Inducing a Grammar Without an Explicit Teacher: Incremental Distributed Prediction Feedback

Michael Spivey-Knowlton
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14627
spivey@psych.rochester.edu

Jenny R. Saffran
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14627
saffran@psych.rochester.edu

Abstract

A primary problem for a child learning her first language is that her ungrammatical utterances are rarely explicitly corrected. It has been argued that this dearth of negative evidence regarding the child's grammatical hypotheses makes it impossible for the child to induce the grammar of the language without substantial innate knowledge of some universal principles common to all natural grammars. However, recent connectionist models of language acquisition have employed a learning technique that circumvents the negative evidence problem. Moreover, this learning strategy is not limited to strictly connectionist architectures. What we call Incremental Distributed Prediction Feedback refers to when the learner simply listens to utterances in its environment and makes internal predictions on-line as to what elements of the grammar are more or less likely to immediately follow the current input. Once that subsequent input is received, those prediction contingencies (essentially, transitional probabilities) are slightly adjusted accordingly. Simulations with artificial grammars demonstrate that this learning strategy is faster and more realistic than depending on infrequent negative feedback to ungrammatical output. Incremental Distributed Prediction Feedback allows the learner to *produce its own negative evidence from positive examples of the language* by comparing incrementally predicted input with actual input.

Introduction

Recently, connectionist models have begun to use time as a critical factor. Rather than receiving an explicit training signal for associating arbitrary inputs with arbitrary outputs, irrespective of any temporal relationship, the model is exposed to sequences of inputs and incrementally attempts to predict what the subsequent input will be. Without an explicit teacher, the model compares its predicted subsequent input with the actual subsequent input, and uses the difference as an error signal. Some models of this type use recurrent connections to compute a prediction based on a "Gestalt" of several timesteps (e.g., Elman, 1990; Juliano & Tanenhaus, 1995; St. John & McClelland, 1990), but this is not a necessary condition in order to use this learning strategy. Standard feed-forward networks can also learn by predicting the subsequent input based on the current input (Schütze, 1994). In fact, this learning strategy need not be restricted to connectionist architectures at all. The work we present in this paper is a test case in which we compare this learning strategy with one that requires explicit corrective feedback in terms of their ability to induce a simple grammar matrix (for an example, see Figure 1).

The learning model consists of a reproduction of the matrix with initially equal values in all cells (i.e., identical connection weights for all possible sequential pairings). Starting off with this *tabula rasa* assumes no initial predisposition toward particular kinds of connectivity (i.e., no innate constraints devoted to likely patterns). The first learning strategy we simulate is derived from a standard assumption that explicitly correcting the child's ungrammatical utterances would be an optimal method of grammar induction. Such corrections are known as 'negative evidence'. The corresponding idealized model, implemented in simulations 1A and 2A, is called Explicit Negative Evidence Feedback (ENEF), in which the learner randomly produces sequential pairings that it thinks are grammatical (initially, any pairing) and occasionally receives corrective feedback from a teacher when the pairing is in fact ungrammatical.

t \ t+1	*a*	*b*	*c*	*d*	*e*	*f*
a	0	1	0	1	0	0
b	1	0	0	0	0	1
c	0	0	0	0	1	0
d	0	0	1	1	0	0
e	1	1	0	0	0	1
f	0	1	1	0	0	0

Figure 1. Elements of the language are cross-indexed with one another, and some sequential pairings (from t to t+1) are deemed grammatical (1) while others are considered ungrammatical (0). For example, along the top row, *a* can be followed by *b* or by *d*, but not by *a*, *c*, *e* or *f*. (This kind of local transition system is equivalent to a finite automaton language or a Markov chain with equal probabilities from a given state to the possible next states. It follows that results of modeling the learning of such a grammar may not generalize to that of natural grammars which are considerably more complex.)

We compare this learning strategy, within the same model architecture, to one based on the learning method employed by the connectionist models that were discussed above. Simulations 1B and 2B implement Incremental Distributed Prediction Feedback (IDPF), in which the learner listens to grammatical pairings in its environment and makes on-line predictions (i.e., multiple bets of varying magnitude) about what elements at time t+1 are more or less likely to follow the current element. The bet that wins is rewarded, thus increasing the strength of that prediction in future instances. All other pairings from that initial element are punished, with a corresponding decrease in their prediction strength.

Both learning strategies will eventually get the learner to some arbitrary criterion of accuracy (say, 95%) in its internalization of the grammar. The case for the child learning her first language, however, is special because the scarcity and variability of corrective feedback regarding the child's utterances suggests that relying on negative evidence alone would simply take far longer than children generally require to learn their first language (Marcus, 1993). The purpose of our simulations is to compare how many utterances must typically be produced by ENEF to reach 95% accuracy in its encoding of the grammar with how many utterances must typically be heard by IDPF to reach 95% accuracy in its encoding of the grammar.

Negative Evidence

In principle, negative evidence serves to correct children's misconstrued linguistic rules and to drive change in the developing grammar. The critical role of negative evidence was demonstrated in a classic paper by Gold (1967), who proved that natural languages cannot be learned in a finite period of time from a finite set of *positive* examples alone (grammatical sentences from the language). Given that children do, however, learn language with all due speed, there are two possible means to circumvent this learnability problem: 1) to receive corrections (negative evidence), or 2) to restrict the hypothesis search space through innate constraints. Within the learnability framework, then, a language acquisition device unencumbered by innate constraints requires negative evidence in response to its ungrammatical output. Without negative evidence, such a device could not converge upon the target grammar -- this fact is a primary feature of the "poverty of the stimulus" argument for innate linguistic structure (Chomsky, 1972; Pinker, 1984).

For these reasons, the question of the existence of negative evidence has received much debate. While early studies indicated that parents tend to correct the truth-value of children's utterances rather than grammatical errors (Brown & Hanlon, 1970), more recent research has found that some parents respond differentially to children's errors, for example by repeating ungrammatical sentences more often than grammatical sentences (e.g., Bohannon, MacWhinney & Snow, 1990; Bohannon & Stanowicz, 1988; Hirsh-Pasek, Treiman, & Schneiderman, 1984). However, such effects are generally small, and are not seen consistently either across subjects or across studies (Gordon, 1990; Marcus, 1993; Morgan & Travis, 1989). Moreover, even when negative evidence is available, it is not necessarily *used* by children (e.g., McNeill, 1970, Chapter 7). To date, both the availability and effectiveness of explicit negative evidence remains controversial.

There exists a different learning method, however, that employs a type of feedback which has not been widely considered in discussions of the logical problem of language acquisition. In principle, a child could recover from an overly general hypothesis -- such superset cases are exactly where Gold's proof required negative evidence -- by observing that the predictions generated by the hypothesis are not borne out in the speech she hears. For example, a child might learn that all dative verbs do not alternate ('I donated the book to him/*I donated him the book', as opposed to 'I gave the book to him/I gave him the book') by observing that 'donate' never occurs in a double-object frame. Of course, this powerful mechanism must be constrained, as there are an infinite number of sentences which the child will never hear. This might be accomplished by embedding such a mechanism in a prediction framework: as the child listens to others speak, she predicts that certain elements will follow one another. When the predictions are incorrect, such as a prediction that 'donate' will be followed by the indirect object (based on an overly general rule resulting from the observation that give-type verbs can be followed by the indirect object), the hypothesized sequential pairing which gave rise to the incorrect prediction is decremented. Thus, the child learns by *listening* to utterances rather than by producing them, and generates her own negative evidence (or error signal) by comparing her predicted inputs with the actual input.

Simulation 1A: Explicit Negative Evidence Feedback

We begin with ENEF because it is based on a commonly held view of language acquisition, in which the learner entertains discrete hypotheses about which grammar relations are allowed and which are disallowed. Within this perspective, a method by which the learner rules out hypotheses (that is, those hypotheses not already ruled out by innate constraints) is by producing a hypothetically grammatical utterance and receiving corrective feedback indicating that it is ungrammatical. We wanted to make the implementation of ENEF unrealistically strong in order to determine the near limit in how fast it can induce a grammar with 95% accuracy. Therefore, it is given a generous amount (compared to what children typically receive) of *noise-free* explicit negative evidence, and the model requires *only one instance* of corrective feedback in order to rule out any particular hypothesized sequential pairing.

These first simulations are based on 6X6 grammars, similar to the one in Figure 1. The model starts out assuming that every sequential pairing between every element of the grammar is allowed (1's in all 36 cells). The model randomly produces a sequential pairing that it hypothesizes is grammatical. If this output is, in fact, ungrammatical according to the target grammar, there is a 20% chance that the model will receive corrective feedback, turning the strength of that sequential pairing from one to

A.

	a	*b*	*c*	*d*	*e*	*f*
a	0	0	1	1	1	0
b	0	1	0	0	0	1
c	0	0	0	1	1	0
d	1	1	0	0	1	0
e	1	1	1	0	0	1
f	0	1	1	1	0	0

B.

	a	*b*	*c*	*d*	*e*	*f*
a	0	1	0	0	0	0
b	0	0	1	0	0	0
c	0	0	0	1	0	0
d	0	0	0	0	1	0
e	0	0	0	0	0	1
f	1	0	0	0	0	0

Figure 2. The richest and the sparsest grammars on which the model was trained, respectively. The grammar in panel A has 19 ungrammatical pairings and the one in panel B has 30.

zero, thus discretely ruling out that specific hypothesis. If the output happens to be grammatical, no learning takes place. This process is repeated until the model's internalization of the grammar is at least 95% accurate. In a 6X6 grammar, this means the model must learn all but one of the zeroes in the target grammar. The number of utterances required to achieve this criterion of accuracy is referred to as the "learning time". (It is important to note that the last few zeroes take the longest to learn because, by that time, there is a slim chance that the model's random output of hypothesized sequential pairings will come upon an ungrammatical one. And this slim chance is *multiplied* by the .2 probability of negative evidence!)

The model was trained on twelve different grammars (see Figure 2). 100 simulations were conducted per grammar in order to compute a reliable average learning time. With the richest grammar, one containing many grammatical sequential pairings (Figure 2A), learning was slowest. The average learning time for this grammar was 315 timesteps (or produced utterances). In contrast, the absolute sparsest grammar (Figure 2B) took an average of only 242 timesteps. The remaining grammars of intermediate densities had average learning times between those extremes, gradually decreasing as density decreased.

As mentioned above, the probability of ruling out an hypothesized sequential pairing on any one timestep is equal to the probability of receiving negative evidence when an ungrammatical pairing is produced multiplied by the probability of producing an ungrammatical pairing. Thus, at the onset of learning the grammar in Figure 2B, there is a relatively high probability of ruling out an hypothesis that is ungrammatical: P(c) = .2(30/36), where P(c) is the probability of a correction, and 30/36 is the ratio of ungrammatical pairings over the total number of pairings. If a correction does not take place (either because the utterance happened to be grammatical, or the teacher failed to provide negative evidence), then the probability of a correction increases on the next timestep, according to probability summation over time (Watson, 1979). Using probability summation over time[1], the ENEF's mean learning time for the twelve grammars is approximated; r^2=.90. See also the open and filled triangles of Figure 4.

Simulation 1B: Incremental Distributed Prediction Feedback

IDPF learns in a much more realistic fashion than ENEF. It requires no explicit negative evidence regarding produced utterances. Instead, it *listens* to random utterances comprised of a randomly generated element followed by a randomly selected grammatical subsequent element. Incrementally, it makes distributed predictions about what elements should follow the initial input element. It typically requires about 10-20 exposures to an utterance in order to reduce the prediction strength of alternative ungrammatical pairings from that first element to less than .05.

In the next set of simulations, we trained the IDPF model on the same twelve grammars. IDPF is probabilistic, in that it starts out with prediction strengths that sum to 1 across any given row of the matrix (like a typical Markov chain). Beginning with a blank slate, the learner has .167

1

$$(1) \qquad \text{cumP(C)} = 1-\prod_{\tau}(1-\eta\frac{\Omega}{\psi})$$

where η is the probability of receiving negative evidence (should the learner produce an ungrammatical utterance), Ω is the number of *ungrammatical* pairings still hypothesized by the learner to be grammatical, ψ is the *total* number of sequential pairings still hypothesized by the learner to be grammatical, and τ is the timestep or utterance number. The cumulative probability of a correction increases as τ increases. When the cumulative probability of a correction exceeds .632 (that is, 1-1/e), the signal has, on average, occurred (cf. Watson, 1979). Thus, we iterated this equation, decrementing Ω and ψ by 1 each time cumP(C) exceeded .632 (as that meant there was one less ungrammatical pairing left to learn, and one less pairing hypothesized by the learner), and starting again at τ=1. This was repeated until Ω=1,which meant that all but one incorrect hypothesis had been ruled out; >95% accuracy. The sum of the Ω-1 values of τ corresponds to how many timesteps probability summation over time predicts the ENEF model will require to learn the grammar.

	a	b	c	d	e	f
a	.020	.324	.020	.596	.020	.020
b	.519	.004	.004	.004	.004	.465
c	.010	.010	.010	.010	.950	.010
d	.024	.024	.500	.404	.024	.024
e	.308	.255	.005	.005	.005	.422
f	.010	.481	.479	.010	.010	.010

Figure 3. IDPF's 95% accurate probabilistic representation of the grammar in Figure 1. (achieved in 107 timesteps.)

prediction strength in every cell of the 6X6 matrix. And reaching 95% accuracy in internalizing the grammar means that the summed error in each row averages .05. Figure 3 shows the learner's prediction strengths after reaching 95% accuracy on the grammar shown in Figure 1.

The model modifies its prediction strengths by comparing its weighted predictions of the input at time t+1 with the actual (discrete) input at time t+1. The one weighted prediction that was correct is increased by 15% of the difference between 1 and its current value, and the five predictions that were incorrect are decreased by 15% of their current values. This learning procedure is a version of the generalized delta rule (Rumelhart, Hinton & Williams, 1986), or "back-propagation", with a .15 learning rate.[2]

As before, 100 training simulations were run on each grammar. The IDPF learner reached criterion accuracy well before ENEF on all twelve grammars. [In fact, in order to achieve performance equivalent to IDPF's, ENEF requires an extremely unrealistic 45-65% explicit negative evidence.] An important observation here is that ENEF learns cell-by-cell, that is, it has a chance at each utterance of learning completely about one particular sequential pairing. In contrast, IDPF (because its predictions are distributed across the entire matrix row) learns gradually row-by-row, thus every utterance is a learning experience with respect to *all* of the elements that could follow the element at time t.

Learning time results for IDPF are shown in Figure 4, combined with the results of ENEF. A lower limit on the learning time for IDPF is easily computed by determining the minimum number of exposures to an initial element required to bring the sum of that row's ungrammatical cells to less than or equal to .05.[3] By adding to this lower limit the number of rows in the matrix minus 1, we can closely approximate IDPF's mean learning time (learning time = $(\Sigma\ \tau_r) + N-1$); $r^2=.98$. See the open and filled circles in Figure 4.

[2] In fact, the IDPF model is equivalent to a perceptron (with no hidden units) using the delta rule, but with the "desired output" signal being provided by the subsequent input, rather than by an explicit teacher.

Sim 2A: ENEF with 10X10 Grammars

To test how well these models scale up to larger grammars, this next set of simulations trained ENEF on a larger size of the same type of grammar (see Figure 5). As before, 100 simulations were run on each grammar to compute mean learning times. ENEF learned in the same fashion described in Simulation 1A, but this time all but *five* ungrammatical sequential pairings needed to be learned; in a 10X10 grammar, this was sufficient to produce 95% accuracy.

For the richest grammar tested (Figure 5A), ENEF was slowest, producing an average of 777 pairings in order to achieve 95% accuracy. As the grammar became sparser, ENEF learned faster, with the sparsest grammar (Figure 5B) being learned in 679 timesteps. As in Simulation 1A, probability summation over time approximated the mean learning times for the twelve grammars; $r^2=.945$ (Figure 6).

Sim 2B: IDPF with 10X10 Grammars

IDPF was trained on the same twelve grammars, in the same fashion as in Simulation 1B. The 95% accurate internalization of the 10X10 grammar, at the end of learning, was analogous to the 6X6 version shown in Figure 3. IDPF scaled up to this larger grammar much more gracefully than did ENEF. IDPF learned the richest grammar in an average of 159 timesteps, and the sparsest grammar in 181 timesteps. As before, Equation (2) (in footnote 3) closely approximated the mean learning times for the twelve grammars; $r^2=.99$ (see Figure 6).

General Discussion

A common assumption in the field of language acquisition is that if innate constraints did not encode certain grammatical relationships, then the standard negative evidence model, with its extremely slow learning, is all the learner could resort to. The results of these simulations suggest that an alternative, and cognitively plausible, method of learning (IDPF) is a great deal faster than the standard model. IDPF's distributed prediction of temporal associations is analogous in mechanism to the priming of semantic or syntactic associations. However, it remains an empirical question whether children learning their first language actually use this kind of passive incremental prediction of temporal associations between inputs.

[3]

$$(2) \qquad \frac{\Omega_r}{N} \prod_{\tau}(1-\eta) \leq .05$$

where Ω_r is the number of ungrammatical cells in row r, N is the number of elements in the language, τ is the timestep, and η is the learning rate [playing a similar role here as did the negative evidence term in Equation 1 for probability summation over time in ENEF (see footnote 1)]. By solving for τ for each row of the matrix, and summing those N values of τ, we get a lower limit of learning time for IDPF (lower limit = $\Sigma\tau_r$).

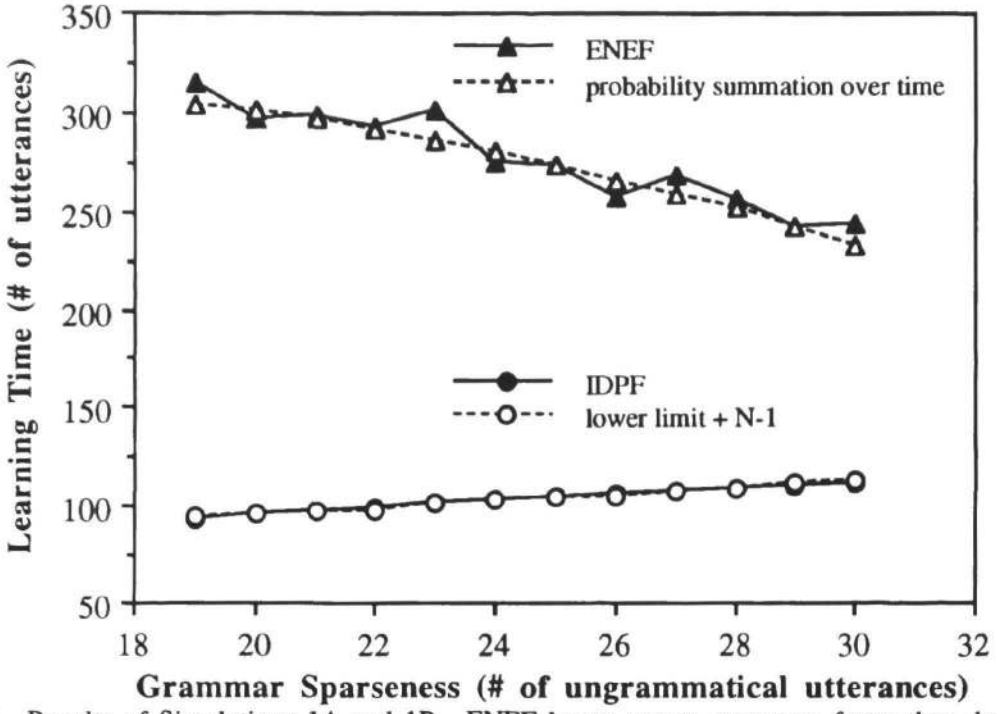

Figure 4. Results of Simulations 1A and 1B. ENEF learns sparse grammars faster than dense ones, and is approximated by probability summation over time. IDPF learns much faster than ENEF, with a modest increase in learning time as a function of the ratio of ungrammatical cells to total cells in the grammar.

A.

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*
a	1	1	1	1	0	1	1	0	0	1
b	0	0	0	0	1	0	1	1	0	1
c	1	0	0	1	0	0	0	1	0	0
d	0	0	1	0	0	1	0	1	1	1
e	0	0	1	1	1	1	1	1	1	0
f	0	1	0	0	1	0	1	0	0	0
g	0	1	0	0	0	1	0	0	1	1
h	1	0	1	1	1	0	0	1	0	0
i	0	1	0	0	1	1	0	0	0	1
j	1	0	1	0	0	0	1	0	1	0

B.

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*
a	0	1	0	1	0	0	0	1	0	0
b	1	0	0	0	0	0	0	0	1	0
c	0	1	0	0	1	0	0	0	0	0
d	0	0	0	1	0	0	0	0	0	0
e	1	0	0	0	0	0	0	0	0	1
f	0	0	0	1	0	0	1	0	0	0
g	0	1	0	0	0	0	0	1	0	0
h	0	0	0	0	1	0	1	0	1	0
i	1	0	1	0	1	0	0	0	0	0
j	0	0	0	1	0	1	1	0	1	0

Figure 5. Simulations 2A and 2B: The richest and sparsest grammars on which the model was trained, respectively. The grammar in panel A has 54 ungrammatical pairings and the one in panel B has 76.

With Incremental Distributed Prediction Feedback, the learner can take advantage of the conspicuous *absence* of certain grammatical relationships in the input. Moreover, if the input contains graded statistical biases for some sequential elements over others, IDPF's probabilistic encoding will cause it to reflect those graded preferences, just as adult comprehenders do (e.g., Juliano & Tanenhaus, 1993; Saffran, Newport & Aslin, submitted). The standard negative evidence model cannot produce such graded preferences, as its coding is discrete.

It is certainly possible that evolution has caused our DNA to encode certain constraints devoted to language learning (Batali, 1994; Pinker & Bloom, 1990). However, our results cast some doubt on whether the apparent lack of negative evidence can be used as a valid motivation for such a claim. In fact, what IDPF provides is a mechanism by which the child can *produce her own negative evidence* by comparing predicted (or primed) input with actual input.

Certainly, these simulations are a simplified test case, and do not apply to natural grammars, which contain multiple-contingency relationships of far greater complexity than pairwise sequences. Future work on this issue will require comparing the two learning styles ("overt testing" vs. "passive predicting") in their ability to learn more complex, natural grammars.

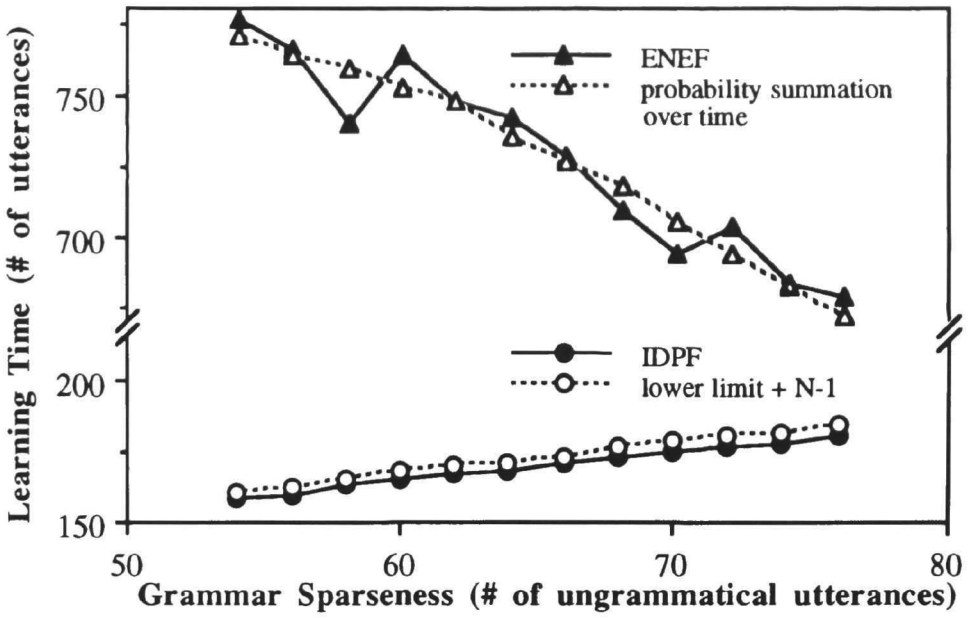

Figure 6. Results of Simulations 2A and 2B. The gap between ENEF and IDPF is even greater.

Acknowledgments: This research was supported by NSF Graduate Research Fellowships to both authors. We are grateful to Gail Mauner, Toby Mintz, Elissa Newport, Whitney Tabor and Mike Tanenhaus for helpful discussions of the work.

References

Batali, J. (1994). Artificial evolution of syntactic aptitude. *Proceedings of the 16th Annual Conference of the Cognitive Science Society* .

Bohannon, J. & Stanowicz, L. (1988). The issue of negative evidence: Adult responses to children's language errors. *Developmental Psychology*, 24, 684-689.

Bohannon, J., MacWhinney, B. & Snow, C. (1990). No negative evidence revisited: Beyond learnability or who has to prove what to whom. *Developmental Psychology*, 26, 221-226.

Brown, R. & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. Hayes (Ed.), *Cognition and the development of language*. New York: Wiley.

Chomsky, N. (1972). *Language and mind*. New York: Harcourt Brace Jovanovich.

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

Gold, E. (1967). Language identification in the limit. *Information and Control*, 10, 447-474.

Gordon, P. (1990). Learnability and feedback. *Developmental Psychology*, 26, 217-220.

Hirsh-Pasek, K., Treiman, R. & Schneiderman, M. (1984). Brown and Hanlon revisited: Mother's sensitivity to ungrammatical forms. *Journal of Child Language*, 11, 81-88.

Juliano, C. & Tanenhaus, M. (1993). Contingent frequency effects in syntactic ambiguity resolution. *Proceedings of the 15th Conference of the Cognitive Science Society*.

Juliano, C. & Tanenhaus, M. (1995). A constraint-based lexicalist account of the subject/object attachment preference. *Journal of Psycholinguistic Research*, 23, 459-471.

Marcus, G. (1993). Negative evidence in language acquisition. *Cognition*, 46, 53-85.

McNeill, D. (1970). *The acquisition of language*. New York: Harper and Row.

Morgan, J. & Travis, L. (1989). Limits on negative information in language input *Journal of Child Language*, 16, 531-552.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pinker, S. & Bloom, P. (1990). Natural language and natural selection. *Brain and Behavioral Sciences*, 13, 707-784.

Rumelhart, D., Hinton, G. & Williams, R. (1986). Learning internal representations by error propagation. In D. Rumelhart, J. McClelland, and the PDP Research Group (Eds.), *Parallel Distributed Processing. Volume 1: Foundations*. Cambridge, MA: MIT Press.

Saffran, J., Newport, E. & Aslin, R. (submitted). Word segmentation: The role of distributional cues.

Schütze, H. (1994). A connectionist model of verb subcategorization. *Proceedings of the 16th Annual Conference of the Cognitive Science Society* .

St. John, M. & McClelland, J. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46, 217-257.

Watson, A. B. (1979). Probability summation over time. *Vision Research*, 19, 515-522.