Cognitive Science 29 (2005) 1007–1028

Uncovering the Richness of the Stimulus: Structure Dependence and Indirect Statistical Evidence

Florencia Reali, Morten H. Christiansen

Department of Psychology, Cornell University

Received 12 August 2004; received in revised form 9 March 2005; accepted 22 March 2005

Abstract

The *poverty of stimulus argument* is one of the most controversial arguments in the study of language acquisition. Here we follow previous approaches challenging the assumption of impoverished primary linguistic data, focusing on the specific problem of auxiliary (AUX) fronting in complex polar interrogatives. We develop a series of corpus analyses of child-directed speech showing that there is indirect statistical information useful for correct auxiliary fronting in polar interrogatives and that such information is sufficient for distinguishing between grammatical and ungrammatical generalizations, even in the absence of direct evidence. We further show that there are simple learning devices, such as neural networks, capable of exploiting such statistical cues, producing a bias toward correct AUX questions when compared to their ungrammatical counterparts. The results suggest that the basic assumptions of the poverty of stimulus argument may need to be reappraised.

Keywords: Poverty of stimulus; Distributional information; Corpus analysis; Neural networks; Language acquisition

1. Introduction

How do children learn aspects of their language for which there appears to be no evidence in the input? This question lies at the heart of the most enduring and controversial debates in cognitive science. Ever since Chomsky (1965), it has been argued that the information in the linguistic environment is too impoverished for a human learner to attain adult competence in language without the aide of innate linguistic knowledge. Although this *poverty of the stimulus argument* (Chomsky, 1980a, 1980b; Crain & Pietroski, 2001) has guided most research in linguistics, it has proved to be much more contentious within the broader context of cognitive science.

The poverty of stimulus argument rests on specific assumptions about the nature of the input to the child, the properties of computational learning mechanisms, and the learning

Requests for reprints should be sent to Florencia Reali, Department of Psychology, Cornell University, Ithaca, NY 14853. E-mail: fr34@cornell.edu

abilities of young infants. A growing bulk of research in cognitive science has begun to call each of these three assumptions into question. Thus, whereas the traditional nativist perspective suggests that statistical information may be of little use for syntax acquisition (e.g., Chomsky, 1957), recent research indicates that distributional regularities may provide an important source of information for bootstrapping syntax (e.g., Mintz, 2002; Redington, Chater, & Finch, 1998)—especially when integrated with prosodic or phonological information (e.g., Christiansen & Dale, 2001; Morgan, Meier & Newport, 1987; Reali, Christiansen, & Monaghan, 2003). And although the traditional approach only tends to consider learning in highly simplified forms, such as "move the first occurrence of *X* to *Y*," progress in statistical natural language processing and connectionist modeling has revealed much more complex learning abilities of potential relevance for language acquisition (e.g., Christiansen & Chater, 1999; Elman, 1993; Lewis & Elman, 2001; Manning & Schütze, 1999). Finally, little attention has traditionally been paid to what young infants may be able to learn, perhaps because it has implicitly been assumed that such learning would be negligible. However, recent research has demonstrated that young infants are quite competent statistical learners (e.g., Gómez, 2002; Saffran, Aslin, & Newport, 1996; Saffran & Wilson, 2003; for reviews, see Gómez & Gerken, 2000; Saffran, 2003).

These research developments suggest the need for a reappraisal of the poverty of stimulus argument, centered on whether they can answer the question of how a child may be able to learn aspects of linguistic structure. In this article, we approach this question in the context of structure dependence in language acquisition, specifically in relation to auxiliary fronting in polar interrogatives. We first outline the poverty of stimulus debate, describing how it has played out with respect to forming grammatical questions with auxiliary fronting. Second, we conduct a corpus analysis to show that there is sufficiently rich statistical information available in child-directed speech for differentiating between correct and incorrect auxiliary (AUX) questions—even in the absence of any such constructions in the corpus. We additionally demonstrate how the same approach is consistent with results from studies of auxiliary fronting in children ages 3.0 to 5.0 years (Crain & Nakayama, 1987). Finally, we address the issue of whether simple learning devices are capable of utilizing such information. We therefore conduct a set of connectionist simulations to illustrate that neural networks are capable of using statistical information to distinguish between correct and incorrect AUX questions. Further analysis of the networks' patterns of prediction provides insights into the nature of statistical learning that are consistent with production data from children. Finally, we consider our results in the broader context of theories of language acquisition.

1.1. Poverty of stimulus argument and auxiliary fronting

The poverty of stimulus argument suggests that learning a language requires arriving at the correct generalization of grammatical structure given insufficient data. The logic of the argument is powerful: If the premises are granted, the conclusion seems airtight. If the data in the primary linguistic input are too impoverished to allow correct generalization, then convergence to adult grammatical competence requires a more endogenous, biological explanation: innateness of linguistic knowledge (e.g., Boeckx & Hornstein, 2004; Chomsky, 1980a, 1980b; Crain & Pietroski, 2001; Fodor & Crowther, 2002; Hornstein & Lightfoot, 1981; Laurence &

Margolis, 2001; Yang, 2002). Thus, the following question arises: How good are the premises? We suggest that the weakness of the argument lies in the difficulty of assessing the input, and in the imprecise and intuitive definition of "insufficient data."

One of the often-used examples to support the poverty of stimulus argument concerns auxiliary fronting in polar interrogatives (e.g., Boeckx & Hornstein, 2004; Chomsky, 1980a; Crain & Nakayama, 1987; Crain & Pietroski, 2001; Legate & Yang, 2002). In most generative grammar frameworks, declaratives are turned into questions by fronting the correct auxiliary. For example, in the declarative form *The man who is hungry is ordering dinner,* it is correct to front the main clause auxiliary as in (1a), but fronting the subordinate clause auxiliary produces an ungrammatical sentence as in (1b) (Chomsky, 1965).

(1a) *Is the man who is hungry ordering dinner?*
(1b) **Is the man who hungry is ordering dinner?*

It has been suggested that children can generate two types of rules: a structure-independent rule where the first *is* is moved, or the correct structure-dependent rule, where only the movement of the *is* from the main clause is allowed (Chomsky, 1980a). Crucially, children do not appear to go through a period when they erroneously move the first *is* to the front of the sentence (Crain & Nakayama, 1987). It has moreover been asserted that a person might go through much of his or her life without ever having been exposed to the relevant evidence for inferring correct auxiliary fronting (e.g., Chomsky, 1980a; Crain & Pietroski, 2001; Legate & Yang, 2002).

The purported absence of evidence in the primary linguistic input regarding auxiliary fronting is not without debate. Intuitively, as suggested by Lewis and Elman (2001), it is perhaps unlikely that a child would reach kindergarten without being exposed to sentences such as (2a) to (2c).

(2a) *Is the boy who was playing with you still there?*
(2b) *Will those who are hungry raise their hand?*
(2c) *Where is the little girl full of smiles?*

These examples have an auxiliary verb within the subject noun phrase, and thus the auxiliary that appears initially would not be the first auxiliary in the declarative, providing evidence for correct auxiliary fronting. Pullum and Scholz (2002) explored the presence of relevant examples for auxiliary fronting including *wh* questions similar to (2b) and (2c) in the *Wall Street Journal* corpus and found that at least five crucial examples occur in the first 500 interrogatives in the *Wall Street Journal.* They also analyzed two corpora from the CHILDES database (MacWhinney, 2000), reporting two relevant *wh*-question examples in a transcript from TV programs aimed at children and three additional examples in a corpus of utterances addressed to a girl when she was between 1 year 11 months and 3 years 3 months of age. These results suggest that the assumption of complete absence of evidence for correct auxiliary fronting is overstated (see also Cowie, 1998; Lewis & Elman, 2001). However, in defense of the poverty of stimulus argument, it has been argued that the positive evidence that young children could encounter in the primary linguistic data might not be considered sufficient to support data-driven learning theories. Indeed, more comprehensive studies of CHILDES corpora show that even though interrogatives constitute a large percentage of the corpus, relevant examples

of auxiliary fronting in polar interrogatives represent less than 1% of them (Legate & Yang, 2002; but see Scholz & Pullum, 2002, for discussion). However, it is worth reappraising the notion of relevant evidence that was entailed in this literature: Only *explicit* examples of particular grammatical constructions are being considered as evidence. We want to argue that this notion of "relevant evidence" is too narrow. Our claim is that, although direct examples of correct auxiliary fronting in complex polar interrogatives may be too infrequent to be helpful in acquisition—as suggested by Legate & Yang (2002)—other more *indirect sources* of statistical information may provide an additional cue for making the appropriate grammatical generalizations.

Recent connectionist simulations provide preliminary data in this regard. Lewis and Elman (2001) trained simple recurrent networks (SRN; Elman, 1990) on data from an artificial grammar that generated questions of the form "AUX NP ADJ?" and sequences of the form "A_i NP B_i" (where A_i and B_i represent a variety of different material) but no relevant examples of polar interrogatives. The SRNs were better at making predictions for multiclause questions involving correct auxiliary fronting compared to those involving incorrect auxiliary fronting.

However, the SRNs in the Lewis and Elman (2001) simulation studies were exposed to an artificial grammar without the complexity and noisiness that characterize actual child-directed speech. The question thus remains whether the indirect statistical regularities in an actual corpus of child-directed speech are strong enough to support grammatical generalizations over incorrect ones. Next, in our first experiment, we conduct a corpus analysis to address this question.

2. Experiment 1: Measuring indirect statistical information

We trained simple statistical models based on pairs (bigrams) and triples (trigrams) of words drawn from the Bernstein-Ratner (1984) corpus of child-directed speech. After training, the models were tested on sentences that consisted of correct polar interrogatives (e.g., *Is the man who is hungry ordering dinner?*) and incorrect ones (e.g., *Is the man who hungry is ordering dinner?*)—neither of which was present in the training corpus. We reasoned that if indirect statistical information—in the form of co-occurrences of pairs–triples of words—provides a possible cue for generalizing correctly to the grammatical AUX questions, then we should find a difference in the likelihood of these two alternative hypotheses.

Bigram–trigram models are simple statistical models that use the previous one or two words to predict the next one. Given a string of words, or a sentence, the associated *cross-entropy* for that string of words is a simple measure of probability according to the bigram–trigram model trained on a particular corpus (from Chen & Goodman, 1999). Cross-entropy is used for comparing the *likelihood* of test sentences in computational linguistics (Jurafsky & Martin, 2000), and it is inversely correlated to co-occurrence of the words in the sentences found in the training corpus. Thus, given two competing sentences, we can compare the probability of each of them as indicated by their associated cross-entropy, within the context of a particular corpus. Specifically, we can contrast the two alternative generalizations of AUX questions by compar-

ing the cross-entropy associated with grammatical (e.g., *Is the man who is in the corner smoking?*) and ungrammatical forms (e.g., *Is the man who in the corner is smoking?*). This will allow us to determine whether there may be sufficient indirect statistical information available in actual child-directed speech to decide between these two forms. More important, the Bernstein-Ratner (1984) corpus contains no explicit examples of auxiliary fronting in polar interrogatives.

The intuitive idea is to first break down a sentence into word *chunks*—that is, bigrams and trigrams—and then determine their frequency of occurrence in the corpus. The probability of a sentence can then be calculated as a product of its component word-chunk frequencies. The higher the frequencies are of the component word chunks, the more probable the sentence will be. For example, in Sentences (1a) and (1b), the number of occurrences of the component word chunks in each sentence is first counted across the corpus (i.e., *Is the, the man, man who* … and so on, according to the bigram model, and *Is the man, the man who, man who is* … and so on, according to the trigram model). The overall probability of each sentence is then computed, and we can thus directly compare the grammatical (1a) and ungrammatical (1b) and choose the most probable of the two.

2.1. Method

2.1.1. Models

For the corpus analysis, we used bigram and trigram models (see, e.g., Jurafsky & Martin, 2000) to measure how frequently pairs or triples of adjacent words occur in a corpus. Based on the probability of its fragments, the probability of a sentence, $p(s)$, was expressed as the product of the probabilities of the words (w_i) that compose the sentence, with each word probability conditional to the last $n - 1$ words. Then, if s = $w_1 \ldots w_k$ we have $p(s)$ as in Equation 1:

$$P(s) = \prod_i P\left(w_i \mid w_{i\text{-}n+1}^{i-1}\right) \tag{1}$$

As defined in Equation 2, we used *maximum likelihood* to estimate $p(w_i|w_{i-1})$ (here considering the bigram model):

$$P_{ML}(w_i|w_{i\text{-}1}) = P(w_{i\text{-}1}w_i)\,/P(w_{i\text{-}1}) = (c(w_{i\text{-}1}w_i)/Ns)/(c(w_{i\text{-}1})/Ns) \tag{2}$$

where *N*s denote the total number of tokens and $c(\alpha)$ is the number of times the string α occurs in the corpus. In this context, maximum likelihood provides a measure of how often fragments—triples of words in the case of trigrams, and pairs of words in the case of bigrams—occur in a corpus.

The Bernstein-Ratner (1984) corpus is small, thus, sometimes two or three words never occur together. We therefore used the *interpolation smoothing technique* defined in Chen and Goodman (1999). This method ensures that those pairs or triples of words in test sentences that never co-occur in the corpus receive a positive probability based on the probability of pairs of words or isolated words or both. The smoothing technique consists of the interpolation of the

bigram model with the unigram model, and the trigram model with the bigram model. Thus, if the probability of a word (w_i) (or unigram model) is defined as in Equation 3:

$$P_{ML}(w_1) = c(w_i)/Ns \tag{3}$$

for the bigram model we calculate the interpolated probability in Equation 4:

$$P_{\text{interp}}(w_i|w_{i\text{-}1}) = \lambda P_{ML}(w_i|w_{i\text{-}1}) + (1\text{-}\lambda)P_{ML}(w_i) \tag{4}$$

Accordingly for trigram models, the interpolated probability is computed as in Equation 5:

$$P_{\text{interp}}(w_i|w_{i\text{-}1}w_{i\text{-}2}) = \lambda P_{ML}(w_i|w_{i\text{-}1}w_{i\text{-}2}) + (1\text{-}\lambda)(\lambda P_{ML}(w_i|w_{i\text{-}1}) + (1\text{-}\lambda)P_{ML}(w_i)) \tag{5}$$

where λ is a value between 0 and 1 that determines the relative importance of each term in the equation. We used a standard $\lambda = 0.5$ so that all terms are equally weighted. We measured the likelihood of a given set of sentences using the measure of cross-entropy (Chen & Goodman, 1999). The cross-entropy of a set of sentences is defined in Equation 6:

$$1/N_T \, \Sigma_i \, \text{-}\log_2 P(s_i) \tag{6}$$

where N_T is the total number of words in the set of sentences and s_i is the ith sentence. Thus, the cross-entropy value is inversely correlated with the likelihood of the test sentences. For example, given a training corpus and two sentences "A" and "B," we can compare the cross-entropy of both sentences to estimate which one is more probable according to the statistical information inherent in the corpus.[1]

2.1.2. Materials

The Bernstein-Ratner (1984) corpus contains speech from nine mothers addressing their children and recorded over a 4- to 5-month period when children were between the ages 1 year 1 month and 1 year 9 months. This is a relatively small and very noisy corpus, mostly containing short sentences with simple grammatical structure. The sentences incorporate a number of different types of grammatical structures, showing the varied nature of the linguistic input to children. Utterances range from declarative sentences to *wh* questions to one-word utterances. Representative sample sentences are shown in (3a) to (3c).

(3a) *Oh you need some space.*
(3b) *Where is my apple?*
(3c) *Oh. That's it.*

2.1.3. Procedure

The bigram–trigram models were trained on 10,705 sentences from the Bernstein-Ratner (1984) corpus. The sentences contained 35,505 word tokens distributed over 1,856 word types. To compare the cross-entropy of grammatical and ungrammatical polar interrogatives, we created two novel sets of sentences. The first set contained grammatically correct multiclause polar interrogatives generated using an algorithm that randomly selected words from the corpus and created sentences according to syntactic and semantic constraints. The test sets only contained polar interrogatives of the form "*Is* NP (who/that) *is* A_i B_i?," where A_i denotes the re-

maining material in the relative clause (e.g., PNP, ADJP, ADVP, NP, VPG), and B_i the remaining material in the main clause (e.g., PNP, ADJP, ADVP, NP, VPG). Sentences (4a) to (4c) provide examples of the grammatical test sentences.

(4a) *Is the lady who is there eating?*
(4b) *Is the dog that is on the chair black?*
(4c) *Is the goose that is hungry smelling?*

A second set of matching ungrammatical sentences was created by moving the incorrect auxiliary to the front of the sentence. For example, for the grammatical sentence (4a) the corresponding ungrammatical sentence is *"*Is the lady who there is eating?"* We generated 100 of these sentence pairs and computed the mean cross-entropy per sentence across the 100 sentences in each set. Finally, we contrasted the likelihood of pairs of grammatical and ungrammatical sentences by comparing their cross-entropy and choosing the version with the lowest value.

2.2. Results

We found that the mean cross-entropy of grammatical sentences was lower than the mean cross-entropy of ungrammatical sentences. As paired *t*-test comparisons revealed, the cross-entropy difference between the two sentence types was highly significant, $t(99) = 15.03$, $p = .0001$ for the bigram model, and $t(99) = 11.74$, $p = .0001$ for the trigram model (see Table 1). These results indicate that grammatical sentences have a higher probability than ungrammatical ones.

The probability of a sentence is inversely correlated with its cross-entropy value (see Equation 6). Fig. 1 shows the comparison of mean probability of grammatical and ungrammatical sentences. We found that the mean probability of grammatical polar interrogatives was almost twice as high as the mean probability of ungrammatical polar interrogatives, according to the bigram model, and it was more than twice as high according to the trigram model.

To compare each grammatical–ungrammatical pair of sentences, we defined the following criterion: When deciding between each grammatical versus ungrammatical polar interrogative example, choose the one that has lower cross-entropy (the most probable one). A sentence is defined as *correctly classified* if the chosen form is grammatical. Using that criterion, we found that the percentage of correctly classified sentences using the bigram and trigram models is 96%. Fig. 2 shows the performance of the models according to the defined classification criterion.

Table 1
Comparison of mean cross-entropy in Experiment 1

	Mean Cross-Entropy			
	Grammatical	Ungrammatical	Mean Difference	*t*(99)
Bigram	22.42	23.24	0.82	15.03*
Trigram	21.29	22.50	1.21	11.75*

*$p < .0001$.

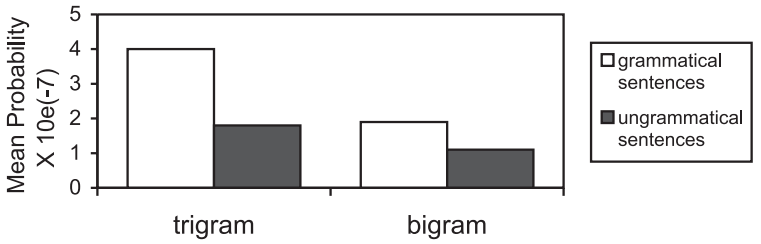

Fig. 1. Mean probability of grammatical sentences versus ungrammatical sentences estimated by the trigram (left) and bigram (right) models trained on the corpus.

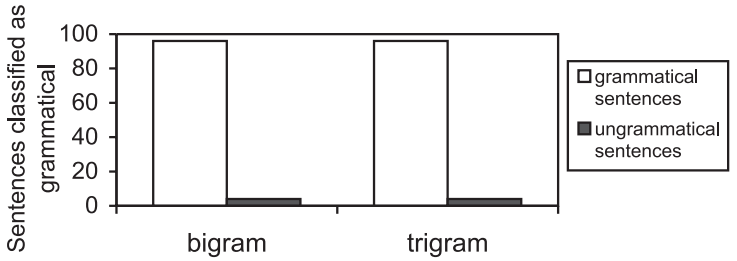

Fig. 2. Number of sentences classified correctly (white bars) and incorrectly as grammatical (black bars).

Of the 100 test sentences, the bigram and trigram model only misclassified the four sentences in (5a) to (5d).

(5a) *Is the jacket that is on the chair lovely?*
(5b) *Is the lady who is here drinking?*
(5c) *Is the alligator that is standing there red?*
(5d) *Is the dog that is on the chair black?*

These results indicate that it is possible to distinguish between grammatical and ungrammatical AUX questions based on the indirect statistical information in a noisy child-directed speech corpus containing no explicit examples of such constructions.

These results are based on the combined input from nine different mothers, but the question remains whether similar results can be obtained if the analyses focus on the input to individual children.[2] To explore this question, we repeated our analyses by conducting individual analyses for each child.

To compute the bigram and trigram analyses, it is necessary that the words in the sentences are included in the corpus. Thus, for each individual corpus analysis we only used test sentences that were supported by the relevant corpus in the sense that all words were present therein. Due to variability in word use by different mothers, the number of supported test sentences therefore varies across each of the nine individual child-directed speech corpora. Table 2 shows the number of test sentences supported by each child corpus and how many of these were misclassified, along with the mean cross-entropy for the grammatical and ungrammatical test sentence, and whether the difference between these two was significant. In all cases, even

Table 2
Comparison of mean cross-entropy across individual-child corpora

	Sentences		Mean Cross-Entropy			
	Total	MC	G	UG	Mean Difference	*t*
Dale						
Bigram	43	7	22.23	22.68	0.49	5.99**
Trigram	43	7	21.32	21.93	0.75	5.31**
Marie						
Bigram	3	2	18.90	18.91	< 0.1	< 1
Trigram	3	1	17.72	18.20	0.75	1.06
Cindy						
Bigram	52	2	22.13	22.82	0.61	8.24**
Trigram	52	7	21.07	22.08	1.00	6.10**
Gail						
Bigram	38	6	22.06	22.47	0.45	5.59**
Trigram	38	6	21.14	21.72	0.58	4.91**
Anne						
Bigram	10	0	19.95	20.41	0.45	3.00*
Trigram	10	0	19.17	19.63	0.45	3.00*
Lena						
Bigram	11	3	20.74	20.74	< 0.001	< 1
Trigram	11	3	19.79	20.03	0.23	< 1
Amelia						
Bigram	1	0	15.25	15.62	0.37	—
Trigram	1	0	15.25	15.62	0.37	—
Kay						
Bigram	1	0	16.50	17.02	0.52	—
Trigram	1	0	16.50	17.02	0.52	—

Note. Total = total of sentences available for comparison in individual corpora; MC = misclassified; G = grammatical; UG = ungrammatical. The *Alice* corpus did not support any test sentences and could therefore not be analyzed.

*$p < .05$. **$p < .0001$.

when the analyses were conducted with the quite small individual child corpora, the models only misclassified a small fraction of the test sentences. Thus, these additional analyses indicate that the statistical pattern found in our first analyses is stable across input to different individual children. Of course, under natural circumstances the primary linguistic input to each child will surpass the size of our combined corpus by orders of magnitude, and this is likely to lead to even more robust indirect statistical information.

3. Experiment 2: Testing sentences produced by children

Although Experiment 1 shows that there is sufficient indirect statistical information available in child-directed speech to differentiate reliably between the grammatical and un-

grammatical AUX questions that we had generated, it could be argued that the real test for our approach is whether it works for actual sentences produced by children. We therefore tested our models on a small set of sentences elicited from children under experimental conditions.

Crain and Nakayama (1987) conducted an experiment designed to elicit complex AUX questions from children between 3 and 5 years of age. The participants were involved in a game in which they asked questions of Jabba the Hutt, a creature from *Star Wars*. During the task the experimenter gives an instruction to the child: *Ask Jabba if the boy who is watching Mickey Mouse is happy.* Children produced sentences such as (7a), but they never produced sentences such as (7b):

(7a) *Is the boy who is watching Mickey Mouse happy?*
(7b) *Is the boy who watching Mickey Mouse is happy?*

The authors concluded that the lack of structure-independent errors suggested that children entertain only structure-dependent hypotheses, supporting the existence of innate grammatical structure. Here we explore whether our model is capable of distinguishing between structure-dependent and structure-independent hypotheses in Crain and Nakayama's (1987) study, based purely on the statistical information of the Bernstein-Ratner (1984) corpus.

3.1. Method

3.1.1. Models

Same as in Experiment 1.

3.1.2. Materials

Six example pairs were derived from the declarative sentences used in Crain and Nakayama (1987)[3]:

(8a) *The ball that the girl is sitting on is big*
(8b) *The boy who is unhappy is watching Mickey Mouse*
(8c) *The boy who is watching Mickey Mouse is happy*
(8d) *The boy who is being kissed by his mother is happy*
(8e) *The boy who was holding the plate is crying*
(8f) *The dog that is sleeping is on the blue bench*

The grammatical and ungrammatical AUX questions were derived from the declaratives in (8a) to (8f). Thus, the sentence *Is the ball that the girl is sitting on big?* belonged to the grammatical test set, whereas the sentence **Is the ball that the girl sitting on is big?* belonged to the ungrammatical test set. Consequently, grammatical and ungrammatical test sets contained 6 sentences each.

3.1.3. Procedure

The bigram–trigram models were trained on the Bernstein-Ratner (1984) corpus, as in Experiment 1, and tested on the material derived from Crain and Nakayama (1987).

3.2. Results

Consistent with Experiment 1, we found that the mean cross-entropy of grammatical sentences was significantly lower than the mean cross-entropy of ungrammatical sentences both for bigram, $t(5) = 3.88$, $p = .011$, and trigram models, $t(5) = 2.97$, $p = .031$. Table 3 summarizes these results.

Using the classification criterion defined in Experiment 1, we found that all six sentences were correctly classified using the bigram model. When using the trigram model, we found that five out of six sentences were correctly classified, and only Sentence (8f) was misclassified.

As in Experiment 1, we repeated the analysis for each of the individual-child corpora. Only five out of the nine individual corpora contained the appropriate words to support a subset of the test sentences in 8. Table 4 shows the results of these analyses in terms of mean

Table 3
Comparison of mean cross-entropy in Experiment 2

	Mean Cross-Entropy			
	Grammatical	Ungrammatical	Mean difference	$t(5)$
Bigram	27.13	28.00	0.87	3.88*
Trigram	26.15	26.96	0.81	2.97*

*$p < .05$.

Table 4
Experiment 2: Comparison across individual-child corpora

	Sentences		Mean Cross-Entropy			
	Total	MC	G	UG	Mean difference	$t(5)$
Dale						
Bigram	3	0	26.31	26.95	0.63	1.21
Trigram	3	0	25.68	26.31	0.63	1.21
Marie						
Bigram	2	0	25.50	25.50	—	—
Trigram	2	0	25.50	25.50	—	—
Cindy						
Bigram	5	0	27.69	28.56	0.87	3.31*
Trigram	5	0	27.15	28.03	0.87	3.31*
Gail						
Bigram	3	0	26.00	26.64	0.64	1.22
Trigram	3	0	25.35	25.99	0.64	1.22
Lena						
Bigram	3	0	25.25	25.92	0.67	< 1
Trigram	3	0	24.88	25.55	0.23	< 1

Note. Total = total of sentences available for comparison in individual corpora; MC = misclassified; G = grammatical; UG = ungrammatical. The *Alice, Anne, Amelia,* and *Kay* corpora did not support any test sentences and could therefore not be analyzed.

*$p < .05$.

cross-entropy and classification. For each individual-child analysis, every test sentence was classified correctly (i.e., the grammatical sentence had a higher probability than the ungrammatical version).

4. Experiment 3: Learning to produce correct sentences

Previous simulations by Lewis and Elman (2001) showed that SRNs trained on data from an artificial grammar were better at predicting the correct auxiliary fronting in AUX questions. Here, we explore whether the results shown using artificial-language models will scale up to deal with the full complexity and the general disorderliness of speech directed at young children. It is not clear whether a simple learning device may be able to exploit the statistical information established in Experiments 1 and 2 to develop an appropriate bias toward the grammatical forms. To investigate this question, we took a previously developed SRN model of language acquisition (Reali et al., 2003), which had also been trained on the same corpus, and tested its ability to deal with AUX questions.

SRNs have been used widely as a psychological model of human learning (e.g., Botvinick & Plaut, 2004; Christiansen & Chater, 1999; Cleeremans, 1993; Elman, 1990). This type of network is well suited for our simulations because it has previously been successfully applied to the modeling of language learning (e.g., Elman, 1990, 1993) and has been shown to be sensitive to bigram–trigram information (Christiansen & Chater, 1999; Reali et al., 2003). More important, 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).

4.1. Method

4.1.1. Networks

We used the same 10 SRNs that Reali et al. (2003) had trained to predict the next lexical category given this one. An SRN is essentially a standard feed-forward neural network equipped with an extra layer of so-called context units. At a particular time step, *t,* an input pattern is propagated through the hidden unit layer to the output layer. At the next time step, $t + 1$, the activation of the hidden unit layer at time t is copied back to the context layer and paired with this input. This means that this state of the hidden units can influence the processing of subsequent inputs, providing a limited ability to deal with integrated sequences of input presented successively.

The initial weights of the networks were randomized within the interval (–0.1; 0.1). A different random seed was used for each network. Learning rate was set to 0.1, and momentum to 0.7. Each input to the network contained a localist representation of the lexical category of the incoming word. With a total of 14 different lexical categories and a pause marking boundaries between utterances, the network had 15 input units. The network was trained to predict the lexical category of the next word, and thus the number of output units was 15. Each network had 30 hidden units and 30 context units. Fig. 3 provides an illustration of the network used in the

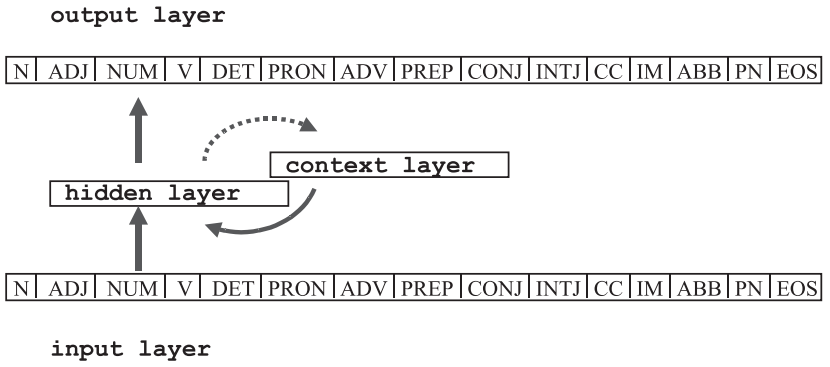

Fig. 3. Network configuration. The arrows indicate full connectivity between layers. Dashed lines indicate fixed connection weights (with a value of 1), and solid lines indicate learnable connection weights.

simulations. No changes were made to the original networks from Reali et al. (2003) and their parameters.[4]

4.1.2. Materials

We trained and tested the networks on the Bernstein-Ratner (1984) corpus similarly to the bigram–trigram models. Each word in the corpus was assigned one of the 14 following lexical categories from CELEX database (Baayen, Pipenbrock & Gulikers, 1995): nouns (N: 19.5%), verbs (including auxiliaries; V: 18.5%), adjectives (ADJ: 4%), numerals (NUM: < 0.1%), adverbs (ADV: 6.5%), determiners (DET: 6.5%), pronouns (PRON: 18.5%), prepositions (PREP: 5%), conjunctions (CONJ: 4%), interjections (INTJ: 7%), complex contractions (CC: 8%), abbreviations (ABB: < 0.1%), infinitive markers (IM: 1.2%), and proper names (PN: 1.2%). Each word in the corpus was replaced by a vector encoding the lexical category to which it belonged. We used the two sets of test sentences used in Experiment 1, containing grammatical and ungrammatical polar interrogatives, respectively. However, due to replacing each of the individual words with their respective lexical categories, some of the original test sentences from Experiment 1 ended up mapping onto the same string of lexical categories. For simplicity, we only considered unique strings, resulting in 30 sentences in each test set (grammatical and ungrammatical).

4.1.3. Procedure

The 10 SRNs (with different random weight initializations) from Reali et al. (2003) were trained on one pass through the Bernstein-Ratner (1984) corpus and tested on the AUX questions described previously. To compare network predictions for the ungrammatical versus the grammatical AUX questions, we measured the networks' mean square error (*MSE*) recorded during the presentation of each test sentence pair. The *MSE* is calculated by measuring the difference between the network's output activity for the next lexical category and the *target* output (i.e., N, ADJ), given the previous sentential context. Because the *MSE* is measured for each word in the sentence, we calculate the average *MSE* per sentence, by averaging the *MSE* value

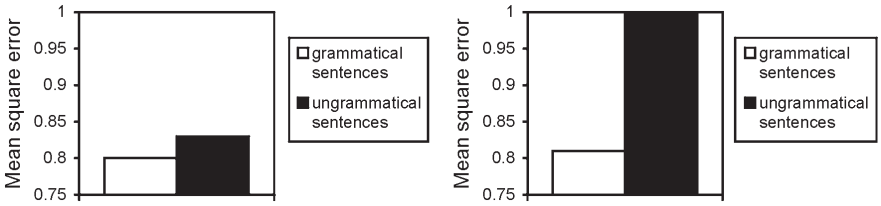

Fig. 4. (a) Left side: Average MSE across sentences in the grammatical set (white bar) and ungrammatical set (black bar). (b) Right side: Average MSE elicited after the sequence of lexical categories corresponding to *Is NP that/who …* in the grammatical set (white bar) and the ungrammatical set (black bar).

across all words. Then we can compare the average *MSE* elicited by grammatical versus ungrammatical sentences.

4.2. Results

We found that in all 10 simulations, the grammatical set of AUX questions produced a lower average *MSE* compared to the ungrammatical ones. As illustrated by Fig. 4a, when network predictions for the next lexical class were averaged across sentences, the *MSE* for the grammatical set was significantly lower than the *MSE* for the ungrammatical set (0.80 vs. 0.83; $t(9) = 11.93$, $p < .0001$). Because the MSE is averaged across all the words in a sentence, the small numerical difference between grammatical and ungrammatical sentences is partly explained by the fact that the two test sets form minimal pairs in that they are almost identical save for the position of the fronted *is.* Fig. 4b shows the difference in *MSE* for predicting the next lexical category at the crucial point where grammatical and ungrammatical test sentences diverge after the string of lexical categories corresponding to *Is NP who/that … ,* indicating a significantly lower error for the grammatical continuations compared to the ungrammatical ones (0.82 vs. 1.02; $t(9) = 13.08$, $p < .0001$).

To further elucidate why the networks produced more accurate predictions for the grammatical test sentences, we looked at the networks' predictions of the different lexical categories at the point of grammatical–ungrammatical divergence. For example, consider the sentences in (9a) and (9b):

(9a) *Is the boy who is hungry nearby?*
(9b) **Is the boy who hungry is nearby?*

Fig. 5 shows the mean activation of the 14 lexical output units averaged across the 10 networks after being presented with the sequence of lexical categories corresponding to *Is the boy who …* The prediction of the well-formed relative clause continuation, V (i.e., *is*), is highly preferred over the ill-formed version, ADJ (i.e., *hungry*).[5] This pattern of predictions reflects the networks' sensitivity to the statistical properties of the corpus. The networks are capable of distinguishing chunks of lexical categories that are more frequent in the training input from less frequent ones (that is, the lexical categories corresponding to PRON V ADJ [*who is hungry*] versus PRON ADJ V [*who hungry is*]).

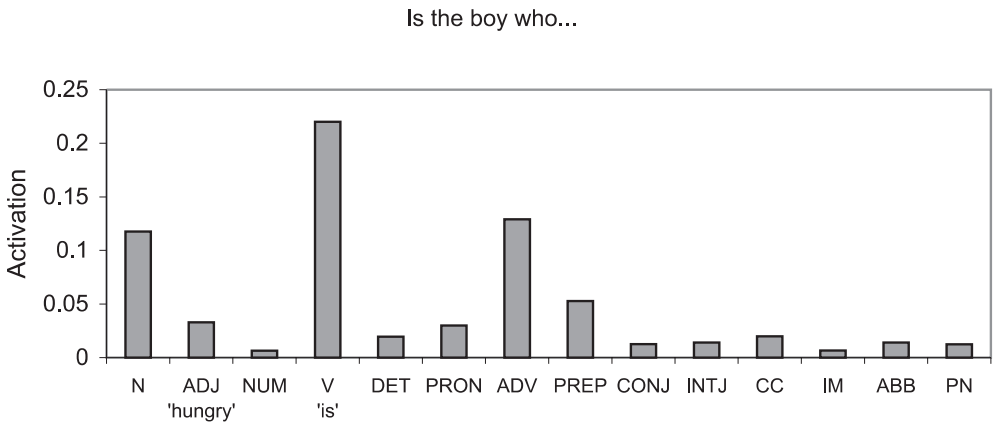

Fig. 5. Network prediction after the presentation of the lexical classes corresponding to: *"Is he boy who …"* The prediction of the well-formed relative clause continuation, V (i.e., *is*), is highly preferred over the ill-formed version, ADJ (i.e., *hungry*).

As a further indication of the networks' correct generalizations to grammatical AUX questions, we found that out of the 30 test sentences, 27 grammatical sentences produced a lower error than their ungrammatical counterparts (Fig. 6). Thus, on the assumption that sentences with lower error would be preferred, the SRNs showed a bias toward the grammatical forms in 90% of the test cases.

These results show that SRNs are able to pick up on the implicit statistical regularities demonstrated in Experiment 1. Moreover, in contrast to Experiment 1, the networks were only exposed to the distributional information of the lexical categories and not to the potentially richer distributional information present in word co-occurrences. Yet, it is also clear that children are not provided with input "tagged" for lexical categories. Rather, the child has to bootstrap both lexical categories and syntactic constraints concurrently. Fortunately, recent research has demonstrated that lexical categories can be learned from a combination of distributional patterns of word co-occurrence (e.g., Mintz, 2002, 2003; Mintz, Newport, & Bever, 2002; Redington et al., 1998), the phonological properties of words (e.g., Kelly, 1992; Monaghan, Chater, & Christiansen, 2005; Reali et al., 2003) as well as other cues (see Christiansen & Monaghan, in press, for discussion). Consequently, we hypothesize that the kind of learning modeled in our

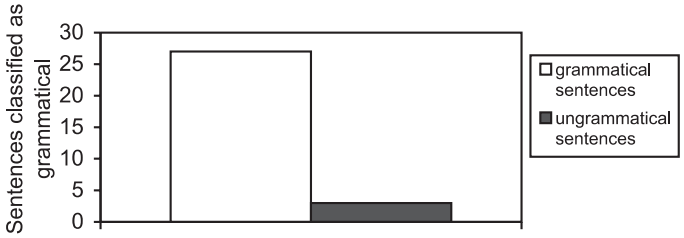

Fig. 6. Number of sentences classified correctly by the SRNs (white bars) and incorrectly as grammatical (black bars).

connectionist simulations is building on top of already learned lexical category information acquired through such multiple-cue integration.

5. General discussion

The corpus analyses indicate that there is sufficiently rich statistical information available *indirectly* in child-directed speech for making appropriate generalizations about complex AUX questions. Therefore, our results challenge the classic notion of *evidence* in the primary linguistic input, which presupposes that only explicit examples of a certain grammatical construction constitute useful evidence for its correct generalization (Boeckx & Hornstein, 2004; Chomsky, 1965; Crain & Pietroski, 2001; Legate & Yang, 2002).

Previous results suggest that children are sensitive to the same kind of statistical evidence that we found in this study. Saffran et al. (1996) demonstrated that 8-month-old children are particularly sensitive to transitional probabilities (similar to our bigram model). Sensitivity to transitional probabilities seems to be present across different domains, for instance in the segmentation of streams of tones (Saffran, Johnson, Aslin, & Newport, 1999) and visual sequences (Kirkham, Slemmer, & Johnson, 2002). These and other results on infant statistical learning (for reviews, see Gómez & Gerken, 2000; Saffran, 2003) suggest that children have mechanisms for relying on implicit statistical information.

This type of statistical learning is also evident in SRNs whose learning properties have been shown to be consistent with human learning abilities (e.g., Christiansen, Conway, & Curtin, 2005; Elman, 1990; see Christiansen & Chater, 2001, for a review). Even though the SRN model in Experiment 3 was originally developed in a different context (Reali et al., 2003), it proved to develop an appropriate bias toward the correct forms of AUX questions. Moreover, because the networks were trained to predict the next lexical category in a sentence, the pattern of network predictions (Fig. 5) can be construed as providing statistical constraints on production as well. Christiansen and Chater (1999) showed how output predictions could be construed as a set of possible sentence *continuations.* Similarly, we envisage that the output predictions of our SRN model could be used as a probabilistic basis for sentence production—when combined with other sources of constraints from semantics, and so on. The AUX questions generated in this fashion would be consistent with children's production data (Crain & Nakayama, 1987).

More important, this account of production does not require children to generate and subsequently choose between grammatical and ungrammatical versions of AUX-fronting constructions. Rather, this kind of statistically driven production would result in a single construction that would tend to be the most probable one (as illustrated by Fig. 5). However, it makes sense to evaluate the usefulness of this statistical knowledge by adopting the traditional linguistic paradigm of minimal pair comparison—in this case between grammatical and ungrammatical AUX questions. Consequently, the specific entropy values are therefore not important in an absolute sense but only in a relative sense inasmuch as they reflect what may be more likely to be produced given the underlying statistical information.

More generally, the statistical knowledge acquired by the networks is learned from positive examples only. There is no need for "negative" evidence in the form of ungrammatical sen-

tences explicitly marked as such. During learning, the networks adjust their weights to best reflect the statistical properties of the input, creating a bias toward the positive examples but at the same time biasing the network against other (ungrammatical) responses that are not likely to occur. Similarly, Schütze (1997) demonstrated how a connectionist model can learn verb subcategorization information from positive evidence alone (e.g., in contrast to prior assertions by Pinker, 1989). Of course, this emphasis on positive evidence underscores the importance of ensuring that the input provides a reasonably accurate reflection of the statistics governing the linguistic phenomena being modeled.[6]

Both bigram–trigram models and SRNs are sensitive to the probability of co-occurrence of sequences of words found in the training corpus. The grammatical and ungrammatical sentences used to test generalizations to AUX questions were almost identical, only differing in the position of the fronted *is.* Because the sequences of words in grammatical relative clauses, such as *who is hungry* or *that is there,* tend to occur frequently as chunks of words in child-directed speech, the model is able to distinguish between grammatical and ungrammatical polar interrogatives. From this we can predict that children may also be more likely to make errors involving frequently co-occurring chunks when producing complex sentences, including AUX questions. This prediction is consistent with both types of errors produced by the children in the Crain and Nakayama (1987) study. Thus, the "prefix" errors (which account for 58% of the children's errors) in, for example, **Is the boy who is being kissed by his mother is happy?* may be explained by the frequent occurrence in the input of chunks such as *mother is happy.* Similarly, the "restarting" errors (accounting for 22% of errors) as in **Is the boy that is watching Mickey Mouse, is he happy?* may derive from often-heard questions such as *Is he happy?*

This partial account of children's errors in production depends on the theoretical notion that young learners to a large degree rely on word chunks to organize their early language. This is supported by recent studies showing that much of young children's language is organized in terms of *item-based* linguistic schemas (e.g., Bybee, 2002; Cameron-Faulkner, Lieven, & Tomasello, 2003; Lieven, Pine, & Baldwin, 1997; Pine & Lieven, 1997; Tomasello, 1992), suggesting that children might construct their early language system around chunks of words. For example, Lieven et al. (1997) found that children between 2 and 3 years of age use virtually all their verbs and predicative terms in a unique sentence frame in early language development (for a review, see Tomasello, 2000, 2003). The results we have presented here can thus be seen as building on, as well as contributing to, the item-based approach to language acquisition.

Given a statistically based approach to language acquisition, an important remaining issue is where universal patterns of language structure and use may derive from. Indeed, a language without reliable structural regularities would be unlearnable from a statistical perspective because it is exactly the constituent properties of well-formed sentences that make distributional cues useful in the first place. It seems clear that such linguistic universals are likely to be due to innate constraints. However, it is an open question whether such innate constraints have to be specifically linguistic in nature, or whether they may derive from more general learning constraints not specific to language. Our view is consistent with the latter perspective and supported by recent work in language evolution demonstrating how language universals can emerge through processes of cultural transmission across many generations of learners (e.g., Batali, 2002; Brighton, 2002; Christiansen, Dale, Ellefson, & Conway, 2002; Kirby &

Christiansen, 2003) and through grammaticalization in diachronic change (e.g., Bybee, 2002; Givón, 1998; Heine & Kuteva, 2002). For example, Christiansen and Devlin (1997) demonstrated how constraints on the learning of sequential structure in an SRN may explain the emergence of word order universals. Ellefson and Christiansen (2000) further demonstrated that limitations on SRN learning of sequential material can help explain certain subjacency constraints on complex question formation. Thus, from our perspective the linguistic universals that make language learnable by statistical means derive from innate nonlinguistic constraints on the statistical learning mechanisms themselves and from general functional properties of communicative interactions.

6. Conclusion

Our corpus analyses and connectionist simulations have underscored the importance of statistical properties of word co-occurrences in the process of grammatical generalization. However, a complete model of language acquisition cannot be developed on the basis of this distributional cue alone. Young learners are likely to rely on many additional sources of information (e.g., semantic, phonological, prosodic) to be able to infer different aspects of the structure of the target language. Previous work has shown that syntactic acquisition is greatly facilitated when distributional information is integrated with other sources of probabilistic information (e.g., Christiansen & Monaghan, in press; Monaghan et al., 2005; Morgan et al., 1987). More important, the SRN has been shown to provide a useful powerful basis for such multiple-cue integration (Christiansen & Dale, 2001; Reali et al., 2003), suggesting that these results can be incorporated into a more comprehensive computational account of language acquisition.

On the theoretical side, our results indicate that the poverty of stimulus argument may not apply to the classic case of auxiliary fronting in polar interrogatives, previously a cornerstone in the argument for the innateness of grammar. Although this study only pertains to a single construction—AUX fronting in polar interrogatives—we anticipate that there are likely to be other cases in which indirect statistical information (and/or other cues) can lead to correct generalization of structure. This highlights the important issue of what counts as *sufficient evidence* for learning a particular linguistic structure, and it suggests that the general assumptions of the poverty of stimulus argument may need to be reappraised in the light of the statistical richness of language input to children.

Notes

1. We used PERL programming in a UNIX environment to implement the corpus analysis. This includes the simulation of bigram and trigram models and cross-entropy calculation and comparisons.
2. We thank an anonymous reviewer for raising this question and Eve Clark for urging that we pursue it.
3. As some of the words in the examples were not present in the Bernstein-Ratner (1984) corpus, we substituted them for semantically related ones: Thus, the words *mother,*

plate, watching, unhappy, and *bench* were replaced, respectively, by *mommy, ball, looking at, crying,* and *chair.*

4. All networks were simulated using the Lens simulator (Rohde, 1999) in a UNIX environment.
5. The relatively high activation of the noun and adverb lexical categories—N and ADV—is due to the relatively frequent occurrence of lexical category fragments such as N PRON N (*boy whose father*) and N PRON ADV (*boy who happily*). In the context of this sequence of lexical categories—V DET N PRON N—this would be consistent with the sentence fragments such as *Is the boy whose father came …* and *Is the boy who happily ate his …* We note here that the latter two AUX question examples did not appear in the corpus.
6. In this context, it is important to note that all languages differ considerably in terms of their distributional characteristics. Thus, one cannot draw any conclusions about how our approach may generalize to other languages by simply changing a few lexical items to introduce a specific aspect of some other language. Rather, to apply our approach to other languages it is crucial that appropriate corpora of child-directed speech be used; otherwise no conclusions can be drawn.

Acknowledgments

The research reported in this chapter was supported in part by a grant from the Human Frontiers Science Program (RGP0177/2001-B) to MHC.

We thank Chris Conway, Thomas Farmer and Luca Onnis for commenting on an earlier version of this article. We are also grateful to Janet Fodor, Bill Sakas and Xuan-Nga Kam for their feedback on the work presented in this article.

References

Baayen, R. H., Pipenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* [*CD-ROM*]. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Batali, J. (2002). The negotiation and acquisition of recursive grammars as a result of competition among examples. In E. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (pp. 111–172). Cambridge, England: Cambridge University Press.

Bernstein-Ratner, N. (1984). Patterns of vowel modification in motherese. *Journal of Child Language, 11,* 557–578.

Boeckx, C., & Hornstein, N. (2004). *The varying aims of linguistic theory.* Unpublished manuscript, Harvard University, Department of Linguistics, and University of Maryland, Department of Linguistics, College Park.

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.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life, 8,* 25–54.

Bybee, J. (2002). Sequentiality as the basis of constituent structure. In T. Givón & B. Malle (Eds.), *The evolution of language out of pre-language* (pp. 107–132). Philadelphia: Benjamins.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science, 27,* 843–873.

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language, 13,* 359–394.

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

Chomsky, N. (1965). *Aspects of the theory of syntax.* Boston: MIT Press.

Chomsky, N. (1980a). The linguistic approach. In M. Piatelli-Palmarini (Ed.), *Language and learning: The debate between Jean Piaget and Noam Chomsky.* Cambridge, MA: Harvard University Press.

Chomsky, N. (1980b). *Rules and representations.* Cambridge, MA: MIT Press.

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

Christiansen, M. H., & Chater, N. (2001). Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences,* 5, 82–88.

Christiansen, M. H., Conway, C. M., & Curtin, S. (2005). Multiple-cue integration in language acquisition: A connectionist model of speech segmentation and rule-like behavior. In J. W. Minett & W. S.-Y. Wang (Eds.), *Language acquisition, change and emergence: Essay in evolutionary linguistics* (pp. 205–249). Hong Kong: City University of Hong Kong Press.

Christiansen, M. H., & Dale, R. A. (2001). Integrating distributional, prosodic and phonological information in a connectionist model of language acquisition. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the twenty-third annual conference of the Cognitive Science Society* (pp. 220–225). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Christiansen, M. H., Dale, R. A. C, Ellefson, M. R., & Conway, C. M. (2002). The role of sequential learning in language evolution: Computational and experimental studies. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 165–187). London: Springer-Verlag.

Christiansen, M. H., & Devlin, J. T. (1997). Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations. In M. Shafto & P. Langley (Eds.), *Proceedings of the nineteenth annual Cognitive Science Society conference* (pp. 113–118). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Christiansen, M. H., & Monaghan, P. (in press). Discovering verbs through multiple-cue integration. In K. Hirsh-Pasek & R. M. Golinkoff (Eds.), *Action meets words: How children learn verbs*. New York: Oxford University Press.

Cleeremans, A. (1993). Attention and awareness in sequence learning. In W. Kintsch (Ed.), Proceedings of the fifteenth annual conference of the Cognitive Science Society (pp. 330–335). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cowie, F. (1998). *What's within? Nativism reconsidered.* New York: Oxford University Press.

Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language, 63,* 522–543.

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

Ellefson, M. R., & Christiansen, M. H. (2000). Subjacency constraints without universal grammar: Evidence from artificial language learning and connectionist modeling. In L. R. Gleitman & A. K. Josh (Eds.), *Proceedings of the twenty-second annual conference of the Cognitive Science Society* (pp. 645–650). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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.

Fodor, F. D., & Crowther, C. (2002). Understanding stimulus poverty arguments. *Linguistic Review, 19,* 105–145.

Givón, T. (1998). On the co-evolution of language, mind and brain. *Evolution of Communication, 2,* 45–116.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13,* 431–436.

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

Heine, B., & Kuteva, T. (2002). On the evolution of grammatical forms. In A. Wray (Ed.), *Transitions to language* (pp. 376–397). Oxford, England: Oxford University Press.

Hornstein, N., & Lightfoot, D. (1981). Introduction. In N. Hornstein & D. Lightfoot (Eds.), *Explanation in linguistics: The logical problem of language acquisition* (pp. 9–31). London: Longman.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing.* Upper Saddle River, NJ: Prentice Hall.

Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review, 99,* 349–364.

Kirby, S., & Christiansen, M. H. (2003). From language learning to language evolution. In M. H. Christiansen & S. Kirby (Eds.), *Language evolution* (pp. 272–294). Oxford, England: Oxford University Press.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence of a domain general learning mechanism. *Cognition, 83,* B35–B42.

Laurence, S., & Margolis, E. (2001). The poverty of the stimulus argument. *British Journal for the Philosophy of Science, 52,* 217–276.

Legate, J. A., & Yang, C. (2002). Empirical re-assessment of stimulus poverty arguments. *Linguistic Review, 19,* 151–162.

Lewis, J. D., & Elman, J. L. (2001). Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In B. Skarabela, S. Fish, & A. H. J. Do (Eds.), *Proceedings of the twenty-sixth annual Boston University Conference on Language Development* (pp. 359–370). Somerville, MA: Cascadilla.

Lieven, E., Pine, J., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language, 24,* 187–219.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

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

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90,* 91–117.

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26,* 393–424.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition, 96,* 143–182.

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.

Pine, J., & Lieven, E. (1997). Slot and frame patterns in the development of the determiner category. *Applied Psycholinguistics, 18,* 123–138.

Pinker, S. (1989). *Learnability and cognition.* Cambridge, MA: The MIT Press.

Pullum, G. K., & Scholz, B. (2002). Empirical assessment of stimulus poverty arguments. *Linguistic Review, 19,* 9–50.

Reali, F., Christiansen, M. H., & Monaghan, P. (2003). Phonological and distributional cues in syntax acquisition: Scaling up the connectionist approach to multiple-cue integration. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the twenty-fifth annual conference of the Cognitive Science Society* (pp. 970–975). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Rohde, D. L. T. (1999). *LENS: The light, efficient network simulator* (Tech. Rep. CMU–CS–99-164). Pittsburgh, PA: Carnegie Mellon University, Department of Computer Science.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science, 12,* 110–114.

Saffran, J. R., Aslin, R., & Newport, E. L. (1996, December 13). Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70,* 27–52.

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multi-level statistical learning by 12-month-old infants. *Infancy, 4,* 273–284.

Scholz, B., & Pullum, G. K. (2002). Searching for arguments to support linguistic nativism. *Linguistic Review, 19,* 185–223.

Schütze, H. (1997). *Ambiguity resolution in language learning.* Stanford, CA: CSLI Publications.

Tomasello, M. (1992). *First verbs: A case study of early grammatical development.* Cambridge, England: Cambridge University Press.

Tomasello, M. (2000). The item-based nature of children's early development. *Trends in Cognitive Science, 4,* 156–163.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Cambridge, MA: Harvard University Press.

Yang, C. (2002). *Knowledge and learning in natural language.* Oxford, England: Oxford University Press.