Poverty of the Stimulus? A rational approach

Amy Perfors\(^1\) (perfors@mit.edu) Josh Tenenbaum\(^1\) (jbt@mit.edu) Terry Regier\(^2\) (regier@uchicago.edu)
\(^1\)MIT Department of Brain and Cognitive Sciences; \(^2\)Department of Psychology, University of Chicago

Abstract

The Poverty of the Stimulus (PoS) argument holds that children do not receive enough evidence to infer the existence of core aspects of language, such as the dependence of linguistic rules on hierarchical phrase structure. We reevaluate the PoS argument by using a rational Bayesian model of grammar to show that an unbiased learner choosing between structure-dependent and structure-independent explanations of the input will ultimately prefer the structure-dependent one for reasons of simplicity and generalizability. This choice enables the learner to master subtle aspects of syntax, such as the auxiliary fronting rule in interrogative formation, even without having heard directly relevant data (e.g., interrogatives with relative clauses in the subject NP).

Introduction

Modern linguistics was strongly influenced by Chomsky’s observation that language learners make grammatical generalizations that do not appear justified by the evidence in the input (Chomsky, 1965, 1980). The notion that these generalizations can only be explained by innate knowledge, known as the argument from the Poverty of the Stimulus (henceforth PoS), has led to an enduring debate that is central to many of the key issues in cognitive science and linguistics.

The original formulation of the Poverty of the Stimulus argument rests critically on assumptions about simplicity, the nature of the input children are exposed to, and how much evidence is “enough” to support the generalizations that children make. The phenomenon of auxiliary fronting in interrogative sentences is one example used in support of the PoS; it states that children must be innately biased to favor structure-dependent rules that operate using grammatical constructs like phrases and clauses over structure-independent rules that operate only on the sequence of words.

In auxiliary fronting, interrogatives are formed from declaratives by fronting the main clause auxiliary. Thus, given a declarative sentence like “The dog in the corner is hungry”, the interrogative is formed by moving the is to make the sentence “Is the dog in the corner hungry?” Chomsky considered two types of operation that can explain auxiliary fronting (Chomsky, 1965, 1971). The simplest is independent of the hierarchical phrase structure of the sentence: take the leftmost (first) occurrence of the auxiliary in the sentence and move it to the beginning. The structure-dependent operation (move the auxiliary from the main clause of the sentence) is more complex since it considers a sentence’s phrasal structure and not just its sequence of elements.

The “poverty” part of this form of the PoS argument claims that children do not see the data they would need to in order to rule out the structure-independent hypothesis. An example of such data would be an interrogative sentence such as “Is the man who is hungry ordering dinner?” In this sentence, the main clause auxiliary is fronted in spite of the existence of another auxiliary that would come first in the corresponding declarative sentence. Chomsky argued that this type of data is not accessible in child speech, going so far as to maintain that “it is quite possible for a person to go through life without having heard any of the relevant examples that would choose between the two principles” (Chomsky, 1971).

It is indeed accepted that children do not appear to go through a period where they consider the structure-independent hypothesis in auxiliary fronting (Crain and Nakayama, 1987). However, two other aspects of the PoS argument are the topic of much debate. The first is about the nature of poverty, questioning what evidence there is in the input and what constitutes “enough” (Pul-lum and Scholz, 2002; Legate and Yang, 2002). Unfortunately, this type of argument is inconclusive: while there is some agreement that the critical forms are rare in child-directed speech, they do occur (Legate and Yang, 2002; Pullum and Scholz, 2002). Thus the debate consists of an unsatisfying back-and-forth about how much evidence is enough and how rare is too rare, which is difficult to resolve without a clear specification of how a child’s language learning mechanism might work.

The second basic type of argumentation is about the nature of the stimulus, suggesting that regardless of whether there is enough direct syntactic evidence available, there may be enough distributional and statistical regularities in language to explain children’s behavior (Redington et al., 1998; Lewis and Elman, 2001; Reali and Christiansen, 2004). Most of the work focusing specifically on auxiliary fronting uses connectionist simulations or n-gram models to argue that there is enough information in child-directed language to predict the grammatical status of aux-fronted interrogatives (Reali and Christiansen, 2004; Lewis and Elman, 2001).

While both of these arguments are useful and the research on statistical learning in particular is new and promising, there are still notable shortcomings. Most
profoundly, the statistical models do not engage with the primary intuition and issue raised by the PoS argument. The intuition is that language has a hierarchical structure—it uses symbolic notions like syntactic categories and phrases that are hierarchically organized within sentences, which are recursively generated by a grammar. The issue is whether knowledge about this structure is learned or innate. A purely statistic or connectionist approach that denies the explicit representation of structure has two problems addressing this issue. First of all, most linguists and many cognitive scientists tend to discount these results because they ignore a principle feature of linguistic knowledge, namely that is based on structured symbolic representations. Secondly, connectionist networks and n-gram models tend to be difficult to understand analytically. For instance, the models used by Reali and Christiansen (2004) and Lewis and Elman (2001) measure success by whether they predict the next word in a sequence, rather than based on examination of an explicit grammar or on the production of grammatical but not ungrammatical sequences. Though the models perform above chance, it is difficult to tell why and what precisely they have learned.

In this work we present a Bayesian account of structure learning in language in order to engage with the PoS argument on its own terms—taking the existence of structure seriously and asking whether, and to what extent, knowledge of that structure can be inferred by a rational statistical learner. Doing this enables us to achieve a number of important goals. (1) We demonstrate that a learner equipped with the capacity to explicitly represent both structure-dependent and structure-independent grammars—but without any initial biases—will infer that the structure-dependent grammar is a better fit to typical child-directed input. (2) We show that inferring this structure-dependent grammar results in the mastery of subtle aspects of auxiliary fronting, even if no direct evidence is available in the input. (3) Our approach provides a clear and objectively sensible metric of simplicity, as well as a way to explore what sort of data and how much is required to make these structure-dependent generalizations. And (4) our results suggest that PoS arguments are sensible only when phenomena are considered as part of a linguistic system, rather than taken in isolation.

This work addresses the exact challenge posed by Chomsky, which is still salient in much of language acquisition: not whether there is enough statistical information in the input to make predictions of the next word in a sequence, but whether there is enough information to infer the kind of structure over which linguistic rules operate. A rational Bayesian approach is useful here because it gives us the capability to combine structure and statistics: we can make inferences about different kinds of grammatical structure, on the basis of statistical information in the input.

Method

We formalize the problem of picking the grammar that best fits a corpus of child-directed speech as an instance of Bayesian model selection. We assume that language is generative, i.e., that the set of grammatical sentences is generated from some grammar G. Inspired by recent work of Goldwater et al. (2005), we use a generative model for language that is divided into two components. The first is the grammar, which assigns a probability distribution over the potentially infinite set of syntactic forms that are accepted in the language. The second generates a finite observed corpus from the infinite set of forms produced by the grammar, and can account for the characteristic power-law distributions found in language (Zipf, 1932). In essence, this two-component model assumes separate generative processes for the allowable types of syntactic forms in a language and for the frequency of specific sentence tokens.

One advantage of this two-component model is that grammars are analyzed based on individual sentence types rather than by looking at the frequencies of different sentence forms. This model therefore parallels standard linguistic practice: grammar learning and comparison is based on how well each grammar accounts for the all of the types of sentence forms rather than their frequency distribution. But because the second component of the model is sensitive to token frequency, this is also compatible with usage based accounts, which hold that production frequencies can have an effect on other aspects of language. Since we are concerned with grammar comparison rather than corpus generation, we focus in this work on the first component of the model.

We compare grammars G according to a probabilistic score that combines the prior probability of G and the likelihood of a corpus C given that grammar, in accordance with Bayes’ rule:

$$p(G|C) \propto p(G)p(C|G)$$

The prior is calculated assuming a generative model of grammars (a “grammar of grammars”). The likelihood reflects the probability of a type-based corpus of child-directed speech C given G. Because the grammars we analyse are those that can successfully parse our corpora, we first consider the corpora before moving on to an explanation of the grammars.

The corpora

The corpora consist of the sentences spoken by adults in the Adam corpus (Brown, 1973) in the CHILDES database (MacWhinney, 2000). In order to focus on grammar learning rather than lexical acquisition, each word was replaced by its syntactic category. Ungrammatical sentences, sentence fragments, and the most grammatically complex sentence types were removed, leaving 6152 individual sentence tokens and 1317 unique sentence types in the final corpus. Removing the complicated sentence types is if anything a conservative move,

1Parts of speech used included determiners (det), nouns (n), adjectives (adj), prepositions (prep), pronouns (pro), proper nouns (prop), infinitives (to), participles (part), infinitive verbs (vinf), conjugated verbs (v), auxiliary verbs (aux), complementizers (comp), and wh-question words (wh). Adverbs and negations were removed from all sentences.

2Removed types included topicalized sentences as well as sentences containing subordinate phrases, sentential comple-
since the hierarchical, structure-dependent grammar is more preferred as the input grows more complicated. It also makes the data analysis more tractable.

In order to explore how the preference for a grammar is dependent on the level of evidence in the input, six corpora were created as subsets of the main corpus. Under the reasoning that the most frequent sentences are most available as evidence, the Level 1 corpus contained only those sentence forms that occurred over 100 times in the corpus (six unique sentence types, all declarative). The Level 2 corpus, containing 19 types (including aux-fronted interrogatives and wh-questions), consisted of items that occurred at least 50 times in the input. Level 3 contained forms that occurred 25 or more times (39 types). Level 4 was the corpus of forms occurring 10 or more times (100 types), and Level 5 included all forms occurring at least twice (531 types). The complete corpus, Level 6, contained 1317 unique types, including interrogatives, wh-questions, relative clauses, prepositional and adverbial phrases, command forms, and auxiliary as well as non-auxiliary verbs.

### The grammars

Because this work was motivated by the distinction between structure-dependent and structure-independent rules, we wanted to compare grammars that clearly differed from each other structurally. The structure-dependent (hierarchical) grammar was context-free, since CFGs generate parse trees with hierarchical structure and are accepted as a reasonable “first approximation” to the grammars of natural language (Chomsky, 1959). We chose two different kinds of grammars to represent structure-independent systems. The first, which we call the linear grammar, is simply a memorized list of each of the syntactic sentences that occur in the corpus. Because Chomsky often compared language to a Markov model in which each word depended only on the word immediately before it in the sequence, we compared these two grammars to a regular grammar as well. All grammars are probabilistic, meaning that each production is associated with a probability and the probability of any given parse is the product of the probabilities of the productions involved in the derivation.

The probabilistic context-free grammar (PCFG) was the most linguistically accurate grammar we could devise that could parse all of the forms in the corpus. The full grammar, used for the Level 6 corpus, contained the 13 terminals, 14 non-terminals, and 50 productions. All grammars at other levels included only the subset of productions and items necessary to parse that corpus.

The probabilistic regular grammar (PRG) was derived directly from the context-free grammar by converting all productions not already of the form $A \rightarrow a$ or $A \rightarrow aB$. When it was possible to do so without loss of generalizeability, the resulting productions were simplified, and any productions that were not used to parse the corpus were eliminated. The final regular grammar contained 13 terminals, 65 non-terminals, and 249 productions. The number of productions was so much greater than in the PCFG because each context-free production containing two nonterminals in a row had to be expanded into a series of productions (e.g. $NP \rightarrow NP PP$ expands to e.g., $NP \rightarrow pro PP$, $NP \rightarrow n PP$, etc.). To illustrate this, Table 1 compares how the NP phrases are handled in the context-free and regular grammar.

<table>
<thead>
<tr>
<th>Context-free grammar</th>
<th>Regular grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>$NP \rightarrow NP PP</td>
<td>NP CP</td>
</tr>
<tr>
<td>pro</td>
<td>prop</td>
</tr>
<tr>
<td>$N \rightarrow n</td>
<td>adj N</td>
</tr>
<tr>
<td>$NP PP \rightarrow n PP</td>
<td>adj N_{PP}$</td>
</tr>
<tr>
<td>$N_{CP} \rightarrow n CP</td>
<td>adj N_{CP}$</td>
</tr>
<tr>
<td>$N_{C} \rightarrow n C</td>
<td>adj N_{C}$</td>
</tr>
</tbody>
</table>

Table 1: Sample NP productions from two grammar types.

The linear grammar was a comprehensive list of each sentence form that appeared in the corpus. It contained zero non-terminals (aside from $S$) and 1317 productions.

### Scoring the grammars: prior probability

Because both regular and linear grammars are special cases of context-free grammars in general, we assume a generative model for creating the grammars under which each grammar is selected from the space of PCFGs with a vocabulary size $V$ (including terminals and nonterminals). Each symbol in each production is created by selecting a term from a uniform distribution of the vocabulary items used in that grammar, and therefore has probability $\frac{1}{V}$ of being chosen. The prior probability for a grammar with $P$ productions and $N_i$ symbols for production $i$ is thus given by:

$$p(G) = p(P) \prod_{i=1}^{P} p(N_i) \prod_{j=1}^{N_i} \frac{1}{V} \quad (1)$$

Because of the small numbers involved, all calculations are done in the log domain. The probability of $P$ productions ($p(P)$) and $N_i$ symbols ($p(N_i)$) is calculated assuming that the distribution of the number of productions and items are exponential functions with mean 1. Thus, simpler grammars – those with fewer productions and symbols – are given higher prior probability. Grammars with the fewest vocabulary items and productions and the shortest right hand sides are the thus the simplest according to this prior.

---

3Partitioning in this way, by frequency alone, allows us to stratify the input in a principled way; additionally, the higher levels include not only rarer forms but also more complex ones, and thus levels may be thought of as loosely corresponding to complexity.

4The full grammars are available at [http://www.mit.edu/~perfor/cogsci06/archive.html](http://www.mit.edu/~perfor/cogsci06/archive.html).
Scoring the grammars: likelihood

The likelihood $p(C|G)$ reflects how likely the corpus $C$ is to be generated by the grammar $G$. It is calculated as the product of the likelihoods of each sentence type $S$ in the corpus. If the set of sentences is partitioned into $k$ unique types, the log likelihood is given by:

$$\log(p(C|G)) = \sum_{i=1}^{k} \log(p(S_i|G))$$  \hspace{1cm} (2)

The probability of any sentence type $k$ given the grammar $(p(S_k|G))$ is the product of the probability of each production in the grammar used to derive the sentence type; thus sentences with longer derivations will tend to be less probable – another way in which simplicity is naturally favored by this model. The probability of each production in each grammar is set to the maximum likelihood value.

Basic results

The posterior probability of a grammar $G$ is the product of the likelihood and the prior, or the sum of the log likelihood and the log prior. All scores are presented as log probabilities and thus are negative, meaning that lower absolute values are more probable.

Prior probability

Table 2 shows the prior probability of each grammar type on each corpus. In keeping with intuitive notions of simplicity, when there is little evidence available in the input the simplest grammar that accounts for all the data is the structure-independent linear grammar. However, by Level 3, the simplest grammar that could account for the data is structure-dependent. As the number of unique sentences and the length of the average sentence increases, the linear grammar becomes too costly to compete with the abstraction offered by the structure-dependent PCFG. The regular grammar is the best of neither world: it has too many productions and vocabulary items even on the smallest corpus, but the generalization ability of the grammar is poor enough that additional sentences in the input necessitate adding so many new productions that this early cost is never regained. The context-free grammar is the most instructive. For the simple sentences in Level 1, it requires a full 16 productions and 15 vocabulary items to handle just six sentences, and thus has the lowest relative prior probability. However, its generalization ability is sufficiently great that additions to this corpus require the addition of very few additional productions: as a result, it quickly overtake both of the structure-independent grammars in terms of simplicity.

Comparing the Level 2 and Level 3 corpora allows us to analyse which kind of input is most crucial for making the transition from structure-independent to structure-dependent grammars. The Level 2 corpus contains interrogatives and sentences involving both auxiliary verbs and normal verbs, but lacks any sentences with prepositional phrases or NPs containing nested adjectives (e.g. det adj adj n). Both of these are examples of recursive productions (N → adj N and NP → NP PP), and in general, this is where the advantage of the PCFG comes through most clearly. For new recursive elements in a corpus, a regular grammar must often add an entire new subset of productions, as is evident in the subset of the grammar shown in Table 1. This limits its generalizability and also means that adding new levels is proportionally more costly for a regular grammar. Because any new sentence form requires adding more productions to a linear grammar, it is also costly there. However, additional levels of recursion in the input add no productions to a context-free grammar, and thus represent a real gain in its relative probability.

Likelihoods

The likelihood scores for each grammar on each corpus are shown in Table 2. It is not surprising that the linear grammar has the highest likelihood score on all six corpora – after all, as a perfectly memorized list of each of the sentence types, it does not generalize beyond the data at all. This is an advantage when calculating strict likelihood, though of course a disadvantage for a language-learner wishing to make generalizations that go beyond the data. Another reason that the linear grammar is preferred is that grammars with recursive productions are penalized when calculating likelihood scores based on finite input. This is because recursive grammars will generate an infinite set of sentences that do not exist in any finite corpus, and some of the probability mass will be wrongly allocated to those sentences.

The likelihood preference for a structure-independent grammar does not mean that it should be preferred overall. Most importantly, preference is based on the the posterior probability – the combination of likelihood and prior – rather than likelihood alone. And for larger corpora, the slight disadvantage of the PCFG in the likelihood is vastly outweighed by its large advantage due to the simplicity of the grammar. Secondly, as the corpus size increases, all the trends increasingly favor the context-free grammar: it becomes ever simpler relative to the increasingly unwieldy regular and linear grammars required to handle new productions.

Generalizeability

Perhaps most interestingly for language learning, the context-free grammar offers the best generalizeability. One measure of this is what percentage of larger corpora that a grammar based on a smaller corpus can parse. If the smaller grammar can parse sentences in the larger corpus that did not exist in the smaller corpus, it has generalized beyond the input in the smaller corpus. Table 3 shows the percentage of sentence types and tokens in the full (Level 6) corpus that can be parsed by each
grammar corresponding to each of the smaller levels of evidence. At all levels of evidence, the PCFG shows the highest level of generalizeability, followed by the PRG. The linear grammar does not generalize at all: at each level it can only parse the sentence types it has direct experience with.

More concretely, we can also see the improved generalizeability of the PCFG in auxiliary fronting. The PCFG can parse aux-fronted interrogatives with relative clauses in subject NP position – Chomsky’s critical forms – despite never having seen an example of that form in the input. Table 4 compares the performance of each grammar on different sentence types, including the critical forms. The PCFG can parse the critical form because it has seen simple declaratives and interrogatives, which allowed it to add productions in which the interrogative production is an aux-initial sentence that does not contain the auxiliary in the main clause. By Level 5 the grammar has also seen relative clauses, which are parsed as part of the noun phrase using the production NP → NP CP. Thus, the PCFG will correctly generate an interrogative with a relative clause in the subject NP.

Unlike the PCFG, the PRG is incapable of making the correct generalization. Although the regular grammar has productions corresponding to a relative clause in an NP, it has no way of encoding whether or not a verb phrase lacking in the main clause auxiliary should follow that NP. This is because there was no input in which such a verb phrase did occur, so the only relative clauses occur either at the end of a sentence in the object NP, or followed by a normal verb phrase. It would require further evidence from the input – namely, examples of exactly the sentences that Chomsky argues are lacking – to be able to make the correct generalization.

### Conclusion and discussion

Our model of language learning suggests that there is sufficient evidence in the input for a rational learner to conclude that language is structure-dependent without being innately biased to do so. Because of this, such a learner can correctly form interrogatives by fronting the main clause auxiliary, even if they hear none of the crucial data Chomsky identified. Furthermore, our account makes a clear suggestion about what properties of the input – namely, sentences with nested phrase structure and recursive productions – are responsible for this transition. It thus makes predictions that can be tested either by analyzing child input or studying artificial grammar learning in adults.

Our findings here also make an important general point that has frequently been overlooked in considering stimulus poverty arguments, namely that children learn grammatical rules as a part of a system. As with auxiliary fronting, most PoS arguments consider some isolated linguistic phenomenon and conclude that because there is not enough evidence for that phenomenon in isolation, it must be innately specified. We have shown here that while there might not be direct evidence for a phenomenon taken individually, there may be enough accumulated evidence about the system of which it is a part to explain the phenomenon itself.

One advantage of the account we present here is that it allows us to formally engage with the notion of simplicity. In making the simplicity argument Chomsky appealed to the notion of a neutral scientist who would rationally should first consider the structure-independent hypothesis because it is a priori less complex (Chomsky, 1971). The question of what a “neutral scientist” would do is especially interesting in light of the fact that Bayesian models are considered by many to be an implementation of inductive inference (Jaynes, 2003). Our model incorporates an automatic notion of simplicity that favors hypotheses with fewer parameters over more complex ones. We use this notion to show that, for the sparsest levels of evidence, a basic structure-independent model is

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Linear</th>
<th>PRG</th>
<th>PCFG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>-74</td>
<td>-144</td>
<td>-161</td>
</tr>
<tr>
<td>Level 2</td>
<td>-236</td>
<td>-310</td>
<td>-255</td>
</tr>
<tr>
<td>Level 3</td>
<td>-589</td>
<td>-526</td>
<td>-385</td>
</tr>
<tr>
<td>Level 4</td>
<td>-1654</td>
<td>-906</td>
<td>-506</td>
</tr>
<tr>
<td>Level 5</td>
<td>-11048</td>
<td>-2034</td>
<td>-621</td>
</tr>
<tr>
<td>Level 6</td>
<td>-31858</td>
<td>-4178</td>
<td>-621</td>
</tr>
</tbody>
</table>

Table 2: Log prior, likelihood, and posterior probabilities of each grammar for each level of evidence in the corpora.

<table>
<thead>
<tr>
<th>Grammar</th>
<th>% types</th>
<th>% tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>0.4%</td>
<td>19%</td>
</tr>
<tr>
<td>Level 2</td>
<td>1%</td>
<td>35%</td>
</tr>
<tr>
<td>Level 3</td>
<td>3%</td>
<td>47%</td>
</tr>
<tr>
<td>Level 4</td>
<td>8%</td>
<td>62%</td>
</tr>
<tr>
<td>Level 5</td>
<td>41%</td>
<td>87%</td>
</tr>
</tbody>
</table>

Table 3: Proportion of sentences in the full corpus that are parsed by smaller grammars of each type (linear, PRG, PCFG). The Level 1 grammar is the smallest grammar of that type that can parse the Level 1 corpus. All Level 6 grammars can parse the full (Level 6) corpus.
The boy who is reading is happy. (det n comp aux part aux adj)

He is happy. (pro aux adj)

Example

Table 4: Likelihood of specific sentences under each grammar. Empirical frequencies refer to the number of declarative and interrogative sentence tokens in the full corpus with complex subject NPs (those with a relative clause) and simple subject NPs (those without). Reported likelihoods are for the specific example, not the average of all sentences of that kind.

<table>
<thead>
<tr>
<th>Type</th>
<th>Subject NP</th>
<th># in input</th>
<th>Example</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decl</td>
<td>Simple</td>
<td>4244</td>
<td>He is happy. (pro aux adj)</td>
<td>-3.673 -3.295</td>
</tr>
<tr>
<td>Int</td>
<td>Simple</td>
<td>1907</td>
<td>Is he happy? (aux pro adj)</td>
<td>-5.029 -4.747</td>
</tr>
<tr>
<td>Decl</td>
<td>Complex</td>
<td>1</td>
<td>The boy who is reading is happy. (det n comp aux part aux adj)</td>
<td>-16.351 -12.316</td>
</tr>
<tr>
<td>Int</td>
<td>Complex</td>
<td>0</td>
<td>Is the boy who is reading happy? (aux det n comp aux part adj)</td>
<td>-17.075 -∞</td>
</tr>
</tbody>
</table>

We have demonstrated that a child equipped with the resources to learn a range of symbolic grammars that differ in structure and the ability to find the best fitting grammars of various types, can in principle infer the appropriateness of hierarchical phrase-structure grammars without the need for innate biases to that effect. Much harder questions remain: can we devise an automated learning algorithm that efficiently searches a vast space of grammar hypotheses and reliably converges on the best grammar, and does its learning trajectory resemble in interesting ways the development of children’s grammatical knowledge?

Acknowledgments

Thanks to Virginia Savova for helpful comments. The first author is funded by an NDSEG grant.

References


