Poverty of the Stimulus? A Rational Approach

Amy Perfors[1] (perfors@mit.edu), Joshua B. Tenenbaum[1] (jbt@mit.edu), and Terry Regier[2] (regier@uchicago.edu)
[1]Department of Brain and Cognitive Sciences, MIT; [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 one version of this argument with a Bayesian model of grammar induction, and show that a rational learner without any initial language-specific biases could learn this dependency given typical child-directed input. This choice enables the learner to master aspects of syntax, such as the auxiliary fronting rule in interrogative formation, even without having heard directly relevant data (e.g., interrogatives containing an auxiliary in a relative clause 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 best 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 Stimulus argument rests critically on assumptions about simplicity, the nature of the input children are exposed to, and how much evidence is sufficient to support the generalizations that children make. The phenomenon of auxiliary fronting in interrogative sentences is one example of the PoS argument; here, the argument 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.

English interrogatives are formed from declaratives by fronting the main clause auxiliary. 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 (linear) rule 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 (hierarchical) rule – move the auxiliary from the main clause of the sentence – is more complex since it operates over 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 (linear) 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, maintaining 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 mostly accepted that children do not appear to go through a period where they consider the linear hypothesis (Crain and Nakayama, 1987). However, two other aspects of the PoS argument are the topic of much debate. The first considers what evidence there is in the input and what constitutes "enough" (Pullum and Scholz, 2002; Legate and Yang, 2002). Unfortunately, this approach 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). Lacking a clear specification of how a child's language learning mechanism might work, it is difficult to determine whether that input is sufficient.

The second issue concerns the nature of the stimulus, suggesting that regardless of whether there is enough direct syntactic evidence available, there may be sufficient 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 child-directed language contains enough information to predict the grammatical status of aux-fronted interrogatives (Reali and Christiansen, 2004; Lewis and Elman, 2001).

While both of these approaches are useful and the research on statistical learning in particular is promising, there are still notable shortcomings. First of all, 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. An approach that lacks an explicit representation of structure has two problems addressing this issue. First of all, many linguists and cognitive scientists tend to discount these results because they ignore a principal feature of linguistic knowledge, namely that it 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. 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 linguistic structure learning 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. This is an ideal learnability analysis: our question is not whether a learner without innate language-specific biases *must* be able infer that linguistic structure is hierarchical, but rather whether it is *possible* to make that inference. It thus addresses the exact challenge posed by the PoS argument, which holds that such an inference is not possible.

The Bayesian approach provides the capability of combining structured representation with statistical inference, which enables us to achieve a number of important goals. (1) We demonstrate that a learner equipped with the capacity to explicitly represent both hierarchical and linear grammars – but without any initial biases – could infer that the hierarchical grammar is a better fit to typical child-directed input. (2) We show that inferring this hierarchical grammar results in the mastery of aspects of auxiliary fronting, even if no direct evidence is available. (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 hierarchical 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.

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. The model assumes that linguistic data is generated by first picking a type of grammar T, then selecting as an instance of that type a specific grammar G from which the data D is generated. We compare grammars according to a probabilistic score that combines the prior probability of G and T and the likelihood of corpus data D given that grammar, in accordance with Bayes' rule:

$$p(G, T|D) \propto p(D|G, T)p(G|T)p(T)$$

Because this analysis takes place within an ideal learning framework, we assume that the learner is able to effectively search over the joint space of G and T for grammars that maximize the Bayesian scoring criterion. We do not focus on the question of whether the learner can successfully search the space, instead presuming that an ideal learner can learn a given G, T pair if it has a higher score than the alternatives. Because we only compare grammars that can parse our corpus, we first consider the corpus before explaining the grammars.

The corpus

The corpus consists 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 is replaced by its syntactic category.[1] Ungrammatical sentences and the most grammatically complex sentence types are removed.[2] The final corpus contains 21792 individual sentence tokens corresponding to 2338 unique sentence types out of 25876 tokens in the original corpus.[3] Removing the complicated sentence types, done to improve the tractability of the analysis, is if anything a conservative move since the hierarchical grammar is more preferred as the input grows more complicated.

In order to explore how the preference for a grammar is dependent on the level of evidence in the input, we create six smaller corpora as subsets of the main corpus. Under the reasoning that the most frequent sentences are most available as evidence,[4] different corpus *Levels* contain only those sentence forms that occur with a certain frequency in the full corpus. The levels are: *Level 1* (contains all forms occurring 500 or more times, corresponding to 8 unique types); *Level 2* (300 times, 13 types); *Level 3* (100 times, 37 types); *Level 4* (50 times, 67 types); *Level 5* (10 times, 268 types); and the complete corpus, *Level 6*, with 2338 unique types, including interrogatives, wh-questions, relative clauses, prepositional and adverbial phrases, command forms, and auxiliary as well as non-auxiliary verbs.

[1] Parts of speech used included determiners (*det*), nouns (*n*), adjectives (*adj*), comments like "mmhm" (*c*, sentence fragments only), 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.

[2] Removed types included topicalized sentences (66 utterances), sentences containing subordinate phrases (845), sentential complements (1636), conjunctions (634), serial verb constructions (459), and ungrammatical sentences (444).

[3] The final corpus contained forms corresponding to 7371 sentence fragments. In order to ensure that the high number of fragments did not affect the results, all analyses were also performed for the corpus with those sentences removed. There was no qualitative change in any of the findings.

[4] Partitioning 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.

The grammars

Because this work is motivated by the distinction between rules operating over linear and hierarchical representations, we would like to compare grammars that differ structurally. The hierarchical grammar is 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 choose two different types of linear (structure-independent) grammars. The first, which we call the *flat grammar*, is simply a list of each of the sentences that occur in the corpus; it contains zero non-terminals (aside from S) and 2338 productions corresponding to each of the sentence types. Because Chomsky often compared language to a Markov model, we consider a *regular grammar* as well.

Though the flat and regular grammars may not be of the precise form envisioned by Chomsky, we work with them because they are representative of simple syntactic systems one might define over the linear sequence of words rather than the hierarchical structure of phrases; additionally, it is straightforward to define them in probabilistic terms in order to do Bayesian model selection. 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) is the most linguistically accurate grammar we could devise that could parse all of the forms in the corpus: as such, it contains the syntactic structures that modern linguists employ, such as noun and verb phrases. The full grammar, used for the *Level 6* corpus, contains 14 terminals, 14 nonterminals, and 69 productions. All grammars at other levels include only the subset of productions and items necessary to parse that corpus.

The probabilistic regular grammar (PRG) is derived directly from the context-free grammar by converting all productions not already consistent with the formalism of regular grammar ($A \to a$ or $A \to aB$). When possible to do so without loss of generalization ability, the resulting productions are simplified and any unused productions are eliminated. The final regular grammar contains 14 terminals, 85 non-terminals, and 390 productions. The number of productions is greater than in the PCFG because each context-free production containing two non-terminals in a row must be expanded into a series of productions (e.g. NP → NP PP expands to NP → pro PP, NP → n PP, etc). To illustrate this, Table 1 compares NPs in the context-free and regular grammars.[5]

Context-free grammar	
NP → NP PP | NP CP | NP C | N | det N | adj N	
pro | prop	
N → n | adj N	
Regular grammar	
NP → pro | prop | n | det N | adj N	
pro PP | prop PP | n PP | det N_{PP} | adj N_{PP}	
pro CP | prop CP | n CP | det N_{CP} | adj N_{CP}	
pro C | prop C | n C | det N_C | adj N_C	
N → n | adj N	N_{PP} → n PP | adj N_{PP}
N_{CP} → n CP | adj N_{CP}	N_C → n C | adj N_C

Table 1: Sample NP productions from two grammar types.

Scoring the grammars: prior probability

We assume a generative model for creating the grammars under which each grammar is selected from the space of grammars by making a series of choices: first, the grammar type T (flat, regular, or context-free); next, the number of non-terminals, productions, and number of right-hand-side items each production contains. Finally, for each item, a specific symbol is selected from the set of possible vocabulary (non-terminals and terminals). The prior probability for a grammar with V vocabulary items, n nonterminals, P productions and N_i symbols for production i is thus given by:[6]

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

Because of the small numbers involved, all calculations are done in the log domain. For simplicity, $p(P)$, $p(n)$, and $p(N_i)$ are all assumed to be geometric distributions with parameter 0.5.[7] Thus, grammars with fewer productions and symbols are given higher prior probability.

Notions such as minimum description length and Kolmogorov complexity are also used to capture inductive biases towards simpler grammars (Chater and Vitanyi, 2003; Li and Vitanyi, 1997). We adopt a probabilistic formulation of the simplicity bias because it is efficiently computable, derives in a principled way from a clear generative model, and integrates naturally with how we assess the fit to corpus data, using standard likelihood methods for probabilistic grammars.

Scoring the grammars: likelihood

Inspired by Goldwater et al. (2005), the likelihood is calculated assuming a language model that is divided into two components. The first component, the grammar, assigns a probability distribution over the potentially infinite set of syntactic forms that are accepted in the language. The second component generates a finite

[5] The full grammars are available at `http://www.mit.edu/~perfors/cogsci06/archive.html`.

[6] This probability is calculated in subtly different ways for each grammar type, because of the different constraints each kind of grammar places on the kinds of symbols that can appear in production rules. For instance, with regular grammars, because the first right-hand-side item in each production must be a terminal, the effective vocabulary size V when choosing that item is $\frac{1}{\#\text{ terminals}}$. However, for the second right-hand-side item in a regular-grammar production or for any item in a CFG production, the effective V is $\frac{1}{\#\text{ terminals} + \#\text{ non-terminals}}$, because that item can be either a terminal or a non-terminal. This prior thus slightly favors linear grammars over functionally equivalent context-free grammars.

[7] Qualitative results are similar for other parameter values.

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 approach is that grammars are analyzed based on individual sentence types rather than on the frequencies of different sentence forms. This parallels standard linguistic practice: grammar learning is based on how well each grammar accounts for the set of grammatical sentences rather than their frequency distribution. Since we are concerned with grammar comparison rather than corpus generation, we focus here on the first component of the model.

The likelihood $p(D|G,T)$ reflects how likely the corpus data D was 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:

$$\log(p(D|G,T)) = \sum_{i=1}^{k} \log(p(S_i|G,T)) \quad (2)$$

The probability $p(S_i|G,T)$ of generating any sentence type i is the sum of the probabilities of generating all possible parses of that sentence under the grammar G. The probability of a specific parse is the product of the probability of each production in the grammar used to derive that parse. We assume for simplicity that all productions with the same left-hand side have the same probability, in order to avoid giving grammars with more productions more free parameters to adjust in fitting the data; a more complex analysis could assign priors over these production-probabilities and attempt to estimate them or integrate them out.

Results

The posterior probability of a grammar G is the product of the likelihood and the prior. All scores are presented as log probabilities and thus are negative; smaller absolute values correspond to higher probabilities.

Prior probability

Table 2 shows the prior probability of each grammar type on each corpus. When there is little evidence available in the input the simplest grammar that accounts for all the data is the structure-independent flat grammar. However, by *Level 4*, the simplest grammar that can parse the data is hierarchical. As the number of unique sentences and the length of the average sentence increases, the flat grammar becomes too costly to compete with the abstraction offered by the PCFG. The regular grammar has too many productions and vocabulary items even on the smallest corpus, plus its generalization ability 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 more complicated than necessary on the smallest corpus, requiring 17 productions and 7 nonterminals to parse just eight sentences, and thus has the lowest relative prior probability. However, its generalization ability is sufficiently great that additions to the corpus require few additional productions: as a result, it quickly becomes simpler than either of the linear grammars.

What is responsible for the transition from linear to hierarchical grammars? Smaller corpora do not contain elements generated from recursive productions (e.g., nested prepositional phrases, NPs with multiple adjectives, or relative clauses) or multiple sentences using the same phrase in different positions (e.g., a prepositional phrase modifying an NP subject, an NP object, a verb, or an adjective phrase). While a regular grammar must often add an entire new subset of productions to account for them, as is evident in the subset of the grammar shown in Table 1, a PCFG need add few or none. As a consequence, both linear grammars have poorer generalization ability and must add proportionally more productions in order to parse a novel sentence.

Likelihoods

The likelihood scores for each grammar on each corpus are shown in Table 2. It is not surprising that the flat grammar has the highest likelihood score on all six corpora – after all, as a 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 flat 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 allocated to those sentences.

The likelihood preference for a flat grammar does not mean that it should be preferred overall. Preference is based on the the posterior probability rather than likelihood alone. For larger corpora, the slight disadvantage of the PCFG in the likelihood is outweighed by the large advantage due to its simplicity. Furthermore, as the corpus size increases, all the trends favor the hierarchical grammar: it becomes ever simpler relative to the increasingly unwieldy linear grammars.

Generalizability

Perhaps most interestingly for language learning, the hierarchical grammar generalizes best to novel items. One measure of this is what percentage of larger corpora a grammar based on a smaller corpus can parse. If the smaller grammar can parse sentences in the larger cor-

	Prior			Likelihood			Posterior		
Corpus	Flat	PRG	PCFG	Flat	PRG	PCFG	Flat	PRG	PCFG
Level 1	**-68**	-116	-133	**-17**	-19	-29	**-85**	-135	-162
Level 2	**-112**	-165	-180	**-33**	-36	-56	**-145**	-201	-236
Level 3	-405	-394	**-313**	**-134**	-179	-243	**-539**	-573	-556
Level 4	-783	-560	**-384**	**-281**	-398	-522	-1064	-958	**-906**
Level 5	-4087	-1343	**-541**	**-1499**	-2379	-2891	-5586	-3722	**-3432**
Level 6	-51505	-5097	**-681**	**-18128**	-36392	-38421	-69633	-41489	**-39102**

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

pus 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 corpus that can be parsed by each grammar corresponding to each of the smaller levels of evidence. The context-free grammar always shows the highest level of generalizability, followed by the regular grammar. The flat grammar does not generalize: at each level it can only parse the sentences it has direct experience of.

	% types			% tokens		
Grammar	Flat	RG	CFG	Flat	RG	CFG
Level 1	0.3%	0.7%	2.4%	9.8%	31%	40%
Level 2	0.5%	0.8%	4.3%	13%	38%	47%
Level 3	1.4%	4.5%	13%	20%	62%	76%
Level 4	2.6%	13%	32%	25%	74%	88%
Level 5	11%	53%	87%	34%	93%	98%

Table 3: Proportion of sentences in the full corpus that are parsed by smaller grammars of each type. The *Level 1* grammar is the smallest grammar of that type that can parse the *Level 1* corpus. All *Level 6* grammars parse the full corpus.

PCFGs also generalize more appropriately in the case of auxiliary fronting. The PCFG can parse aux-fronted interrogatives containing subject NPs that have relative clauses with auxiliaries – Chomsky's critical forms – despite never having seen an example in the input, as illustrated in Table 4. The PCFG can parse the critical form because it has seen simple declaratives and interrogatives, allowing it to add productions in which the interrogative production is an aux-initial sentence that does not contain the auxiliary in the main clause. The grammar also has 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 an aux-containing relative clause in the subject NP.

Unlike the PCFG, the PRG cannot make 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 without a 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.

Discussion and conclusions

Our model of language learning suggests that there may be sufficient evidence in the input for an ideal rational learner to conclude that language is structure-dependent without having an innate language-specific bias 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. Our account suggests that certain properties of the input – namely sentences with phrases that are recursively nested and in multiple locations – may be 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 also make a general point that has sometimes been overlooked in considering stimulus poverty arguments, namely that children learn grammatical rules as a part of a *system* of knowledge. 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 innate. We have shown here that while there might not be direct evidence for an individual phenomenon, there may be enough 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 rationally should first consider the linear 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 linear grammar is simpler; but our model also demonstrates that this simplicity is outweighed by

				Can parse?		
Type	Subject NP	in input?	Example	Flat	RG	CFG
Decl	Simple	Y	He is happy. (pro aux adj)	Y	Y	Y
Int	Simple	Y	Is he happy? (aux pro adj)	Y	Y	Y
Decl	Complex	Y	The boy who is reading is happy. (det n comp aux part aux adj)	Y	Y	Y
Int	Complex	N	Is the boy who is reading happy? (aux det n comp aux part adj	N	N	Y

Table 4: Ability of each grammar to parse specific sentences. Only the PCFG can parse the complex interrogative sentence.

the improved performance of a hierarchical grammar on larger quantities of realistic input. Interestingly, the input in the first Adam transcript at the earliest age (27mo) was significantly more diverse and complicated than the frequency-based *Level 1* corpus; indeed, of the three, the hierarchical grammar had the highest posterior probability on that transcript. This suggests that even very young children may have access to the information that language is hierarchically structured.

This work has some limitations that should be addressed with further research. While we showed that a comparison of appropriate grammars of each type results in a preference for the hierarchically structured grammar, these grammars were not the result of an exhaustive search through the space of all grammars. It is almost certain that better grammars of either type could be found, so any conclusions are preliminary. We have explored several ways to test the robustness of the analysis. First, we conducted a local search using an algorithm inspired by Stolcke and Omohundro (1994), in which a space of grammars is searched via successive merging of states. The results using grammars produced by this search are qualitatively similar to the results shown here. Second, we tried several other regular grammars, and again the hierarchical grammar was preferred. In general, the poor performance of the regular grammars appears to reflect the fact that they fail to maximize the tradeoff between simplicity and generalization. The simpler regular grammars buy that simplicity only at the cost of increasing overgeneralization, resulting in a high penalty in the likelihood.

Are we trying to argue that the knowledge that language is structure-dependent *is not* innate? No. All we have shown is that, *contra* the PoS argument, structure dependence *need not* be a part of innate linguistic knowledge. It is true that the ability to represent PCFGs is "given" to our model, but this is a relatively weak form of innateness: few would argue that children are born without the capacity to represent the thoughts they later grow to have, since if they were no learning would occur. Furthermore, everything that is built into the model – the capacity to represent each grammar as well as the details of the Bayesian inference procedure – is domain-general, not language-specific as the original PoS claim suggests.

In sum, we have demonstrated that a child equipped with both 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. How well this ideal learnability analysis corresponds to the actual learning behavior of children remains an important open question.

Acknowledgments Thanks to Virginia Savova for helpful comments. Supported by an NDSEG fellowship (AP) and the Paul E. Newton Chair (JBT).

References

Brown, R. (1973). *A first language: The early stages*. Harvard University Press.

Chater, N. and Vitanyi, P. (2003). Simplicity: A unifying principle in cognitive science? *TICS*, 7:19–22.

Chomsky, N. (1959). On certain formal properties of grammars. *Information and Control*, 2:137–167.

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

Chomsky, N. (1971). *Problems of Knowledge and Freedom*. Fontana, London.

Chomsky, N. (1980). In Piatelli-Palmarini, M., editor, *Language and learning: The debate between Jean Piaget and Noam Chomsky*. Harvard Univ Press, Cambridge, MA.

Crain, S. and Nakayama, M. (1987). Structure dependence in grammar formation. *Language*, 24:139–186.

Goldwater, S., Griffiths, T., and Johnson, M. (2005). Interpolating between types and tokens by estimating power law generators. *NIPS*, 18.

Jaynes, E. (2003). *Probability theory: The logic of science*. Cambridge University Press, Cambridge.

Legate, J. and Yang, C. (2002). Empirical re-assessment of stimulus poverty arguments. *Ling. Review*, 19:151–162.

Lewis, J. and Elman, J. (2001). Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In *Proc. of the 26th BU Conf. on Lang. Devel.* Cascadilla Press.

Li, M. and Vitanyi, P. (1997). *An Intro. to Kolmogorov complexity and its applications*. Springer Verlag, NY.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Ass., third edition.

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

Reali, F. and Christiansen, M. (2004). Structure dependence in language acquisition: Uncovering the statistical richness of the stimulus. In *Proc. of the 26th conference of the Cognitive Science Society*.

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

Stolcke, A. and Omohundro, S. (1994). Introducing probabilistic grammars by bayesian model merging. *ICGI*.

Zipf, G. (1932). *Selective studies and the principle of relative frequency in language*. Harvard University Press.