computationally, HMMs roughly correspond to regular languages at the bottom of the Chomsky hierarchy.

We use the standard Baum-Welch algorithm (Baum et al, 1970) to estimate the HMM's transition and emission matrices from the training corpus (the preceding sequences) for an HMM with 5 hidden states. The trained HMM is then used to compute the incremental posterior probabilities of each symbol given its preceding context. As always, the predictions only used the preceding sequences as a training corpus (so as to be comparable to the human data).

Simple Recurrent Network (SRN)

A simple recurrent network (SRN) is a standard three-layer feed-forward network, with the addition of a context layer that maintains a copy of the hidden layer's state from the previous timestep, and then allows the nodes in this context layer to feed back into the hidden layer during the next timestep, alongside the next input (Elman, 1990). The context layer in an SRN effectively implements time-tapped feedback loops from every node in the hidden layer back to each of the nodes in the hidden layer (delayed by one timestep). The addition of recurrent hidden layer connections allows an SRN to learn to use its hidden layer representations to maintain *task-relevant* contextual information over theoretically unbounded (though in most cases, rapidly decaying) distances.

The SRN used in this paper contained 9 input nodes (one for each symbol, plus a sequence boundary marker), 16 hidden nodes, and 9 output nodes. The network was trained using standard back-propagation, with a learning rate of 0.5 and no momentum, on a single pass through the sequences. Output activations at every timestep were converted into probabilities through the Luce choice rule (in effect, normalizing the network's output vector).

Probabilistic Context-Free Grammar (PCFG)

Context-free grammars (CFGs) have played a central role in linguistic theories of syntax ever since Chomsky (1956) proposed them as being necessary (and almost sufficient) to account for the types of recursive phrase structure observed in human language. A probabilistic context-free grammar adds probabilities to the production rules in a context-free grammar, allowing us to calculate a distribution over strings in the language.

Once we know the parameters of the grammar (see below), incremental predictions can be computed as follows (adapted from Jelinek & Lafferty, 1991):

1. The probability of a string is the sum of the probabilities of all its parse trees.
2. The probability of a string prefix is a sum over the probabilities of all possible completions of the prefix.
3. The probability that a particular symbol w_i will appear following the string prefix $w_1..w_{i-1}$ can be computed by dividing the probability of the prefix with that symbol appended, $P(w_1..w_i)$, by the probability of the prefix, $P(w_1..w_{i-1})$

Stolcke (1995) modified the Earley parsing algorithm to compute the above incremental probabilities efficiently, and we use an implementation by Levy (2008) in the present work.

Learning the parameters of a PCFG from an unparsed corpus is not a trivial task, however. Here, we use a Bayesian framework developed by Mark Johnson[1] that uses Gibbs sampling to learn the probabilities for a set of production rules, given a corpus of training sequences. All combinations of production rules with 8 states (in Chomsky Normal Form, e.g. A->BC) were included in set of candidate rules, and the sampler was given a prior of alpha=0.0001. The counts on the final sample grammar were normalized into probabilities. As with all the other models, the predictions made for every symbol were based on re-training after every sequence, using only on the sequences that occurred prior to that point in the experiment, so that the models have precisely the same information available to them at each timestep as the human subjects. This entire process was repeated 5 times, and the resulting sequences of probabilities were averaged together.

Experiment

Methods

Interface Care was taken in designing and constructing an interface device for the task, due to concerns about measurement noise. The button box (Figure 1) consists of 8 finger-sized push buttons arranged in a 2x4 array, with each button containing its own separately controllable LED for use as a response cue. The buttons and LEDs are interfaced to the PC via a USB-powered LabJack U3 DAQ device, which has very high sampling rates and low command-response latencies, allowing for RTs to be measured to millisecond accuracy.

Figure 1: Button box used in experiment.

[1] http://www.cog.brown.edu/~mj/Software.htm

Participants Eight subjects (mean age 20.5, all right-handed), drawn from the UCSD undergraduate subject pool, received 2 hours of course credit for their participation.

Stimuli Sequences were generated from the following grammar in Table 1.

Table 1: Context-free grammar used to generate stimuli.

Probability	Production Rule
0.193	S → T0 S T0
0.146	S → T1 S T1
0.112	S → T2 S T2
0.128	S → T3 S T3
0.077	S → T4 S T4
0.082	S → T5 S T5
0.159	S → T6 S T6
0.103	S → T7

This grammar generates palindromes, a particular type of "mirror recursion" in which the right-hand side of the sequence is a mirror image (flipped left-to-right) of the left-hand side. The 7th symbol serves as a consistent center marker, making the grammar deterministic. An example sequence would be "0 4 1 3 7 3 1 4 0".

Palindomes are the canonical example of context-free structures, and possibly the simplest type of grammar that is context-free and thus cannot be fully captured by finite state models such as an HMM, or by n-gram statistics.

An experimental session consisted of 16 blocks of 25 sequences each, with sequences ranging in length from 5 to 15. Each of the 8 subjects were presented with the same set of sequences, but with a different mapping of symbols to buttons, shuffled in a Latin-square design such that every symbol was mapped onto each of the 8 buttons for exactly one subject (to balance out any effects of button location or between-button distances).

Procedure Subjects were told that the purpose of the experiment was to study the "effects of practice on reaction times", and were told to "hit each button as quickly as possible when that button's light goes on". No mention was made regarding the structured nature of the stimuli; as far as the subjects were concerned, the sequences were entirely random.

Sequences were presented rapidly, with the next light in a sequence turning on 120ms after the previous button had been released. After the end of an individual sequence there was a 2 second pause before the next sequence began.

In between blocks, subjects were presented with a feedback screen indicating their performance on the block relative to their performance on earlier blocks (plotting their RT contour over time), and also relative to previous subjects, by means of a highscores list derived from earlier pilot testing. Subjects were given a chance to take a short break in between blocks.

After completing the experiment, subjects were interviewed about the strategies they had employed in the

task, the factors they thought affected their performance, and what sorts of patterns (if any) they had noticed in the sequences.

Results and Analysis

Reaction times longer than 1000ms (greater than ~4.2 std above mean) were excluded from analysis, to eliminate extreme outliers caused by events not related to the task (such as distractions, subject sneezing, etc). Only 0.2% of the trials were excluded by this criterion. In addition, the first trial of every sequence was excluded from correlation analyses, as earlier pilot testing using random sequences showed that mean reaction times for these sequence-initial trials were ~70ms slower than for the remainder of the sequence. Reaction times for error trials (when the incorrect button was pressed) were measured from when the light went on to when the correct button was pressed, ignoring the intervening erroneous button press. Subjects made an average of 65 errors each (1.7% of the trials), and these trials were not excluded from the analysis, but doing so has no noticeable effect.

The median reaction time for each trial is calculated across subjects, and then the resulting sequence of reaction times is correlated with the sequences of surprisal values (negative log probability) generated by each of the models. The experiment is divided up into four parts to visualize how the correlations change over the course of training. Standard correlation coefficients and 95% confidence intervals are plotted in Figure 2. Note that each of the models is significantly correlated with the human reaction time data throughout the experiment, though with no model clearly dominating (except perhaps a slight preference for the SRN).

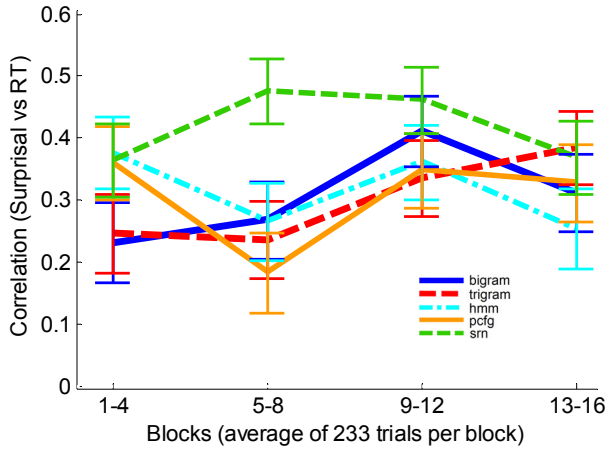

Figure 2: Correlations between models and human reaction times over the course of the experiment.

Several possible interpretations exist at this point. Since the models themselves are quite strongly inter-correlated, it is possible that the correlations for each of the models could be explained by a common shared component. In particular, each of the models is capable of representing n-gram statistics, so perhaps this could explain some portion of the correlation in the other models. To investigate this possibility, partial correlations between the human reaction times and the models are computed after regressing out the bigram and trigram statistics. The residual correlations are plotted in Figure 3.

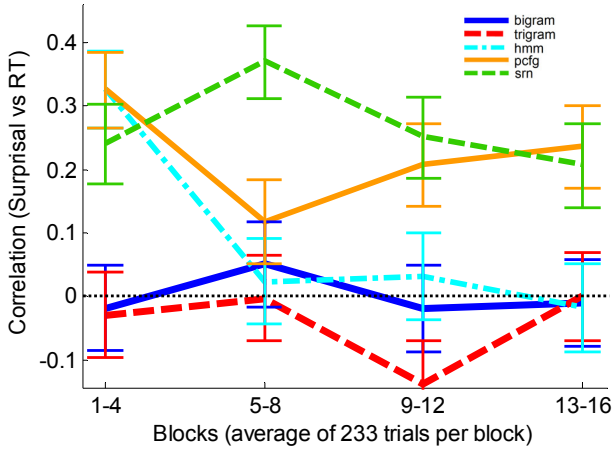

Figure 3: Partial correlations between model probabilities and reactions times, regressing out n-gram probabilities.

As is to be expected, the bigram and trigram correlations become insignificant. The HMM correlations are also eliminated after the first couple of blocks (at which point none of the models have learned very much), suggesting that the HMM was not explaining anything significant about the human behavior beyond n-gram statistics. Both the SRN and PCFG, however, maintain significant correlations throughout, suggesting that they are capturing more about the human reaction times than simply a sensitivity to n-gram statistics.

We might then wonder whether a common component is responsible for both the SRN and PCFG correlations, or if they are each accounting for distinct aspects of the human behavior. To test this, we regress out all models except for the model of interest, and see how much of the variance remains for that model to explain.

Regressing out all the models besides the PCFG reduces its correlations very slightly, but they remain highly over the course of a session, as can be seen in Figure 4 below.

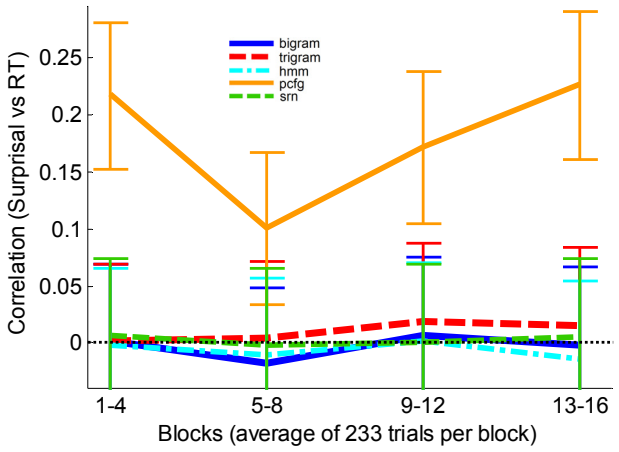

Figure 4: Partial correlations, regressing out all but PCFG.

Similarly, regressing out all models other than the SRN has very little effect on the SRN correlations, which remain strong throughout, despite declining somewhat towards the end (Figure 5).

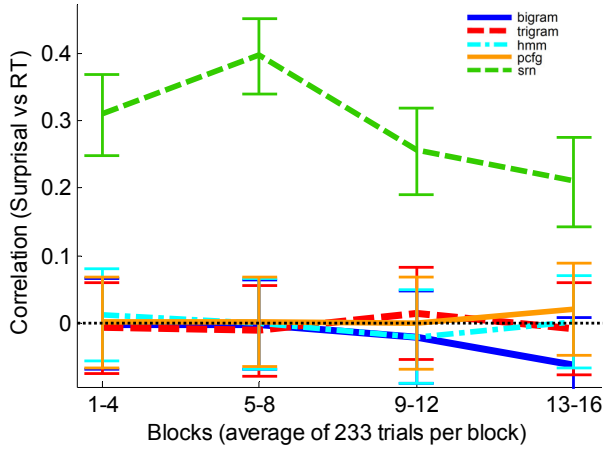

Figure 5: Partial correlations, regressing out all but SRN.

These results seem to suggest that multiple simultaneous processes are playing a role in human behavior on the task; on the one hand, an associative, incremental component captured by the SRN, and on the other hand, a more rule-based, recursive component exemplified by the PCFG. As SRN models have frequently been used to model implicit learning (e.g. Cleeremans, 1993; Misyak et al, 2009), whereas PCFGs are more often associated with explicit rule-based knowledge, we examined individual differences between subjects with regards to implicit and explicit learning, to see if this might help to explain this dissociation.

In the post-testing questionnaire, 3 of the 8 subjects identified some type of structure within the sequences; some referred to it as a "circular" or "mirror" pattern, and one also gave explicit palindromic examples. The 5 remaining subjects had not noticed any regularity to the sequences, even when probed further (2 of these "felt" like there might be some pattern, but could not articulate any details). We separated these two groups from one another and once again calculated partial correlations (regressing out n-grams and the hmm).

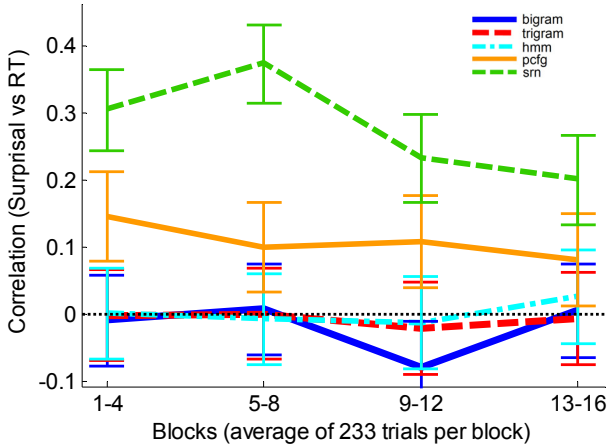

Figure 6: Subjects with no explicit awareness of structure; partial correlations, regressing out n-grams and hmm.

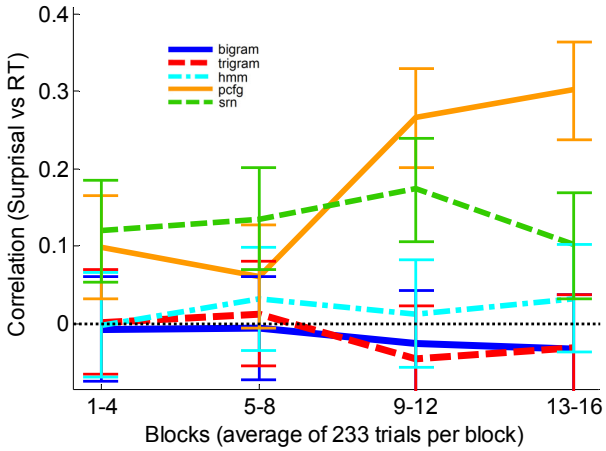

Figure 7: Subjects who were explicitly aware of structure; partial correlations, regressing out n-grams and hmm.

The subjects who were able to report explicit knowledge of aspects of the palindromic structure, by the end of the experiment, showed the strongest correlation with the PCFG (Figure 7), whereas the SRN correlated more strongly with the group that gained no explicit awareness of the structure (Figure 6), indicating that the variance explained by the SRN may reflect a more automatic, implicit processing of the sequential structure (as suggested, for example, by Cleeremans, 1993), whereas the acquisition of recursive, rule-like structures may involve more explicit, conscious processing. It was not possible to query subjects partway through the experiment about whether they had noticed any patterns without drawing their attention to the existence of structure, but the sudden divergence between the PCFG and SRN in Figure 7 lines up well with subjects' comments during the post-test interview that they had begun to notice the pattern somewhere in the "middle of the experiment".

It is also instructive to examine the pattern of reaction times over the course of an average sequence. As the sequences are of different lengths, position on the x-axis is represented as percentage of the way through a sequence (Figure 8).

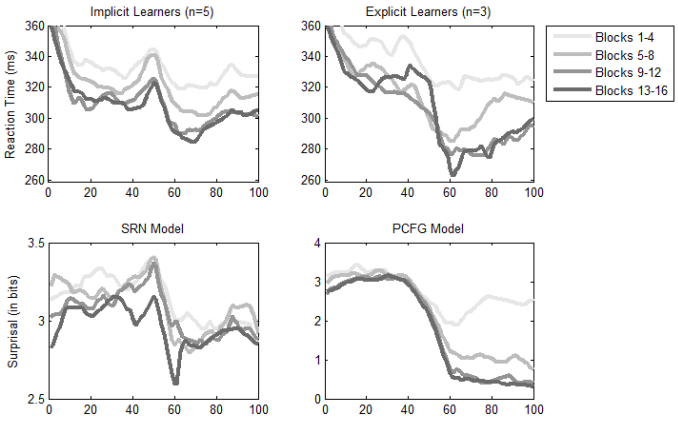

Figure 8: Comparison of RTs and model surprisal over the course of an average sequence (scaled by percentage).

There are several things to note in this reaction time data. Firstly, subjects seem to show a strong advantage in the second half of the sequence, which is consistent with the symbols in the second half being completely determined by the symbols in the first half (due to the palindromic nature of the sequences), and which is seen most strongly both in the PCFG and in the learners with explicit awareness of the structure. Secondly, this advantage is greater immediately following the center symbol and reaction time and then increases slightly as the sequence continues. This is consistent with the fact that later symbols in the second half involve longer-range dependencies, and thus may reflect working memory limitations. The reason for the peak seen halfway through the sequences in both the implicit learners and the SRN is at first unclear, but it is tempting to interpret it as reflecting the cognitive load involved in needing to flip around the first half of the sequence in order to predict the second half, although we might expect this to appear in the explicit rather than the implicit subjects.

Discussion

We attempted to shed light on the mechanisms underlying human processing of recursive structure, by extending the artificial grammar serial reaction time paradigm in two ways; firstly, by training subjects on more complex grammars than are typically used (context-free grammars); and secondly, by comparing performance not only to transitional n-gram probabilities and connectionist models, but also to a Bayesian-induced PCFG model, trained on the exact same set of sequences as the subjects. Evidence was found for a dissociation between implicit and explicit modes of processing, and these modes were seen to correlate most strongly with the predictions of the SRN and the PCFG, respectively.

It may also be fruitful to examine the effects of making subjects explicitly aware of the structure prior to beginning the task, as the results of the present study would suggest this would lead to greater correlation with the predictions of the PCFG. It would also be useful to provide a longer training period, to shed light on how these processes change over the course of more extensive exposure.

References

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164-171.

Chomsky, N. (1956). Three models for the description of language. *IEEE Transactions on Information Theory*, 2(3):113-124.

Christiansen, M. H. and Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23(2):157-205.

Cleeremans, A. (1993). *Mechanisms of implicit learning: connectionist models of sequence processing*. MIT Press.

Cleeremans, A. and McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of experimental psychology. General*, 120(3):235-253.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179-211.

Frank, S. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. *Proceedings of the 31st Annual Cognitive Science Society Conference* (pp. 1139-1144). Austin, TX: Cognitive Science Society.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of NAACL*, vol. 2: 159–166.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569-1579.

Jelinek, F. and Lafferty, J. D. (1991). Computation of the probability of initial substring generation by stochastic context-free grammars. *Computational Linguistics*, 17(3):315-323.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106:1126-1177.

Misyak, J.B., Christiansen, M.H. & Tomblin, J.B. (2009). Statistical learning of nonadjacencies predicts on-line processing of long-distance dependencies in natural language. *Proceedings of the 31st Annual Cognitive Science Society Conference* (pp. 177-182). Austin, TX: Cognitive Science Society.

Smith, N. and Levy, R. (2008). Optimal Processing Times in Reading: a Formal Model and Empirical Investigation. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (oral presentation).

Pollack, J. B. (1990). Recursive distributed representations. *Artif. Intell.*, 46(1-2):77-105.

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

Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics, MIT Press for the Association for Computational Linguistics*, 21.