Reappraising Poverty of Stimulus Argument: A Corpus Analysis Approach

Florencia Reali and Morten H. Christiansen
Cornell University

1. Introduction

The debate between empiricism and nativism goes back to the very beginning of philosophy. More recently, the nature of linguistic structure has been the focus of discussion in the field of psycholinguistics. The *poverty of stimulus argument* for innateness of syntactic knowledge (Chomsky, 1980; Crain & Pietroski, 2001) is one of the most famous and controversial arguments in the study of language and mind. Although it has guided the vast majority of theorizing in linguistics for decades, claims about innate linguistic structure have provoked controversy and the argument is embroiled in dispute.

The poverty of stimulus argument is based on the assumption that the information in the environment is not rich enough to allow a human learner to attain adult competence. This assumption has been previously questioned in that it is based on premature conclusions about the information present on primary linguistic input (e.g., Pullum and Scholz; 2002).

Nativists have tended to dismiss a priori the idea that distributional information could play an important role in syntactic language acquisition. Nevertheless, recent studies show that distributional evidence is a potentially important source of information for word segmentation and syntactic bootstrapping (e.g., Christiansen, Allen and Seidenberg, 1998; Christiansen and Curtin, 1999; Lewis and Elman, 2001; Redington, Chater and Finch, 1999; Seidenberg and MacDonald, 2001). Moreover, infants are very sensitive to the statistical structure of the input (Saffran, Aslin and Newport, 1996), suggesting that they are able to pick up on distributional cues with a high level of accuracy. This growing body of work has provided support for the hypothesis that distributional properties of linguistic input could have a significant role on the acquisition of syntactic structure.

The validity of the poverty of stimulus argument strongly relies on the premise of absence of sufficient information in the primary linguistic input for learning the grammar. A possible approach to address this assumption is to look for statistical evidence that allows a learner to decide between grammatical and ungrammatical hypotheses. In the present study we focus on one of the most used examples to support the argument, the claim concerning auxiliary fronting in polar interrogatives. Using a corpus analysis approach, we estimated the distributional information present in the primary linguistic input, and used it to decide between grammatical and ungrammatical examples of auxiliary fronting in polar interrogatives. We found that there is enough statistical information in the corpus to correctly decide between the grammatical and ungrammatical forms of aux-questions with a high level of accuracy.

2. The poverty of stimulus argument and auxiliary fronting.

Children only hear a finite number of sentences, yet they learn to speak and comprehend sentences drawn from a grammar that can represent an infinite number of sentences. Although

the exact formulation of the poverty of stimulus argument varies it can be briefly set out as follows:

i. Human Language is characterized by a complex grammar.
ii. The information about the grammar that the learner receives is the primary linguistic input (PLI).
iii. No learner can acquire the grammar based on the information in PLI.
iv. Human infants learn their first languages by means of innately-primed learning.

According to this view, at every stage of language acquisition, inferring a syntactical rule, or determining the sub-categorization frame of a new verb, the child can make many logically possible generalizations, but generalizes correctly. The core of the argument is that the grammar cannot be acquired solely on the basis of the input. Instead, infants learn their first languages guided by experience independent internal structures. The crucial assumption of the argument- outlined in (*iii*)- is that children do not have enough data during the early stages of their life to learn the syntactic structure of their language. Thus, learning a language involves the correct generalization of the grammatical structure when insufficient data is available to children. The possible weakness of the argument lies in the difficulty to assess the input, and in the imprecise and intuitive definition of 'insufficient data'.

Given the difficulty of assessing the primary linguistic input, the assumption of lack of information has not been demonstrated, but rather intuitively stated. It also has been suggested that the input is too noisy to afford reliable learning. Thus, the argument may be relying on premature conclusions about: 1) what information is available in the primary linguistic input; 2) what can actually be learned by statistically driven learning mechanisms; and 3) the children's ability to learn statistical information. Based on recent studies that question these assumptions, we suggest that the primary linguistic input and children's learning abilities may have been previously underestimated.

One of the most used examples to support the poverty of stimulus argument concerns auxiliary fronting in polar interrogatives. Declaratives are turned into questions by fronting the correct auxiliary. Thus, 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 1), but fronting the relative clause auxiliary produces an ungrammatical sentence as in 2) (e.g., Chomsky 1965).

1. *Is the man who is hungry ordering dinner?*
2. **Is the man who hungry is ordering dinner?*

Children should be free to generate either of two sorts of rules: a structure independent rule where the first 'is' is moved or the correct structure dependent rule, where only the movement of the second 'is' is allowed. However, children never go through a period when they erroneously move the first *is* to the front of the sentence (Crain and Nakayama, 1987). Chomsky asserts that a person might go through much of his life without ever having been exposed to relevant evidence for correct inference for aux-fronting (Chomsky, in Piatelli-Parmarini, 1980).

The absence of evidence in the primary linguistic input regarding auxiliary fronting in polar interrogatives is a matter of discussion. As suggested in Lewis and Elman (2001), it is quite unlikely that a child reaches kindergarten without being exposed to sentences like:

3. *Is the boy who was playing with you still there*?'
4. *Will those who are hungry raise their hand?*'
5. *Where is the little girl full of smiles?*'

These examples have an auxiliary verb within the subject NP, 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 auxiliary fronting in polar interrogatives in the Wall Street Journal (WSJ). They found that at least five crucial examples occur in the first 500 interrogatives. These results suggest that the assumption of complete absence of evidence for aux-fronting is quite extreme. Nevertheless, Crain and Pietroski (2001) argue that the WSJ corpus is not an ideal source of the grammatical constructions that young children encounter and thus it cannot be considered representative of the primary linguistic data. Moreover, studies of the CHILDES corpus 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 and Yang, 2002).

3. Statistical Properties of Natural Languages.

There is a general agreement on the existence of innate constraints affecting the way the brain processes cognitive information. However, these constraints may not be restricted to language (Kirby and Christiansen, 2003). For example, innate constraints could affect the learning component (e.g. the relative importance of working memory). However, the nature of linguistic structure is a matter of debate. According to the generative view, children are viewed as very poor learners. On the other hand, statistical approaches highlight the role of experience-dependent factors in language acquisition. Learnability of language may be demonstrated by showing not only that there is enough statistical information to generalize the correct grammatical structure but also that children are capable of learning from such information. 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 sets of inputs with random content and length) but no relevant examples of polar interrogatives. The SRNs were able to predict the correct auxiliary fronting in aux-questions but not the incorrect ones, showing that even in total absence of relevant examples the stochastic information in the input data is sufficient to generalize correctly. These results suggest that implicit statistical regularities present in the primary linguistic input can lead to a preferred grammatical structure.

In the present study, we focused on exploring statistical cues for correct fronting in polar interrogatives in a child directed speech corpus. Even if it is the case that children hear a few relevant examples of polar interrogatives, they may be able to rely on other distributional information for learning the correct structure. In order to assess that hypothesis, we trained bigram and trigram models on the Bernstein-Ratner (1984) corpus of child-directed speech and

then tested the likelihood of novel example sentences. The test sentences included correct auxiliary fronting interrogatives (e.g. *Is the man who is hungry ordering dinner?*) and incorrect auxiliary fronting interrogatives (e.g. *Is the man who hungry is ordering dinner?*) – both not present in the training corpus. If transitional probabilities provide any cue for generalizing correctly the grammatical aux-question, then we should find a difference in the likelihood of these two alternative hypotheses.

4. Statistical Methods for Corpus Analysis.

An n-gram model is a statistical model that uses the previous (n-1) words to predict the next one. Given a string of words or a sentence it is possible to study the associated cross-entropy for that string of words according to the n-gram model trained on a particular corpus (Chen and Goodman, 1996). Thus, given two alternative sentences it is possible to measure and compare the probability of each of them, according to a corpus, indicated by the cross-entropy associated to them. Given two alternative hypotheses to formulate a polar interrogative: a) *Is the man who is in the corner smoking?* and b) *Is the man who in the corner is smoking?*, then, using an n-gram model trained in a corpus, it is possible to measure and compare the associated entropy of a) and b).

The main purpose of this study was to address whether there is enough statistical evidence in a child-directed corpus to decide between the two hypotheses a) and b). We did a corpus analysis using n-gram models trained on Bernstein-Ratner (1984) child-directed speech corpus. It contains recorded speech from nine mothers speaking to their children over 4-5 months period when children were between the ages of 1 year and 1 month to 1 year and 9 month. This is a relatively small and very noisy corpus, mostly constituted by short sentences with simple grammatical structure. The following are some example sentences:
'Oh you need some space' | 'Where is my apple?' | 'Oh'| 'That's it'.

An estimation of the information contained in it is likely to be an underestimation of the true information present in the primary linguistic input. Importantly, the corpus contains no examples of auxiliary fronting in polar interrogatives. Our hypothesis was that the corpus contains enough statistical information to decide between correct and incorrect forms of aux-questions at least to some extent.

5. Method

For corpus analysis we used bigram and trigram models of language (see e.g., Jurafsky and Martin, 2000). The probability P(*s*) of a sentence was expressed as the product of the probabilities of the words 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:

$$P(s) = \Pi_i \, P(w_i | w^{i-1}_{i-n+1})$$

To estimate the probabilities of $P(w_i|w_{i-1})$ we used the *maximum likelihood* (ML) estimate for $P(w_i|w_{i-1})$ defined as (considering the bigram model):

$P_{ML}(w_i|w_{i-1}) = P(w_{i-1}w_i) / P(w_{i-1}) = (c(w_{i-1}w_i)/Ns) / (c(w_{i-1})/Ns);$

where Ns denote the total number of tokens and $c(\alpha)$ is the number of times the string α occurs in the corpus. Given that the corpus is quite small, we used the interpolation smoothing technique defined in (Chen and Goodman, 1996).
The probability of a word (w_i) (or unigram model) is defined as:

$P_{ML}(w_i) = c(w_i)/Ns;$

The smoothing technique consists of the interpolation of the bigram model with the unigram model, and the trigram model with the bigram model. Thus, for the
bigram model we have:

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

Accordingly for trigram models we have:

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

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 measure the likelihood of a given set of sentences using the measure of cross-entropy (Chen and Goodman, 1996). The cross-entropy of a set of sentences is defined as:

$1/N_T \Sigma_i -\log_2 P(s_i)$ (where s_i is the i^{th} sentence).

The cross-entropy value of a sentence is inversely correlated with the likelihood of it. Given a training corpus, and two sentences A and B we can compare the cross-entropy of both sentences and estimate which one is more probable according to the statistical information of the corpus. 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.

6. Procedure

We used the Bernstein-Ratner child-directed speech corpus as the training corpus for bigram and trigram models. We trained the models using 10,082 sentences (1,740 words and 34,010 tokens) from the corpus. We were interested in comparing the cross-entropy of correct and incorrect polar interrogatives. For that purpose, we created two novel sets of sentences. The first one contained grammatically correct polar interrogatives and the second one contained the ungrammatical version of each sentence in the first set. For example, set 1 contains the sentence: "*Is the cat that is in the game old?"* and set 2 contains the sentence: "*Is the cat that in the game is old?"*. The sentences were created using a random algorithm that selects words from the corpus, and create sentences according to syntactic and semantic constraints. In other words, we

tried to prevent any possible bias in creating the test sentences. The test sets only contained relevant examples of polar interrogatives of the form: "*Is* / NP/ (who/that)/ *is* / A_i/ B_i?", where A_i and B_i represent a variety of different material including VP, PARTICIPLE,NP, PP, ADJP (e.g.: *"Is the lady who is there eating?*"; "*Is the dog that is on the chair black?*"). Some test sentence examples are shown in Appendix A. Each test set contained 100 sentences. We estimated the mean cross-entropy per sentence by calculating the average cross-entropy of the 100 sentences in each set. Then we compared the likelihood of correct and incorrect sentences by comparing each pair cross-entropy (correct vs. incorrect version) and choosing the version with the lower value of cross-entropy. We studied the statistical significance of the results using paired t-test analyses.

7. Results

We found that the mean cross-entropy of correct sentences was lower than mean cross entropy of incorrect sentences. We performed a statistical analysis of the cross-entropy difference, considering all pairs of correct and incorrect sentences. The cross entropy difference was highly significant $t(99)$, $p < 0.0001$. Table I summarize the results.

Table I : Comparison of mean cross-entropy.

	Mean cross-entropy		Mean difference	t(99) p-value
Bigram	22.92	23.73	0.83	< 0.0001
Trigram	21.81	23.07	1.26	< 0.0001

These results show that grammatical sentences have a higher probability than ungrammatical sentences with a high level of significance. In order to compare each grammatical-ungrammatical pair of sentences, we defined the following criteria: between each correct and incorrect 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 the criteria we found that the percentage of correctly classified sentences using trigram model is 92% and using bigram model is 95%. Figure 1 shows the performance of the models according to the defined classification criterion.

According to the bigram model, we found that only 5 out of 100 sentences were miss-classified. It is possible to calculate the probability of a sentence from the cross-entropy value. Figure 2 shows the comparison of mean probability of grammatical and ungrammatical sentences. We found that the mean probability of correct polar interrogatives is almost twice the mean probability of incorrect polar interrogatives according to the bigram model and it is more than twice according to the trigram model.

Figure 1: Number of correctly classified sentences (total number of sentences = 100)

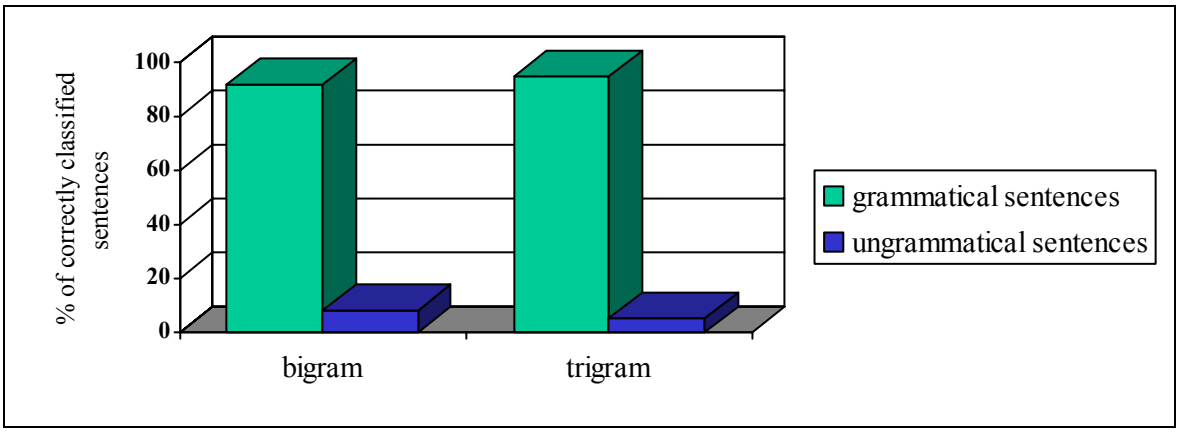
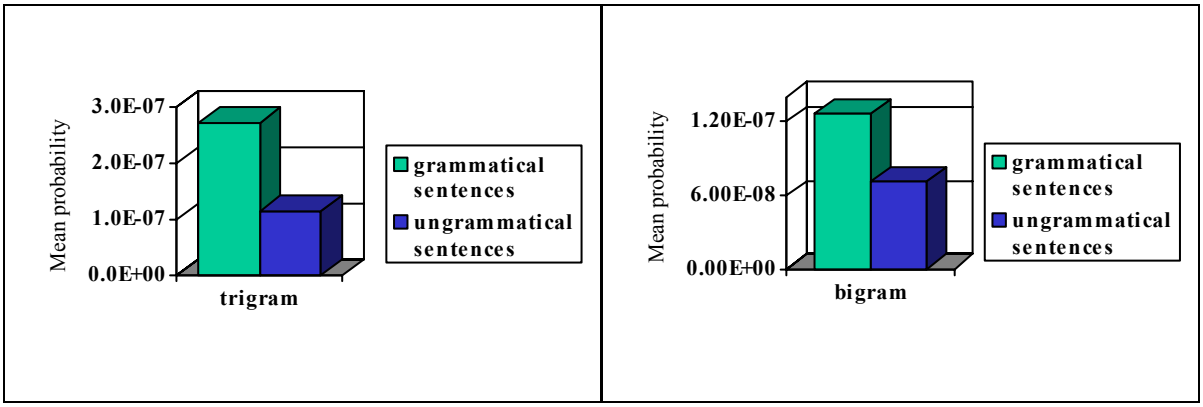


Figure 2: Mean probability of correct sentences vs. mean probability of ungrammatical sentences.



8. Discussion

Even though the corpus is quite noisy and small, it is possible to extract reliable information out of it. We showed that it contains enough statistical information to decide with high accuracy between grammatical and ungrammatical forms of auxiliary fronting in polar interrogatives. These results indicate that the necessary distributional evidence for deciding between the two alternative hypotheses is available given a bigram/trigram analysis. However, it remains to be seen whether children may be sensitive to this kind of difference in distributional information, and whether they may be able to use it for learning the structure of the language. Previous results suggest that in fact children might be sensitive to the same kind of statistical evidence that we found in the present study. Saffran, Aslin and Newport (1996) demonstrated that children as young as 8 month-old are particularly sensitive to transitional probabilities (similar to our bigram model). They confronted learners with a stream of unfamiliar concatenated speech-like sound.

The learners tend to infer word boundaries between two syllables that rarely occur adjacently in the sequence. Sensitivity to transitional probabilities seems to be present across modalities, for instance in the segmentation of streams of tones (Saffran, Johnson, Aslin, and Newport, 1999) and in the temporal presentation of visual shapes (Fiser and Aslin, 2002).

In conclusion, this study provides some insights about the possible nature of implicit statistical information that could be useful for learning the structure of a language. It suggests that, in the statement of the poverty of stimulus argument, nativists might have underestimated the information contained in the primary linguistic input. Since the conclusion of innateness of grammar strongly depends on the assumption of insufficient data available to children, these results point toward a re-assessment of the innateness of grammar.

Appendix A

Example test sentences:

Is the lady who is standing there lovely?
Is the rat that is hungry clean?
Is the kid who is sitting on the chair listening?
Is the girl who is hungry nearby?
Is the machine that is in the school bigger than this?
Is the goose that is here smelling?
Is the bunny that is in the school crying?
Is the dog that is in the corner quiet?
Is the kid who is sleeping better?
Is the door that is in the kitchen old?

References

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

Chen, S.F. and Goodman, J. (1996). An Empirical Study of Smoothing Techniques for Language Modeling. *Proceedings of the 34th Annual Meeting of the ACL.*

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Boston, MA: MIT Press.

Chomsky, N. (1980). *Rules & Representation.* Cambridge, MA: MIT Press.

Christiansen, M.H., Allen, J. and Seidenberg, M.S. (1998). Learning to Segment Speech Using Multiple Cues: A Connectionist Model. *Language and Cognitive Processes* 13: 221-268.

Christiansen, M.H. and Curtin, S.L. (1999). Transfer of learning: Rule acquisition or statistical learning? *Trends in Cognitive Sciences.* 3: 289-290.

Crain, S. and Nakayama, M. (1987). Structure dependence in grammar formation. *Language.* 63: 522-543.

Crain, S. and Pietroski, P. (2001). Nature, Nurture and Universal Grammar. *Linguistics and Philosophy*. 24: 139-186.

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

Fiser, J. and Aslin, R.N. (2002). Statistical learning of higher-order temporal structure from visual shape-sequences. *Journal of Experimental Psychology: Learning, Memory and Cognition.* 130: 658-680.

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

Kirby, S. and Christiansen, M.H. (2003) From Language Learning to Language Evolution. In *Language Evolution* (Christiansen, M.H. and Kirby, S., eds) Oxford University Press.

Legate, J.A. and Yang, C. (2002) Empirical Re-Assessment of Stimulus Poverty Arguments. *Linguistic Review.* 19: 151-162.

Lewis, J.D. and Elman, J.L. (2001). Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In *Proceedings of the 26th Annual Boston University Conference on Language Development* (pp. 359-370). Somerville, MA: Cascadilla Press.

Newport, E. and Aslin, R. (2000). Innate constrained learning: Blending old and new approaches to language acquisition. *Proceedings of the 24th Annual Boston University Conference on Language Development,* Somerville, MA: Cascadilla Press.

Piatelli-Palmerini, M. (ed.) (1980). *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Cambridge: Harvard University Press.

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

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

Redington, M., Chater, N. and Finch S. (1998). Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science.* 22: 425-469.

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

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

Seidenberg, M.S. and MacDonald, M. (2001). Constraint Satisfaction In Language Acquisition and Processing. In *Connectionist Psycholinguistics* (Christiansen, M.H., ed). Westport, Connecticut: Ablex Publishing.