

Running head: LEARNABILITY OF SYNTAX

The learnability of abstract syntactic principles

Amy Perfors and Joshua B. Tenenbaum

Department of Brain & Cognitive Science

Massachusetts Institute of Technology

Terry Regier

Department of Psychology

University of Chicago

Word count: 18745

The learnability of abstract syntactic principles

Children acquiring language infer the correct form of syntactic constructions for which they appear to have little or no direct evidence, avoiding simple but incorrect generalizations that would be consistent with the data they receive. These generalizations must be guided by some inductive bias – some abstract knowledge – that leads them to prefer the correct hypotheses even in the absence of directly supporting evidence. What form do these inductive constraints take? It is often argued or assumed that they reflect innately specified knowledge of language. A classic example of such an argument moves from the phenomenon of auxiliary fronting in English interrogatives to the conclusion that children must innately know that syntactic rules are defined over hierarchical phrase structures rather than linear sequences of words (e.g., Chomsky 1965, 1971, 1980; Crain & Nakayama, 1987). Here we use a Bayesian framework for grammar induction to argue for a different possibility. We show that, given typical child-directed speech and certain innate domain-general capacities, an unbiased ideal learner could recognize the hierarchical phrase structure of language without having this knowledge innately specified as part of the language faculty. We discuss the implications of this analysis for accounts of human language acquisition.

Introduction

Nature, or nurture? To what extent is human mental capacity a result of innate domain-specific predispositions, and to what extent does it result from domain-general learning based on data in the environment? One of the tasks of modern cognitive science is to move past this classic nature/nurture dichotomy and elucidate just how innate biases and domain-general learning might interact to guide development in different domains of knowledge.

Scientific inquiry in one domain, language, was influenced by Chomsky's observation that language learners make grammatical generalizations that appear to go beyond what is immediately justified by the evidence in the input (Chomsky, 1965, 1980). One such class of generalizations concerns the hierarchical phrase structure of language: children appear to favor hierarchical rules that operate on grammatical constructs such as phrases and clauses over linear rules that operate only on the sequence of words, even in the apparent absence of direct evidence supporting this preference. Such a preference, in the absence of direct supporting evidence, is used to suggest that human learners innately know a deep organizing principle of natural language, that grammatical rules are defined on hierarchical phrase structures.

In outline form, this is the “Poverty of the Stimulus” (or PoS) argument for innate knowledge. It is a classic move in cognitive science, but in some version this style of reasoning is as old as the Western philosophical tradition. Plato's argument for innate principles of geometry or morality, Leibniz' argument for an innate ability to understand necessary truths, Hume's argument for innate mechanisms of association, and Kant's argument for an innate spatiotemporal ordering of experience are all used to infer the prior existence of certain mental capacities based on an apparent absence of support for acquiring them through learning.

Our goal in this paper is to reevaluate the modern PoS argument for innate language-specific knowledge by formalizing the problem of language acquisition within a Bayesian framework for rational inductive inference. We consider an ideal learner who comes equipped with two powerful but domain-general capacities. First, the learner has the

capacity to represent structured grammars of various forms, including hierarchical phrase-structure grammars (which the generative tradition argues must be innately known to apply in the domain of language) and simpler non-hierarchical alternatives (which the generative tradition claims must be innately inaccessible for language learning, but that might be an appropriate model for sequential data in non-linguistic domains). Second, the learner has access to a Bayesian engine for statistical inference which can operate over these structured grammatical representations and compute their relative probabilities given observed data. We will argue that a certain core aspect of linguistic knowledge – the knowledge that syntactic rules are defined over hierarchically organized phrase structures – can be inferred by a learner with these capabilities but without a language-specific innate bias favoring this conclusion.

Note that this claim about innateness is really a claim about the domain-specificity of innate linguistic knowledge. Because language acquisition presents a problem of induction, it is clear that learners must have some constraints limiting the hypotheses they consider. The question is whether a certain feature of language – such as hierarchical phrase structure in syntax – must be assumed to be specified innately as part of a language-specific “acquisition device”, rather than derived from more general-purpose representational capacities and inductive biases.

We introduce PoS arguments in the context of a specific example that has sparked many discussions of innateness, from Chomsky’s original discussions to present-day debates (Laurence & Margolis, 2001; Lewis & Elman, 2001; Legate & Yang, 2002; Pullum & Scholz, 2002; Real & Christiansen, 2005): the phenomena of auxiliary fronting in constructing English interrogative sentences. We begin by introducing this example and then lay out the abstract logic of the PoS argument of which this example is a special case. This abstract logic will directly motivate the form of our Bayesian analysis. Our analysis should not be seen as an attempt to explain the learnability of auxiliary fronting (or any specific linguistic rule) *per se*. Rather the goal is to address the general phenomenon of hierarchical phrase structure in syntax – the phenomenon that has been argued to underlie the learning of auxiliary fronting and many other specific rules. We take as data an entire corpus of child-directed speech and evaluate hypotheses about candidate grammars that could account for

the corpus as a whole. As a byproduct of this, our analysis allows us to explore the learnability of auxiliary fronting and other related specific aspects of syntax.

Before moving into the argument itself, we should highlight two aspects of our approach that contrast with other recent analyses of PoS arguments in language, and analyses of auxiliary-fronting in particular (Laurence & Margolis, 2001; Lewis & Elman, 2001; Legate & Yang, 2002; Pullum & Scholz, 2002; Reali & Christiansen, 2005). First, as illustrated in Figure 1, our approach offers a way to study how two fundamental questions of linguistic knowledge interact. The question of whether human learners have (innate) language-specific knowledge is logically separable from the question of whether and to what extent human linguistic knowledge is based on structured representations such as a generative phrase-structure grammar. In practice, however, these issues are often conflated. Within cognitive science, recent computational models of how language might be learned have usually assumed that domain-general learning operates on representations without explicit structure (e.g., Elman et. al., 1996; Rumelhart & McClelland, 1986; Reali & Christiansen, 2005). The main proponents of innate language-specific factors, on the other hand, have typically assumed that the representations involved are structured (e.g., Chomsky 1965, 1980; Pinker, 1984). Few cognitive scientists have explored the possibility that explicitly structured mental representations might be learned via domain-general mechanisms. Our framework offers a way to explore this relatively uncharted territory in the context of language acquisition.

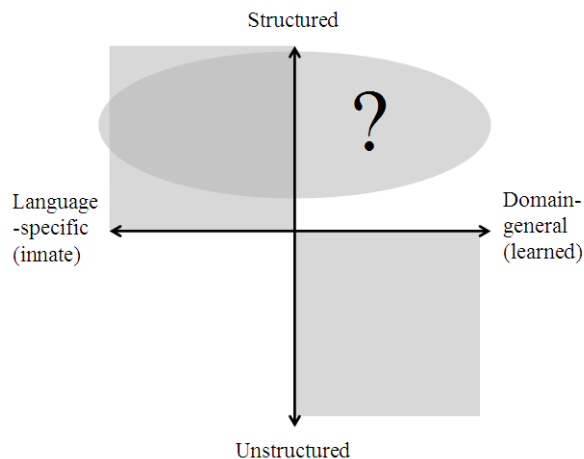


Figure 1: A schematic representation of the theoretical landscape for language acquisition in cognitive science. The vertical axis reflects the nature of the representation. The horizontal axis reflects the source of inductive bias: “innate” and “learned” are in parentheses because they are often conflated with “language-specific” and “domain-general”, which we suggest is closer to the real

issue. The two most prominent views are represented by the two opposite shaded quadrants. We explore another area, represented by the shaded oval. (The oval covers both sides of the horizontal axis because our approach explores both: it is in principle possible that it could yield results suggesting that the particular innate language-specific bias in question is necessary).

Second, while we are somewhat interested in how children master a particular linguistic generalization such as the auxiliary-fronting rule, our real interest is in how and whether children can learn deeper and more abstract principles of linguistic structure, such as the hierarchical phrase-structure basis for syntax. This principle supports an entire class of generalizations that include the auxiliary-fronting rule but also phenomena surrounding agreement, movement, and extraction. Our analysis suggests that it is vital to consider the learnability of all these generalizations as a whole. While the data supporting any one generalization (such as the auxiliary-fronting rule) may be very sparse or even nonexistent, there may be extensive data supporting other, related generalizations; this can bias a rational learner towards making the correct inferences about the cases for which the data is very sparse. To put this point another way, while it may be sensible to ask what a rational learner can infer about language as a whole without any language-specific biases, it is less sensible to ask what a rational learner can infer about any single specific linguistic rule (such as auxiliary-fronting). The need to acquire a whole system of linguistic rules together imposes constraints among the rules, so that an *a priori* unbiased learner may acquire constraints that are based on the other linguistic rules it must learn at the same time.

Auxiliary fronting: a specific PoS argument

At the core of modern linguistics is the insight that sentences, although they might appear to be simply linear sequences of words or sounds, are built up in a hierarchical fashion from nested phrase structures (Chomsky 1965, 1980). The rules of syntax are defined over linguistic elements corresponding to phrases that can be represented hierarchically with respect to one another: for instance, a noun phrase might itself contain a prepositional phrase. By contrast, in a language without hierarchical phrase structure the rules of syntax might make reference only to the individual elements of the sentence as they appear in a linear sequence. Henceforth, when we say that “language has hierarchical phrase structure” we mean, more precisely, that the rules of syntax are defined over hierarchical phrase-structure representations rather than a linear sequence of words. Is the

knowledge that language is organized in this way innate? In other words, is it a part of the initial state of the language acquisition system and a necessary feature of any possible hypothesis that the learner will consider?

Chomsky (1965, 1971, 1980) put forth several arguments for this position, most famously one based on the phenomenon of auxiliary fronting in English. English interrogatives such as “Is the man hungry?” correspond to declaratives with a fronted main clause auxiliary like “The man is hungry”: the auxiliary *is* at the beginning of the interrogative appears to map to the *is* in the middle of the declarative. One might consider two possible rules that could govern this correspondence between declarative and interrogative forms:

- (1a) Linear: Form the interrogative by moving the first occurrence of the auxiliary in the declarative to the beginning of the sentence.
- (1b) Hierarchical: Form the interrogative by moving the auxiliary from the main clause of the declarative to the beginning of the sentence.

The linear rule (1a) can be implemented without reference to the hierarchical phrase structure of the sentence, but the hierarchical rule (1b) cannot. We know that the actual grammar of English follows principles much closer to the hierarchical rule (1b), but how is a child to learn that such a rule is correct as opposed to a linear rule such as (1a)? Although the linear and hierarchical rules result in the same outcome when applied to simple declarative sentences like “The man is hungry”, they yield different results when applied to more complex declaratives such as this:

- (2) The man who is hungry is ordering dinner.

The hierarchical rule predicts the interrogative form in (3a), while the linear rule predicts the form in (3b):

- (3a) Is the man who is hungry ordering dinner?
- (3b) * Is the man who hungry is ordering dinner?

Of course, (3a) is grammatical in English while (3b) is not. This difference could provide a basis for inferring the correct rule: if children learning language hear a sufficient sample of sentences like (3a) and few or no sentences like (3b), they might reasonably infer that the hierarchical rule rather than the linear rule correctly describes the grammar of English. Yet Chomsky argued that complex interrogative sentences such as (3a) do not exist in sufficient quantity in child-directed speech, going so far as to assert 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” (1971). In spite of this paucity of evidence, children three to five years old can form correct complex interrogative sentences like (3a) but appear not to produce incorrect forms such as (3b) (Crain & Nakayama, 1987).

Chomsky further argued that on *a priori* grounds, a general-purpose learning agent who knows nothing specifically about human natural languages would take the linear rule to be more plausible by virtue of its simplicity: it does not assume either the existence of hidden objects (e.g., syntactic phrases) or of a particular organization (e.g., hierarchical rather than linear). If the correct rule cannot be learned from data and is also dispreferred due to a general inductive bias favoring simplicity, the logical conclusion is that children come equipped with some powerful language-specific innate mechanisms that bias them to learn structure-dependent rather than structure-independent syntactic rules.

This version of the PoS argument has been the subject of much debate (Laurence & Margolis, 2001; Lewis & Elman, 2001; Legate & Yang, 2002; Pullum & Scholz, 2002; Reali & Christiansen, 2005). We have considered it here because of its relevance to the learnability of the particular generalization we are concerned with, namely, the generalization that language has hierarchical phrase structure. In the next section we will refer again to the auxiliary fronting example in order to focus on that generalization and to clarify the abstract logical structure of the PoS argument.

A general formulation of Poverty of the Stimulus argument

We formulate the PoS argument more precisely and abstractly as follows:

- (4.i) Children show a specific pattern of behavior B .
- (4.ii) A particular generalization G must be grasped in order to produce behavior B .
- (4.iii) It is impossible to reasonably induce G simply on the basis of the data D that children receive.
- (4.iv) *therefore*, some abstract knowledge T , limiting which specific generalizations G are possible, is necessary.

This form of the PoS argument, also shown schematically in Figure 2, is applicable to a variety of domains and datasets. Unlike other standard treatments (Laurence & Margolis, 2001; Pullum & Scholz, 2002), it makes explicit the distinction between multiple levels of knowledge (G and T); this distinction is necessary to see what is really at stake in arguments about innateness in language and other cognitive domains. In the case of auxiliary fronting, the specific generalization G refers to the hierarchical rule (1b) that governs the formation of interrogative sentences. The learning challenge is to explain how children come to produce only the correct forms for complex interrogatives (B), apparently following a rule like (1b), when the data they observe (D) comprise only simple interrogatives (such as “Is the man hungry?”) that do not discriminate between the correct generalization and simpler but incorrect alternatives such as (1a).

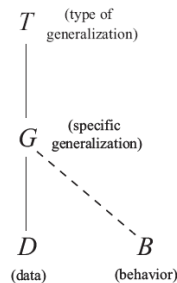


Figure 2. Graphical depiction of the standard Poverty of Stimulus argument. Abstract higher-level knowledge T is necessary to constrain the specific generalizations G that are learned from the data D , and that govern behavior B .

But the interesting claim of innateness here is not about the rule for producing interrogatives (G) *per se*; rather, it concerns some more abstract knowledge T . Note that nothing in the logical structure of the argument requires that T be specific to the domain of language – constraints due to domain-general processing, memory, or learning factors could also limit which generalizations are considered. Nevertheless, many versions of the PoS argument assume that the T is language-specific: in particular, that T is the knowledge that *linguistic* rules are defined over hierarchical phrase structures rather than linear sequences of words. This knowledge constrains the specific rules of grammar that children may posit and therefore licenses the inference to G . Constraints on grammatical generalizations at the level of T may be seen as one aspect of, or as playing the role of, “universal grammar” (Chomsky, 1965).

An advantage of this logical schema is to clarify that the correct conclusion given the premises is *not* that the higher-level knowledge T is innate -- only that it is necessary. The following corollary is required to conclude that T is innate:

- (5.i) (Conclusion from above) Some abstract knowledge T is necessary.
- (5.ii) T could not itself be learned, or could not be learned before the specific generalization G is known.
- (5.iii) *therefore*, T must be innate

Given this schema, our argument here can be construed in two different ways. On one view, we are arguing against premise (5.ii); we suggest that the abstract linguistic knowledge T – that language has hierarchical phrase structure – might be learnable using domain-general mechanisms and representational machinery. Given some observed data D , we evaluate knowledge at both levels (T and G) together by drawing on the methods of hierarchical Bayesian models and Bayesian model selection (Gelman et. al., 2004). Interestingly, our results suggest that less data is required to learn T than to learn the specific grammar G .

On another view, we are not arguing with the form of the PoS argument, but merely clarifying what content the knowledge T must have. We argue that phenomena such as children’s mastery of auxiliary fronting are not sufficient to require that the innate

knowledge constraining generalization in language acquisition be language-specific. Rather it could be based on more general-purpose systems of representation and inductive biases that favor the construction of simpler representations over more complex ones.

Other critiques of the innateness claim dispute the three premises of the original argument, arguing either:

- (6.i) Children do not show the pattern of behavior *B*.
- (6.ii) Behavior *B* is possible without having made the generalization *G*, through some other route from *D*.
- (6.iii) It is possible to learn *G* on the basis of *D* alone, without the need for some more abstract linguistic knowledge *T*.

In the case of auxiliary fronting, one example of the first response (6.i) is the claim that children do not in fact always avoid errors that would be best explained under a structure-independent (linear) rule rather than a structure-dependent (hierarchical) rule. Although Crain & Nakayama (1987) demonstrated that children do not spontaneously form incorrect complex interrogatives such as (c), they make other mistakes that are not so easily interpretable. For instance, one might utter a sentence like “Is the man who is hungry is ordering dinner?”, which is not immediately compatible with the correct hierarchical grammar but might be consistent with a linear rule. Additionally, recent research by Ambridge et. al. (2005) suggests that 6 to 7 year-old children presented with auxiliaries other than *is* do indeed occasionally form incorrect sentences like (c), such as “Can the boy who run fast can jump high?”

A different response (6.iii) accepts that children have inferred the correct hierarchical rule for auxiliary fronting (1b), but maintains that the input data is sufficient to support this inference. If children observe sufficiently many complex interrogative sentences like (3a) while observing no sentences like (3b), then perhaps they could learn directly that the hierarchical rule (1b) is correct, or at least better supported than simple linear alternatives. The force of this response depends on how many sentences like (3a) children actually hear. While it is an exaggeration to say that there are *no* complex interrogatives in typical child-directed speech, they are certainly rare: Legate & Yang (2002) estimate based on two

CHILDES corpora¹ that between 0.045% and 0.068% of all sentences are complex interrogative forms. Is this enough? Unfortunately, in the absence of a specific learning mechanism, it is difficult to develop an objective standard about what would constitute “enough.” Legate & Yang attempt to establish one by comparing how much evidence is needed to learn other generalizations that are acquired at around the same age; they conclude on this basis that the evidence is probably insufficient. However, such a comparison overlooks the role of indirect evidence, which has been suggested to contribute to learning in a variety of other contexts (Landauer & Dumais, 1997; Regier & Gahl, 2004; Reali & Christiansen, 2005).

Indirect evidence also plays a role in the second type of reply, (6.ii), which is probably the most currently popular line of response to the PoS argument. The claim is that children could still show the correct pattern of linguistic behavior – acceptance or production of sentences like (3a) but not (3b) – even without having learned any grammatical rules like (1a) or (1b) at all. Perhaps the data, while poor with respect to complex interrogative forms, are rich in distributional and statistical regularities that would distinguish (3a) from (3b). If children pick up on these regularities, that could be sufficient to explain why they avoid incorrect complex interrogative sentences like (c), without any need to posit the kinds of grammatical rules that others have claimed to be essential (Redington et. al., 1998; Lewis & Elman, 2001; Reali & Christiansen, 2004, 2005).

For instance, Lewis & Elman (2001) trained a simple recurrent network to produce sequences generated by an artificial grammar that contained sentences of the form *AUX NP ADJ?* and *A_i NP B_i*, where *A_i* and *B_i* stand for inputs of random content and length. They found that the trained network predicted sentences like “Is the boy who is smoking hungry?” with higher probability than similar but incorrect sequences, despite never having received that type of sentence as input. In related work, Reali & Christiansen (2005) showed that the statistics of actual child-directed speech support such predictions (though see Kam et. al. (2005) for a response). They demonstrated that simple bigram and trigram models applied to a corpus of child-directed speech gave higher likelihood to correct complex interrogatives formed by structure-dependent fronting than to incorrect interrogatives formed by structure-independent fronting, and that the *n*-gram models correctly classified the grammaticality of

¹ Adam (Brown corpus, 1973) and Nina (Suppes corpus, 1973); for both, see MacWhinney (1995)

96% of test sentences like (3a) and (3b). They also argued that simple recurrent networks could distinguish grammatical from ungrammatical test sentences because they were able to pick up on the implicit statistical regularities between lexical classes in the corpus.

Though these statistical-learning responses to the PoS argument are important and interesting, they have two significant disadvantages. First of all, the behavior of connectionist models tends to be difficult to understand analytically. For instance, the networks used by Reali & Christiansen (2005) and Lewis & Elman (2001) measure success by whether they predict the next word in a sequence or by comparing the prediction error for grammatical and ungrammatical sentences. These networks lack not only a grammar-like representation; they lack any kind of explicitly articulated representation of the knowledge they have learned. It is thus difficult to say what exactly they have learned about linguistic structure.

Second, by denying that explicit structured representations play an important role in children's linguistic knowledge, these statistical-learning models fail to engage with the motivation at the heart of the PoS arguments and most contemporary linguistics. PoS arguments begin with the assumption – taken by most linguists as self-evident – that language does have explicit hierarchical structure, and that linguistic knowledge must at some level be based on representations of syntactic categories and phrases that are hierarchically organized within sentences. The PoS arguments are about whether and to what extent children's knowledge about this structure is learned via domain-general mechanisms, or is innate in some language-specific system. Critiques based on the premise that this explicit structure is not represented as such in the minds of language users do not really address this argument (although they may be valuable in their own right by calling into question the broader assumption that linguistic knowledge is structured and symbolic). Our work here is premised on taking seriously the claim that knowledge of language is based on structured symbolic representations. We can then investigate whether the principle that these linguistic representations are hierarchically organized might be learned. We do not claim that linguistic representations *must* have explicit structure, but assuming such a representation allows us to engage with the PoS argument on its own terms.

Overview

We present two main results. First of all, we demonstrate that a learner equipped with the capacity to explicitly represent both linear and hierarchical grammars – but without any initial bias to prefer either in the domain of language – can infer that the hierarchical grammar is a better fit to typical child-directed input, even on the basis of as little as a few hours of conversation. Our results suggest that at least in this particular case, it is possible to acquire domain-specific knowledge about the form of structured representations via domain-general learning mechanisms operating on type-based data from that domain. Secondly, we show that the hierarchical grammar favored by the model – unlike the other grammars it considers – masters auxiliary fronting, even when no direct evidence to that effect is available in the input data. This second point is simply a by-product of the main result, but it provides a valuable connection to the literature and makes concrete the benefits of learning abstract linguistic principles.

These results emerge because an ideal learner must trade off simplicity and goodness-of-fit in evaluating hypotheses. The notion that inductive learning should be constrained by a preference for simplicity is widely shared among scientists, philosophers of science, and linguists. Chomsky himself concluded that natural language is not finite-state based on informal simplicity considerations (1956, 1957), and suggested that human learners rely on an evaluation procedure that incorporates simplicity constraints (1965). The tradeoff between simplicity and goodness-of-fit can be understood in domain-general terms. Consider the hypothetical data set illustrated in Figure 3. We imagine that data is generated by processes occupying different subsets of space. Models correspond to different theories about which subset of space the data is drawn from; three are shown in A, B, and C. These models fit the data increasingly precisely, but they attain this precision at the cost of additional complexity. Intuitively, the model in B appears to offer the optimal balance, and this intuition can be formalized mathematically using techniques sometimes known as the Bayesian Occam's Razor (e.g., MacKay, 2003). In a similar way, we will argue, a hierarchical phrase-structure grammar yields a better tradeoff than linear grammars between simplicity of the grammar and fit to typical child-directed speech.

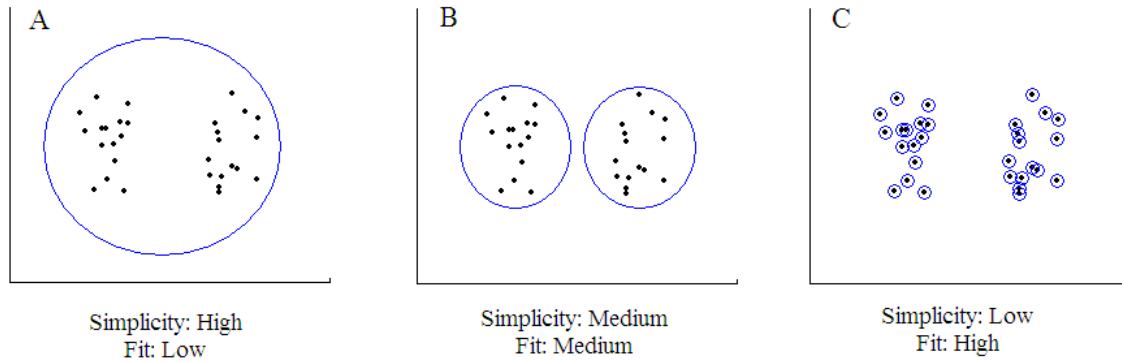


Figure 3. Fitting models of different complexity (represented by the circles) to a dataset (the points). The complexity of a model reflects the number of choices necessary to specify it: the model A can be fully specified by the location and size of only one circle, while model C is more complex because it requires specification of locations and sizes for thirty distinct circles. Model A achieves high simplicity at the cost of poor fit, while C fits extremely closely at the cost of high complexity. The best functional description of the data should optimize a tradeoff between complexity and fit, as shown in B.

Though our findings suggest that the specific feature of hierarchical structure can be learned without an innate language-specific bias, we do not argue or believe that all interesting aspects of language will have this characteristic. Because our approach combines structured representation and statistical inductive inference, it provides a method to investigate the unexplored regions of Figure 1 for a wide range of other linguistic phenomena, as has recently been studied in other domains (e.g., Griffiths et. al., 2004; Kemp et. al. 2004; Yuille & Kersten, 2006).

One finding of our work is that it may require less data to learn a higher-order principle T – such as the hierarchical nature of linguistic rules – than to learn every correct generalization G at a lower level, e.g., every specific rule of English. Though our model does not explicitly use inferences about the higher-order knowledge T to constrain inferences about specific generalizations G , in theory T could provide effective and early-available constraints on G , even if T is not itself innately specified. In the discussion, we will consider what drives this perhaps counterintuitive result and discuss its implications for language acquisition and cognitive development more generally.

Method

We cast the problem of grammar induction within a hierarchical Bayesian framework² whose structure is shown in Figure 4. The goal of the model is to infer from some data D (a corpus of child-directed language) both the specific grammar G that generated the data as well as the higher-level generalization about the type of grammar T that G is an instance of. This is formalized as an instance of Bayesian model selection.

Our framework assumes a multi-stage probabilistic generative model for linguistic utterances, which can then be inverted by a Bayesian learner to infer aspects of the generating grammar from the language data observed. A linguistic corpus is generated by first picking a type of grammar T from the prior distribution $p(T)$. A specific grammar G is then chosen as an instance of that type, by drawing from the conditional probability distribution $p(G|T)$. Finally, a corpus of data D is generated from the specific grammar G , drawing from the conditional distribution $p(D|G)$. The inferences we can make from the observed data D to the specific grammar G and grammar type T are captured by the joint posterior probability $p(G, T|D)$, computed via Bayes' rule:

$$p(G, T|D) \propto p(D|G)p(G|T)p(T). \quad (1)$$

We wish to explore learning when there is no innate bias towards grammars with hierarchical phrase structure. This is implemented in our model by assigning $p(T)$ to be equal for each type T . The prior for a specific grammar $p(G|T)$ is calculated assuming a generative model of grammars that assigns higher prior probability to simpler grammars. The likelihood $p(D|G)$ reflects the probability of the corpus of child-directed speech D given G and T ; it is a measure of how well the grammar fits the corpus data. The Bayesian approach to inferring grammatical structure from data, in the form of the posterior $p(G, T|D)$, thus automatically seeks a grammar that balances the tradeoff between complexity (prior probability) and fit to the data (likelihood).

² Note that the “hierarchical” of “hierarchical Bayesian framework” is not the same “hierarchical” as in “hierarchical phrase structure.” The latter refers to the hierarchical embedding of linguistic phrases within one another in sentences. The former refers to a Bayesian model capable of performing inference at multiple levels, in which not only the model parameters but also the hyperparameters (parameters controlling priors over the parameters) are inferred from the data, rather than being set by the modeler.

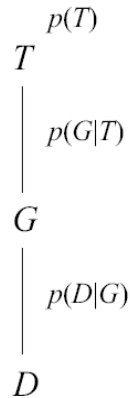


Figure 4. A hierarchical Bayesian model for assessing Poverty of Stimulus arguments. The model is organized around the same structure as Figure 2, but now each level of representation defines a probability distribution for the level below it. Bayesian inference can be used to make inferences at higher levels from observations at lower levels. Abstract principles of the grammar T constrain the specific grammatical generalizations G a learner will consider by defining a conditional probability distribution $p(G|T)$. These generalizations in turn define probabilistic expectations about the data D to be observed, $p(D|G)$. Innate language-specific biases for particular types of grammars can be encoded in the prior $p(T)$, although here we consider an unbiased prior, with $p(T)$ equal for all T .

Relation to previous work

Probabilistic approaches to grammar induction have a long history in linguistics. One strand of work concentrates on issues of learnability (e.g., Solomonoff, 1964, 1978; Horning, 1969; Li & Vitanyi, 1997; Chater & Vitanyi, 2003, 2007). This work is close to ours in intent, because much of it is framed in response to the negative learnability claims of Gold (1967), and demonstrates that, *contra* Gold, learning a grammar in a probabilistic sense is possible if the learner is sensitive to the statistical distribution of the input sentences (Horning, 1969). Part of the power of the Bayesian approach derives from its incorporation of a simplicity metric: an ideal learner with such a metric will be able to predict the sentences of the language with an error that approaches zero as the size of the corpus goes to infinity (Solomonoff, 1978), suggesting that learning from positive evidence alone may be possible (Chater & Vitanyi, 2007). Our analysis is complementary to these previous Bayesian analyses. The main difference is that instead of addressing learnability issues in abstract and highly simplified settings, we focus on a specific question – the learnability of hierarchical structure in syntax – and evaluate it on realistic data: a finite corpus of child-directed speech. As with the input data that any child observes, this corpus contains only a

small fraction of the syntactic forms in the language, and probably a biased and noisy sample at that.

Another strand of related work is focused on computational approaches to learning problems (e.g., Eisner, 2002; Johnson & Riezler, 2002; Light & Grieff, 2002; Klein & Manning, 2004; Alishahi & Stevenson, 2005; Chater & Manning, 2006). Our analysis is distinct in several ways. First, many approaches focus on the problem of learning a grammar given built-in constraints T , rather than on making inferences about the nature of T as well. For instance, Klein & Manning (2004) have explored unsupervised learning for a simple class of hierarchical phrase-structure grammars (dependency grammars) from natural corpora. They assume that this class of hierarchical grammars is fixed for the learner rather than considering the possibility that grammars in other classes, such as linear grammars, could be learned.

A more important difference in our analysis lies in the nature of our corpora. Other work incorporates either on small fragments of (sometimes artificial) corpora (e.g., Dowman, 2000; Alishahi & Stevenson, 2005; Clark & Eyraud, 2006) or on large corpora of adult-directed speech (e.g., Eisner, 2002; Klein & Manning, 2004). Neither is ideal for addressing learnability questions. Large corpora of adult-directed speech are more complex than child-directed speech, and do not have the sparse-data problem assumed to be faced by children. Analyses based on small fragments of a corpus can be misleading: the simplest explanation for limited subsets of a language may not be the simplest within the context of the entire system of linguistic knowledge the child must learn.

An ideal analysis of learnability

Our analysis views learnability in terms of an ideal framework in which the learner is assumed to be able to effectively search over the joint space of G and T for grammars that maximize the Bayesian scoring criterion. We are not proposing a comprehensive or mechanistic account of how children actually acquire language. The full problem of language acquisition poses many challenges that we do not consider here. Rather, our analysis provides a formal framework for analyzing the learnability of some aspects of linguistic syntax, with the goal of clarifying and exploring claims about what language-specific prior knowledge must be assumed in order to make learning possible. The key

component of this analysis is an evaluation metric – a means for the ideal learner to evaluate one G, T pair against another. We assume that an ideal learner is more likely to learn a given G, T pair than an alternative G', T' if the former has a higher posterior probability than the latter. This analysis leaves out significant algorithmic questions of how learners search the space of grammars, but this idealization is valuable in the same spirit as Marr's computational-theory level analyses in vision (Marr, 1982). It allows us to examine rigorously the inductive logic of learning – what constraints are necessary given the structure of the hypothesis space and the data available to learners – independent of the specifics of the algorithms used to search these hypothesis spaces. This formal approach also follows the spirit of how Chomsky and other linguists have considered learnability, as a question of what is learnable *in principle*: is it in principle possible given the data a child observes to learn that language is governed by hierarchical phrase-structure rules, rather than linear rules, if one is not innately biased to consider only hierarchical grammars? If we can show that such learning is in principle possible, then it becomes meaningful to ask the algorithmic-level question of how a system might successfully and in reasonable time search the space of possible grammars to discover the best-scoring grammar.

Of course, the value of this ideal learnability analysis depends on whether the specific grammars we consider are representative of the best hypotheses that can be found in the full spaces of different grammar types we are interested in (the spaces of hierarchical phrase-structure grammars, linear grammars, and so on). We therefore examine grammars generated in a variety of ways:

- (1) The best hand-designed grammar of each grammar type.
- (2) The best grammars resulting from local search, using the grammar from (1) as the starting point.
- (3) The best grammars found in a completely automated fashion.

Because we restrict our analysis to grammars that can successfully parse our corpora, we will explain the corpora before moving on to a more detailed description of the process of inference and search and finally the grammars.

The corpora

The corpus consists of the sentences spoken by adults in the Adam corpus (Brown, 1973) of the CHILDES database (MacWhinney, 2000). In order to focus on grammar learning rather than lexical acquisition, each word is replaced by its syntactic category.³ Although learning a grammar and learning a lexicon are probably tightly linked, we believe that this is a sensible starting assumption for several reasons: first, because grammars are defined over these syntactic categories, and second, because there is some evidence that aspects of syntactic-category knowledge may be in place even in very young children (Booth & Waxman, 2003; Gerken et. al., 2005). In addition, ungrammatical sentences and the most grammatically complex sentence types are removed from the corpus.⁴ The complicated sentence types are removed for reasons of computational tractability as well as the difficulty involved in designing grammars for them, but this is if anything a conservative move since our results suggest that the hierarchical, structure-dependent grammars will be more preferred as the input grows more complex. The final corpus contains 21671 individual sentence tokens corresponding to 2336 unique sentence types, out of 25755 tokens in the original corpus.⁵

In order to explore how the preference for a grammar depends on the amount of data available to the learner, we create six smaller corpora as subsets of the main corpus. Under the reasoning that the most frequent sentences are most available as evidence and are therefore the most likely to be understood, different corpus *Levels* contain only those sentence forms whose tokens occur with a certain frequency or higher in the full corpus. The levels are: *Level 1* (contains all forms occurring 500 or more times, corresponding to 8

³ Parts of speech used included determiners (*det*), nouns (*n*), adjectives (*adj*), comments like “mmhm” (*c*), prepositions (*prep*), pronouns (*pro*), proper nouns (*prop*), infinitives (*to*), participles (*part*), infinitive verbs (*vinf*), conjugated verbs (*v*), auxiliaries (*aux*), complementizers (*comp*), and wh-question words (*wh*). Adverbs and negations were removed from all sentences. Additionally, whenever the word *what* occurred in place of another syntactic category (as in a sentence like “He liked what?”) the original syntactic category was used; this was necessary in order to simplify the analysis of all grammar types, and was only done when the syntactic category was obvious from the sentence.

⁴ Removed types included topicalized sentences (66 individual utterances), sentences containing subordinate phrases (845), sentential complements (1636), conjunctions (634), serial verb constructions (460), and ungrammatical sentences (443).

⁵ 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 replicated for the corpus with those sentences removed. There was no qualitative change in the findings.

unique types); *Level 2* (100 times, 37 types); *Level 3* (50 times, 67 types); *Level 4* (10 times, 268 types); *Level 5* (5 times, 465 types); and the complete corpus, *Level 6*, with 2336 unique types, including interrogatives, wh-questions, relative clauses, prepositional and adjective phrases, command forms, and auxiliary and non-auxiliary verbs. The larger corpora include the rarer and more complex forms, and thus levels roughly correspond to complexity as well as quantity of data.⁶

An additional variable of interest is what evidence is available to the child at different ages. We approximate this by splitting the corpora into five equal sizes by age. The Adam corpus has 55 files, so we define the earliest (*Epoch 1*) corpus as the first 11 files. The *Epoch 2* corpus corresponds to the cumulative input from the first 22 files, *Epoch 3* the first 33, *Epoch 4* the first 44, and *Epoch 5* the full corpus. Splitting the corpus in this way is not meant to reflect the data that children necessarily *use* at each age, but it does reflect the sort of data that is available.

The hypothesis space of grammars and grammar types

Because this work is motivated by the distinction between hierarchical and linear rules, we wish to compare grammar types T that differ from each other structurally in the same way. Different Bayesian approaches to evaluating alternative grammar types are possible. In particular, we could score a grammar type T by integrating the posterior probability over all specific grammars G of that type ($\sum_G p(T, G|D)$) or by choosing the best G of that type ($\max_G p(T, G|D)$). Integrating over all grammars is computationally intractable, and arguably also less relevant. Ultimately it is the specific grammar G that governs how the learner understands and produces language, so we should be interested in finding the best pair of T and G jointly. We therefore compare grammar types by comparing the probability of the best specific grammars G of each type.

There are various formal frameworks we could use to represent hierarchical or linear grammars as probabilistic generative systems. Each of these grammars consists of a set of production rules, specifying how one non-terminal symbol (the left-hand side of the rule) in a string may be rewritten in terms of other symbols, terminal or non-terminal. These grammars can all be defined probabilistically: each production is associated with a

⁶ The mean sentence length of *Level 1* forms is 1.6 words; the mean sentence length at *Level 6* is 6.6.

probability, such that the probabilities of all productions with the same left-hand sides add to one and the probability of a complete parse is the product of the probabilities of the productions involved in the derivation.

To represent hierarchical systems of syntax, we choose context-free grammars (CFGs). Context-free grammars are arguably the simplest approach to capturing the phrase structure of natural language in a way that deals naturally with hierarchy and recursion. They have been treated as a first approximation to the structure of natural language since the early period of generative linguistics (Chomsky, 1959). Probabilistic context-free grammars (PCFGs) are a probabilistic generalization of CFGs commonly used in statistical natural language processing (Manning & Shütze, 1999; Jurafsky & Martin, 2000), and we incorporate powerful tools for statistical learning and inference with PCFGs in our work here. We recognize that there are also many aspects of syntax that cannot be captured naturally in CFGs. In particular, they do not express the sort of movement rules that underlie some standard generative accounts of aux-fronting: they do not represent the interrogative form of a sentence as a transformed version of a simpler declarative form. We work with CFGs because they are the simplest and most tractable formalism suitable for our purposes here – assessing the learnability of hierarchical phrase structure in syntax – but in future work it would be valuable to extend our analyses to richer syntactic formalisms

We consider three different approaches for representing linear grammars. The first is based on regular grammars, also known as finite-state grammars. Regular grammars were originally proposed by Chomsky as a “minimal linguistic theory”, a kind of “null hypothesis” for syntactic theory. They are finite models capable of producing a potentially infinite set of strings, but in a manner that is sensitive only to the linear order of words and not the hierarchical structure of syntactic phrases. A second approach, which we call the FLAT grammar, is simply a memorized list of each of the sentence types (sequences of syntactic categories) that occur in the corpus (2336 productions, zero non-terminals aside from *S*). This grammar will maximize goodness-of-fit to the data at the cost of great complexity. Finally, we consider the one-state (1-ST) grammar, which maximizes simplicity by sacrificing goodness-of-fit. It permits any syntactic category to follow any other and is equivalent to a finite automaton with one state in which all transitions are possible. Though these three approaches may not capture exactly what was originally envisioned as linear

grammars, we work with them because they are representative of simple syntactic systems that can be defined over a linear sequence of words rather than the hierarchical structure of phrases, and they are all easily defined in probabilistic terms.

Hand-designed grammars.

The first method for generating the specific grammars for each type is to design by hand the best grammar possible. The flat grammar and the one-state grammar exist on the extreme opposite ends of the simplicity/goodness-of-fit spectrum: the flat grammar, as a list of memorized sentences, offers the highest possible fit to the data (exact) and the lowest possible compression (none), while the one-state grammar offers the opposite. We design both context-free and regular grammars that span the range between these two extremes (much as the models in Figure 3 do); within each type, specific grammars differ systematically in how they capture the tradeoff between simplicity and goodness-of-fit. Table 1 contains sample productions from each of the specific grammars.⁷

We consider two specific probabilistic context-free grammars in this analysis. The smaller grammar, CFG-S, can parse all of the forms in the full corpus and is based on standard syntactic categories (e.g., noun, verb, and prepositional phrases). The full CFG-S, used for the *Level 6* corpus, contains 14 non-terminal categories and 69 productions. All grammars for other corpus levels and epochs include only the subset of productions and items necessary to parse that corpus.

CFG-L is a larger grammar (14 non-terminals, 120 productions) that fits the data more precisely but at the cost of increased complexity. It is identical to CFG-S except that it contains additional productions corresponding to different expansions of the same non-terminal. For instance, because a sentence-initial V_{inf} may have a different statistical distribution over its arguments than the same V_{inf} occurring after an auxiliary, CFG-L contains both $[V_{\text{inf}} \rightarrow V_{\text{inf}} \text{ PP}]$ and $[V_{\text{inf}} \rightarrow v_i \text{ PP}]$ whereas CFG-S includes the former only. Because of its additional expansions, CFG-L places less probability mass on the recursive productions, which fits the data more precisely. Both grammars have approximately the same expressive power, but balance the tradeoff between simplicity and goodness-of-fit in different ways.

⁷ All full grammars may be found at <http://www.mit.edu/~perfors/posgrammars.html>

We consider three regular grammars spanning the range of the simplicity/goodness-of-fit tradeoff just as the context-free grammars do. All three fall successively between the extremes represented by the flat and one-state grammars, and are created from CFG-S by converting all productions not already of the form $[A \rightarrow a]$ or $[A \rightarrow a B]$ to one of these forms. (It turns out that there is no difference between converting from CFG-S or CFG-L; the same regular grammar is created in any case. This is because the process of converting a production like $[A \rightarrow B C]$ is equivalent to replacing B by all of its expansions, and CFG-L corresponds to CFG-S with some B items replaced.) When possible without loss of generalizability, the resulting productions are simplified and any productions not used to parse the corpus are eliminated.

The “narrowest” regular grammar, REG-N, offers the tightest fit to the data of the three we consider. It has 85 non-terminals and 389 productions, some examples of which are shown in Table 1. The number of productions is greater than in either context-free grammar because it is created by expanding each context-free production containing two non-terminals in a row into a series of distinct productions (e.g. $[NP \rightarrow NP PP]$ expands to $[NP \rightarrow \text{pro PP}]$, $[NP \rightarrow n PP]$, etc). REG-N is thus more complex than either context-free grammar, but it provides a much closer fit to the data -- more like the flat grammar than the one-state.

Just as CFG-S might result from collapsing different expansions in CFG-L into a single production, simpler regular grammars can be created by merging multiple productions in REG-N together. For instance, merging NP_{CP} and NP_{PP} into a single non-terminal such as NP results in a grammar with fewer productions and non-terminals than REG-N. Performing multiple merges of this sort results in a “moderately complex” regular grammar (REG-M) with 169 productions and 13 non-terminals. Because regular grammars are less expressive than context-free grammars, REG-M still requires more productions than either context-free grammar, but it is much simpler than REG-N. In theory, we can continue merging non-terminals to create successively simpler grammars that fit the corpus increasingly poorly until we reach the one-state grammar, which has no non-terminals aside from S. A third, “broader” regular grammar, REG-B, is the best performing of several grammars created in this way from REG-M. It has 10 non-terminals and 117 productions and is identical to REG-

M except that non-terminals NP, AP, PP, and T – which occur in similar contexts as arguments of verbs – are merged to form a new non-terminal HP.

Grammars constructed by automated search.

There is reason to believe that hand-designed grammars provide a good approximation of the best grammar of each type. Both context-free grammars are designed based on linguistic intuition, and the regular grammars are constructed from the context-free grammars in order to preserve as much linguistic structure as possible. Furthermore, grammars of all types have been chosen to reflect the range of tradeoffs between simplicity and goodness-of-fit. It would nevertheless be ideal to search the space of possible grammars and compare the resulting best grammars found for each type, rather than simply comparing the best hand-designed grammars. Unfortunately, this type of search for context-free grammars presents a difficult computational problem, and current search algorithms cannot be relied upon to find the optimal grammar of any given type on large-scale corpora. Fortunately, our argument requires only a search over regular grammars: if our hand-designed context-free grammars are not close to optimal but still have higher probability than the best regular grammars, then the argument is reasonable, but the converse is not true. We perform a fully-automated search of the space of regular grammars by applying an unsupervised algorithm for learning a trigram Hidden Markov Model (HMM) to our corpora (Goldwater & Griffiths, 2007). Though the algorithm was originally developed for learning parts of speech from a corpus of words, it applies to the acquisition of a regular grammar from a corpus of syntactic categories because the formal description of both problems is similar. In both cases, one must identify the hidden variables (parts of speech vs. non-terminals) that best explain the observed data (a corpus of words vs. a corpus of syntactic categories), assuming that the variables depend only on the previous sequence of variables and not on any additional structure. The output of the algorithm is the assignment of each syntactic category in each sentence to the non-terminal that immediately dominates it; this corresponds straightforwardly to a regular grammar containing those non-terminals and no others.⁸

⁸ It is not assumed that each syntactic category has one corresponding non-terminal, or vice versa; both may be ambiguous. Though the algorithm incorporates a prior that favors fewer hidden variables (non-terminals), it

As another comparison, we also perform a partial search over both regular and context-free grammars using the best hand-designed grammar of that type as a starting point. Our partial search was inspired by the work of Stolcke & Omohundro (1994), in which a space of grammars is searched via successive merging of states. States (productions) that are redundant or overly specific are replaced with productions that are not. For more details, see Appendix A.

The probabilistic model

Inferences are calculated using Bayes' rule, which combines the prior probability of G and T with the likelihood that the corpus D was generated by that G and T .

Scoring the grammars: prior probability.

The prior probability of a grammar reflects its complexity. We formalize it using a generative model under which each grammar is selected from the space of all grammars of that type. More complex grammars are those that result from more (and more specific) choices. This method of scoring simplicity is quite general, not restricted to grammars or even language. For instance, the more complex models in Figure 3 are those that require more free parameters to specify. The only parameters for model A are the location and size of one circle, and therefore it is necessary to make only two choices – to set the value of two parameters – in order to precisely specify it. By contrast, the model in B requires two sets of those choices, one for each circle, and therefore twice as many parameters must be set. More choice means more complexity, so C is more complex still.

The simplicity of a probabilistic grammar G is reflected in an analogous way in its prior probability. If one were generating a grammar from scratch, one would have to make the series of choices depicted in Figure 5, beginning with choosing the grammar type: one-state, flat, regular, or context-free. (Since the model is unbiased, the prior probability of each of these is identical). One would then need to choose the number of non-terminals n , and for each non-terminal k to generate P_k productions. These P_k productions, which share a left-

requires the modeler to specify the maximum number of non-terminals considered. We therefore tested all possibilities between 1 and 25. This range was chosen because it includes the number of non-terminals of the best grammars (CFG-L: 21, CFG-S: 21, REG-B: 16, REG-M: 16, REG-N: 86). Since the model is stochastic, we also repeated each run three times, with $N=10000$ iterations each time. The grammars with the highest posterior probability at each level are reported; they have between one and 20 non-terminals.

hand side, are assigned a vector of positive, real-valued production-probability parameters θ_k . Because the productions P_k represent an exhaustive and mutually exclusive set of alternative ways to expand non-terminal k , their parameters θ_k must sum to one. Each production i has N_i right-hand side items, and each of those items must be drawn from the grammar's vocabulary V (set of non-terminals and terminals). If we assume that each right-hand side item of each production is chosen uniformly at random from V , the prior probability is given by:

$$p(G|T) = p(n) \prod_{k=1}^n p(P_k) p(\theta_k) \prod_{i=1}^{P_k} p(N_i) \prod_{j=1}^{N_i} \frac{1}{|V|}. \quad (2)$$

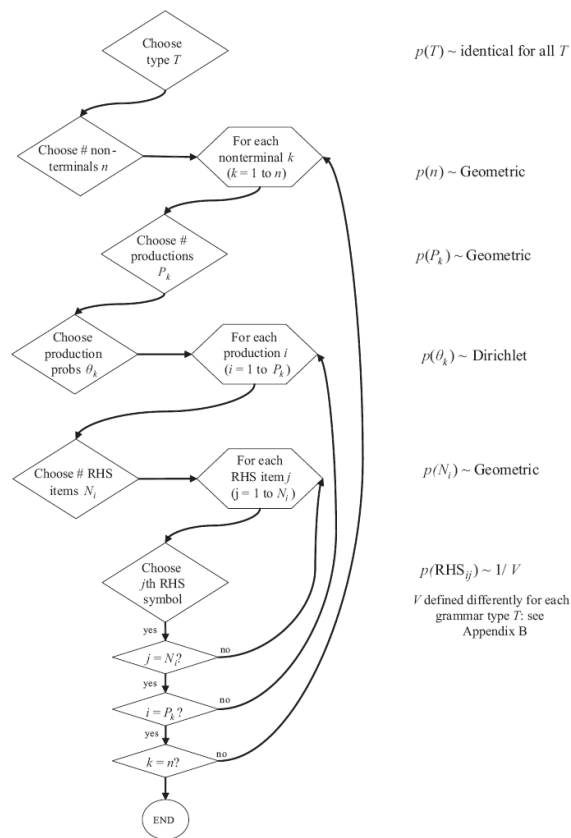


Figure 5. Flowchart depicting the series of choices required to generate a grammar. More subtle differences between grammar types are discussed in Appendix B.

We model the probabilities of the number of non-terminals $p(n)$, productions $p(P_k)$, and items $p(N_i)$ as selections from a geometric distribution; production-probability parameters θ_k are sampled from a discrete approximation of a uniform distribution

appropriate for probability parameters (Dirichlet). This prior gives higher probability to simpler grammars – those with few non-terminals, productions, and items. Because of the small numbers involved, all calculations are done in the log domain. Appendix B contains further details.

The subsets of grammars that can be generated by the several grammar types we consider are not mutually exclusive. A particular grammar – that is, a particular vocabulary and set of productions – might be generated under more than grammar type and would receive different prior probabilities under different grammar types. In general, a grammar with a certain number of productions, each of a certain size, has the highest prior probability if it can be generated as a one-state or flat grammar, next as a regular grammar, and the lowest as a context-free grammar. This follows from the Bayesian Occam's razor that we illustrated with the example in Figure 3. One-state and flat grammars are a subset of regular grammars, which are a subset of context-free grammars (see Figure 6). All other things being equal, one has to make fewer “choices” in order to generate a specific regular grammar from the class containing only regular grammars than from the class of context-free grammars. However, because regular and flat grammars are less expressive, relatively more complex grammars of those types may be required in order to parse all sentences in larger corpora.

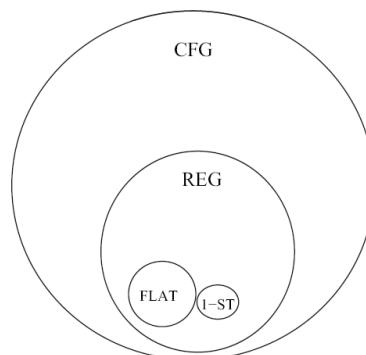


Figure 6. Venn diagram depicting the relation of the grammar types T to each other. The set of context-free grammars contains regular, flat, and one-state grammar types as special cases. Flat and one-state grammars are themselves special cases of regular grammars.

This preference for the simplest grammar type is related to the Bayesian Occam's razor (MacKay, 2003). Other ways to measure simplicity could be based on notions such as minimum description length or Kolmogorov complexity (Li & Vitányi, 1997; Chater &

Vitányi 2003, 2007). These have been useful for the induction of specific context-free grammars G (e.g., Dowman, 1998), and reflect a similar intuitive idea of simplicity.

Scoring the grammars: likelihood.

Inspired by the recent work of 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 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*. The probabilistic grammar is only directly involved in generating the allowable types.

In more psychological terms, this model corresponds to assuming that language users can generate the syntactic forms of sentence tokens either by drawing on a memory store of familiar syntactic types, or by consulting a deeper level of grammatical knowledge about how to generate all and only the legal syntactic forms in the language. Any sentence type generated by the former system would originally have been generated from the latter, but speakers need not consult their deep grammatical knowledge for every sentence token they utter or comprehend. It is the deeper grammatical knowledge used to generate the set of sentence types – not the memory-based generation process that produces the observed frequency of sentence tokens – that we are interested in here.

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 types of sentence forms rather than their frequency distribution. Since we are concerned with grammar comparison rather than corpus generation, we focus in this work on the first component of the model. We thus take the data to consist of the set of sentence types (distinct sequences of syntactic categories) that appear in the corpus, and we evaluate the likelihoods of candidate probabilistic grammars on that dataset.

The likelihood assigned to a grammar based on a dataset of sentence types can be interpreted as a measure of how well the grammar fits or predicts the data. Like simplicity, this notion of “fit” may be understood in intuitive terms that have nothing specifically to do with grammars or language. Consider again Figure 3: intuitively it seems as if model B is more likely to be the source of the data than model A – but why? If A were the correct model, it would be quite a coincidence that all of the data points fall only in the regions covered by B. Likelihood is dependent on the quantity of data observed: it would not be much of a coincidence to see just one or a few data points inside B's region if they were in fact generated by A, but seeing 1000 data points all clustered there – and none anywhere else – would be very surprising if A were correct.

This probabilistic preference for the most specific grammar consistent with the observed data is a version of the size principle in Bayesian models of concept learning and word learning (Tenenbaum & Griffiths, 2001; Regier & Gahl, 2004; Xu & Tenenbaum, 2007). It can also be seen as a probabilistic version of the subset principle (Wexler & Culicover, 1980; Berwick, 1986), a classic heuristic for avoiding the *subset problem* in language acquisition. Many natural hypothesis spaces for grammar induction contain hypotheses which are strictly less general than other hypothesis: that is, they generate languages that are strict subsets of those generated by other hypotheses. If we consider a learner who sees only positive examples of the target grammar, who posits a single hypothesis at any one time and who learns only from errors (sentences which the current hypothesis fails to parse), then if the learner ever posits a hypothesis which generates a superset of the true language, that mistake will never be rectified and the learner will not acquire the correct grammar. The subset principle avoids this problem by mandating that the learner posit only the most restrictive of all possible hypotheses. The Bayesian version becomes equivalent to the subset principle as the size of the dataset approaches infinity because the weight of the likelihood grows with the data while the weight of the prior remains fixed. With limited amounts of data, the Bayesian approach can make different and more subtle predictions, as the graded size-based likelihood trades off against the preference for simplicity in the prior. The likelihood in Bayesian learning can thus be seen as a principled quantitative measure of the weight of implicit negative evidence.

The effective size of the set of sentences that our probabilistic grammars can produce depends on several factors. All other things being equal, a grammar with more productions will produce more distinct sentence types. But the size of the language generated also depends on how those productions relate to each other: how many have the same left-hand side (and thus how much flexibility there is in expanding any one non-terminal), whether the productions can be combined recursively, and other subtle factors. The penalty for overly general or flexible grammars is computed in the parsing process, where we consider all possible ways of generating a sentence under a given grammar and assign probabilities to each derivation. The total probability that a grammar assigns over all possible sentences (really, all possible parses of all possible sentences) must sum to one, and so the more flexible the grammar, the lower probability it will tend to assign to any one sentence.

More formally, the likelihood $p(D|G)$ measures the probability that the corpus data D would be generated by the grammar G . This is given by the product of the likelihoods of each sentence S_l in the corpus, assuming that each sentence is generated independently from the grammar. If there are M unique sentence types in the corpus, the corpus likelihood is given by:

$$p(D|G) = \prod_{l=1}^M p(S_l|G). \quad (3)$$

The probability of any sentence type S_l given the grammar ($p(S_l|G)$) is the product of the probabilities of the productions used to derive S_l . Thus, calculating likelihood involves solving a joint parsing and parameter estimation problem: identifying the possible parse for each sentence in the corpus, as well as calculating the parameters for the production probabilities in the grammar. We use the inside-outside algorithm to integrate over all possible parses and find the set of production probability parameters that maximize the likelihood of the grammar on the observed data (Manning & Schütze, 1999; Johnson, 2006). We evaluate Equation 3 in the same way, using the maximum-likelihood parameter values but integrating over all possible parses of the corpus.⁹ Sentences with longer derivations will tend to be less probable, because each production used contributes a factor less than one

⁹ See Appendix B for a discussion of the subtleties involved in this calculation. One might calculate likelihood under other assumptions, including (for instance) the assumption that all productions with the same left-hand side have the same probability ($g=1$; see Appendix B). Doing so results in lower likelihoods but qualitatively identical outcomes in all cases.

to the product in Equation 3. This notion of simplicity in derivation captures an inductive bias favoring grammars that assign the observed sentences more economical derivations – a bias that is distinct and complementary to that illustrated in Figure 3, which favors grammars generating smaller languages that more tightly cover the observed sentences.

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.

Posterior probability on different grammar types

Hand-designed grammars.

Table 2 shows the prior, likelihood, and posterior probability of each handpicked grammar on each corpus. When there is the least evidence in the input (corpus *Level 1*), the flat grammar is preferred. As the evidence accumulates, the one-state grammar scores higher. However, for the larger corpora (*Level 4* and higher), a hierarchical grammar always scores the highest, more highly than any linear grammar.

If linear grammars are *a priori* simpler than context-free grammars, why does the prior probability favor context-free grammars on more complex corpora? Recall that we considered only grammars that could parse all of the data. Though regular and flat grammars are indeed simpler than equivalently large context-free grammars, linear grammars also have less expressivity: they have to use more productions to parse the same corpus with the same fit. With a large enough dataset, the amount of compression offered by the context-free grammar is sufficient to overwhelm the initial simplicity preference towards the others. This is evident by comparing the size of each grammar for the smallest and largest corpora. On the *Level 1* corpus, the context-free grammars require more productions than do the linear grammars (17 productions for CFG-S; 20 for CFG-L; 17 for REG-N; 15 for REG-M; 14 for REG-B; 10 for 1-ST; 8 for FLAT). Thus, the hierarchical grammars have the lowest initial prior probability. However, their generalization ability is

sufficiently great that additions to the corpus require relatively few additional productions: the context-free grammars that can parse the *Level 6* corpus have 69 and 120 productions, in comparison to 117 (REG-B), 169 (REG-M), 389 (REG-N), 25 (1-ST), and 2336 (FLAT).

The flat grammar has the highest likelihood on all corpora because, as a perfectly memorized list of each of the sentence types, it does not generalize beyond the data at all. The regular grammar REG-N has a relatively high likelihood because its many productions capture the details of the corpus quite closely. The other regular grammars and the context-free grammars have lower likelihoods because they generalize more beyond the data; these grammars predict sentence types which have not (yet) been observed, and thus they have less probability mass available to predict the sentences that have in fact been observed. Grammars with recursive productions are especially penalized in likelihood scores based on finite input. A recursive grammar will generate an infinite set of sentences that do not exist in any finite corpus, and some of the probability mass is allocated to those sentences (although longer sentences with greater depth of recursion are given exponentially lower probabilities). The one-state grammar has the lowest possible likelihood because it accepts any sequence of symbols as grammatical.

As the amount of data accumulates, the posterior increasingly favors the hierarchical grammars: the linear grammars are either too complex or fit the data too poorly by comparison. Our ideal learning analysis thus infers that the syntax of English, at least as represented by this corpus, is best explained using the hierarchical phrase structures of context-free grammars rather than the linear structures of simpler Markovian grammars. In essence, our analysis reproduces one of the founding insights of generative grammar (Chomsky, 1956, 1957): hierarchical phrase-structure grammars are better than Markovian linear grammars as models of the range of syntactic forms found in natural language. Child learners could in principle make the same inference, if they can draw on the same rational inductive principles.

Berwick (1982; Berwick & Weinberg, 1986) presented a different approach to formalizing the simplicity argument for hierarchical structure in syntax, using tools from automata theory and Kolmogorov complexity. At a high level, our analysis and Berwick's are similar, but there are two important differences. First, we evaluate learnability of a sizeable and realistic CFG for English on a natural corpus of child-directed speech, rather

than simple languages (e.g., palindrome or mirror-symmetry languages on a binary alphabet) with idealized corpora (e.g., including all sentences generated by the grammar, or all sentences up to some maximum length or depth of recursion). Second, rather than considering only those grammars that fit the corpus precisely and evaluating them based only on their simplicity, we adopt a Bayesian framework in which simplicity of the grammar (in the prior) trades off against how well the grammar fits the particular corpus (measured by the likelihood). We can see where these differences matter in comparing the scores of particular grammars at different levels of evidence. Although regular grammars never receive the highest score across grammars of all types, they all score higher on the smallest corpus (*Level 1*) than do any of the hierarchical grammars. On most levels of evidence, at least one regular grammar is preferred over CFG-L. CFG-L is in fact the grammar that is ultimately favored on the full corpus, but it overgeneralizes far more than the regular grammars, and for the smaller corpora the tradeoff between simplicity and fit weighs against it. Thus our results cannot be obviously predicted by simplicity-based learnability analyses developed for artificial grammars on idealized corpora. The tradeoff between simplicity and degree of fit, given the sparse and idiosyncratic nature of child-directed language input, is critical to determining what kind of grammar an ideal learner should acquire.

It is interesting that the smallest corpora are best accounted for by the flat and one-state grammars. The smallest corpus contains only eight sentence types, with an average 1.6 words per sentence; thus, it is not surprising that it is optimal to simply memorize the corpus. Why is the one-state grammar preferred on the *Level 2* and *Level 3* corpora? Its simplicity gives it a substantial advantage in the prior, but we might expect it to suffer greatly in the likelihood because it can predict literally any sequence of syntactic categories as a possible sentence. The low likelihood does wind up ruling out the one-state grammar on larger but not smaller corpora: this is because the likelihood is not completely uninformative since it can encode the relative probability of each of the syntactic categories. Though this minimal model never fits the data well, it doesn't fit the smaller corpora so poorly as to overcome the advantage due to the prior. This suggests that simply encoding the statistical distribution of syntactic categories may be helpful at the earliest stages of language learning, even though it is ultimately a poor predictor of natural language.

What kind of input is responsible for the transition from linear to hierarchical grammars? The smallest three corpora contain very few elements generated from recursive productions (e.g., nested prepositional phrases or relative clauses) or sentences using the same kind of 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 these elements, a context-free grammar need add fewer (especially CFG-S). As a consequence, the flat and regular grammars have poorer generalization ability and must add proportionally more productions in order to parse a novel sentence.

The larger context-free grammar CFG-L outperforms CFG-S on the full corpus, probably because it includes non-recursive counterparts to some of its recursive productions. This results in a significantly higher likelihood since less of the probability mass is invested in recursive productions that are used much less frequently than the non-recursive ones. Thus, although both grammars have similar expressive power, the CFG-L is favored on larger corpora because the likelihood advantage overwhelms the disadvantage in the prior.

Local search from hand-designed grammars.

To what extent are these results dependent on our particular hand-designed grammars? We address this question by analyzing the posterior scores of those grammars identified via local search. Figure 7 depicts the posterior probabilities of all of the grammars considered in the search. Many linear grammars found by the automatic search procedures have scores similar to the best hand-designed linear grammars, but none have posterior probabilities close to that of the best context-free grammars.

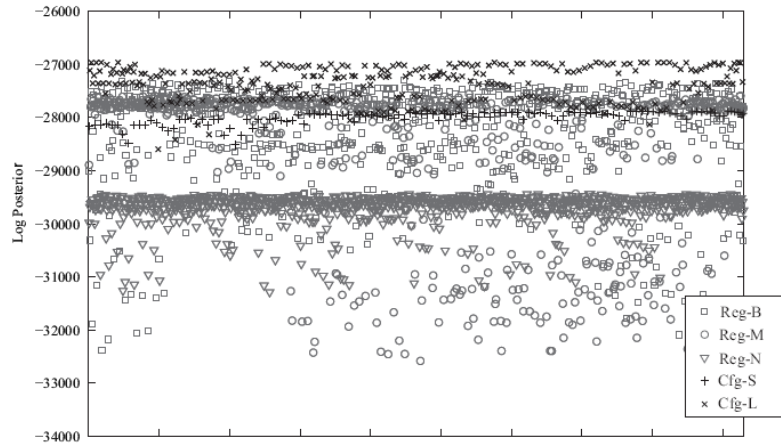


Figure 7. Posterior probabilities of each grammar considered during the search of regular and context-free possibilities. Each point represents one grammar; the x-axis is meaningless.

The posterior probabilities of the best grammars of each type found after the local search are shown in Table 3. The results are qualitatively similar to those obtained with hand-designed grammars: the posterior still favors a context-free grammar once the corpus is large enough, but for smaller corpora the best grammars are linear.

Identifying a regular grammar by automated search.

In addition to identifying the best grammars resulting from a local search, we can also examine the best regular grammar (REG-AUTO) found in a purely automated fashion using the unsupervised learning model developed by Goldwater & Griffiths (2007). The grammars with the highest posterior probability on each corpus are shown in Table 4. All of the REG-AUTO grammars have posterior probabilities similar to those of the other regular grammars, but on the larger corpora none have higher probability than the best context-free grammars. Because the REG-AUTO grammars do not consistently have *higher* probability than the other regular grammars, we cannot conclude that they represent the “true best” from the space of all possible grammars of that type. However, the fact that the best regular grammars found by every method have a similar order-of-magnitude probability – and that none have been found that approach the best-performing context-free grammar – suggests that if better regular grammars do exist, they are not easy to discover.

Summing up these results, we find that a context-free grammar always has the highest posterior probability on the largest corpus, compared to a variety of plausible linear

grammars. Though the ability of the hierarchical grammars to generate a higher variety of sentences from fewer productions typically results in a lower likelihood, this compression helps dramatically in the prior. A hierarchical grammar thus consistently maximizes the tradeoff between data fit and complexity. Drawing the analogy to the models in Figure 3, the best context-free grammar is most analogous to B. The one-state grammar, like A, is very simple but offers a very poor fit to the data, and the flat grammars may be more like C: a closer fit to the data, but too complex to be ideal. The regular grammars span the range between A and C, but none provides as good a tradeoff as in B.

Ungrammatical sentences

One decision made in constructing the corpus was to remove the ungrammatical sentences. This decision was primarily a pragmatic one, but we believe it is justified for several reasons. A child learning a language might be able to identify at least some of the ungrammatical sentences as such, based on pragmatic signals or on portions of the grammar learned so far. Also, if learners disregard sentence forms that occur very rarely, this would minimize the problem posed by ungrammatical sentences: they would be able to ignore the majority of ungrammatical sentences, but relatively few grammatical ones. Finally, since the hierarchical grammar type is preferred on corpora as small as *Level 4* and no ungrammatical sentences occurred 10 times or more, it seemed unlikely that including ungrammatical sentences would alter our main findings.

Nevertheless, it is still useful to compare each of the grammars on the corpus that includes ungrammatical sentences in order to be certain that the decision to exclude them is not critical to the outcome.¹⁰ To the best grammars of each type, we added the minimum number of additional productions required to parse the ungrammatical corpus. The hierarchical grammars still have the highest posterior probability (*Level 6* posterior: CFG-L: -29963; REG-B: -30458; REG-M: -30725; CFG-S: -31008; REG-N: -33466; 1-ST: -43098; FLAT: -92737). Thus, considering the ungrammatical sentences along with the grammatical sentences does not qualitatively alter our findings.

¹⁰ The ungrammatical corpus is the full corpus plus the 191 ungrammatical sentence types that correspond to the 443 ungrammatical sentence tokens.

Sentence tokens vs sentence types

The likelihood was defined under a language model with separate generative processes: one for the allowable types of syntactic forms in a language, another for the frequency of specific sentence tokens. This definition effectively restricts the data considered by the model to only include sentence types, rather than each individual sentence token. Defining the likelihood in this way was a principled choice, paralleling standard linguistic practice, in which grammar learning is based on how well each grammar accounts for which sentences occur, rather than their frequency distribution. It is nevertheless useful to explore precisely what the effect of making this choice is.

Interestingly, the linear grammars were overwhelmingly preferred over the hierarchical grammars on the corpus of sentence tokens (*Level 6* posterior: REG-N: -135704; REG-M: -136965; REG-B: -136389; CFG-L: -145729; CFG-S: -148792; FLAT: -188403; 1-ST: -212551). As before, the context-free grammars had higher prior probability – but unlike before, the linear grammars' goodness-of-fit outweighed the preference for simplicity. Why? The corpus of sentence tokens contains almost ten times as much data, but no concomitant increase in the variety of sentences (as would occur if there were simply more types, corresponding to a larger dataset of tokens). As a result, the likelihood term is weighted much more highly, thus more strongly penalizing the hierarchical grammars (which overgeneralize more).

This result suggests that if the hierarchical structure of syntax is to be inferred from observed data, the learner must have some sort of disposition to evaluate grammars with respect to type-based rather than token-based data. Such a disposition could be language-specific, but need not be, since it could also arise due to memory constraints or other cognitive factors. Indeed, it may be a necessary component for learning in many domains: without it, one would expect adults' representations to grow ever-closer to simple memorization as they get older since the importance of simplicity of representation (the prior) would shrink relative to the importance of closely fitting the ever-larger set of data accrued over a lifetime (the likelihood). That this does *not* happen may suggest that many aspects of learning might be more properly understood as occurring over type rather than token data. This may depend to some extent on particular characteristics of our analysis (e.g., it is unclear to what extent it relies on having grammars that are defined over syntactic

categories rather than individual words); because of this and because the role of type and token data is not generally well-understood, any conclusions in this vein must be tentative. Nevertheless, the apparent importance of type-based rather than token-based reasoning is an interesting implication of this work, and we will briefly return to this issue in the discussion section.

Age-based stratification

Our results may have developmental implications, but these must be interpreted with caution. Our findings do not necessarily imply that children should go through a period of using a simpler flat or one-state grammar, just because those grammar types were found to do best on the smaller type-based corpora. The *Levels* corpora are based on divisions by sentence frequency rather than by age. Though it is plausible that children can parse the simpler and more common sentences before the longer, rarer ones, it is certainly not the case that they acquire an understanding of language sentence by sentence, fully understanding some sentences and not at all understanding everything else. Thus, the different *Levels* corpora probably do not directly correspond to the amount of input available to the children at various ages. Instead, the division into *Levels* allows for an exploration of the tradeoff between complexity and data fit as the quantity of evidence increases.

It is nevertheless worthwhile to estimate, at least approximately, how soon that evidence is available to children. We therefore compare the posterior probabilities of the grammars on the *Epoch* corpora, which were constructed creating age-based divisions in the full corpus. Table 5 shows the probabilities of the best hand-designed linear and hierarchical grammars on these corpora. Strikingly, a context-free grammar is preferred at every age. This is even true for grammars that correspond to just the first file (*Epoch 0*), which consists of one hour of conversation at age 2;3. It is also interesting that the prior probabilities of the CFG-S and CFG-L grammars beginning at *Epoch 3* do not change. Why is this? Recall that at each epoch and level, we evaluate only the subset of each grammar necessary to parse the sentences observed in the corresponding corpus (removing any unnecessary productions). The fact that the CFGs stabilize by *Epoch 3* suggests that only 60% of the corpus is necessary to support the same grammars that are also preferred for the entire corpus. This is a consequence of the powerful generalization capacity that comes

from using a CFG. In contrast, regular grammars generalize less appropriately: the best regular grammar must be supplemented with additional productions at every additional epoch, resulting in a prior probability that continues to change as the corpus grows.

Do these results indicate that English-speaking children, if they are rational learners, can conclude after only a few hours of conversation that language has hierarchical phrase structure? Definitely not. In order to draw such a conclusion the child would minimally need to assign each word to its correct syntactic category and also be able to remember and parse somewhat complex utterances – capacities which are taken for granted in our model. However, this analysis does show that the data supporting a hierarchical phrase structure for English are so ubiquitous that once a learner has some ability to assign syntactic categories to words and to parse sentences of sufficient complexity, it should be possible to infer that hierarchical grammars provide the best description of the language's syntax.

It is interesting and theoretically important that the amount of data required to infer the existence of hierarchical phrase structure is much less than is required to infer all the rules of the correct hierarchical phrase-structure grammar. In terms of Figures 2 and 4, an ideal learner can infer the correct hypothesis at the higher level of abstraction T from less data than is required for inferring the correct hypothesis at a lower level, G . Although we have not demonstrated this here, it is theoretically possible that during the course of acquisition, higher-level knowledge, once learned, may usefully constrain predictions about unseen data. It might also effectively act in ways that are hard to distinguish from innate knowledge or innate constraints, given that it can be learned from such little data. We will return to this point in the discussion below.

Generalizability

Though posterior probability penalizes overgeneralization via the likelihood, it is important for a natural language learner to be able to generalize beyond the input observed, to be able to parse and comprehend novel sentences. How well do the different grammars predict unseen sentences? One measure of this is the percentage of the full (*Level 6*) corpus that can be parsed by the best grammars learned for subsets (*Level 1* to *5*) of the full corpus. If a grammar learned from a smaller corpus can parse sentence types in the full corpus that do not exist in its subset, it has generalized beyond the input it received and generalized in a

correct fashion. Table 6 shows the percentage of sentence types and tokens in the full *Level 6* corpus that can be parsed by each of the best grammars for the smaller *Levels*. The context-free grammars usually generalize the most, followed by the regular grammars. The flat grammar does not generalize at all: at each level it can only parse the sentences it has direct experience of. The one-state grammar can generalize to 100% of sentence types and tokens at every level because it can generalize to 100% of all sentences, grammatical or not.

Do the context-free grammars simply generalize more than the regular grammars, or do they generalize *in the right way*? In other words, would the context-free grammars also recognize and parse more *ungrammatical* English sentences than the regular grammars? Of the 191 ungrammatical sentence types excluded from the full corpus, the REG-B parses the most (107), followed by CFG-L (84), CFG-S (73), REG-M (72), and REG-N (57). Aside from the flat grammar, all of the grammars make some incorrect overgeneralizations. This should not be surprising given that our grammars lack the expressivity needed to encode important syntactic constraints, such as agreement. However, it is interesting that the REG-M grammar, which generalizes less than either context-free grammar to the full corpus in Table 6, generalizes to the ungrammatical sentences similarly to CFG-S: to the extent that REG-M grammar generalizes, it does so more often in the wrong way by making more incorrect overgeneralizations. This is even more striking in the case of the REG-B grammar, which parses somewhat fewer “correct” sentences (in the full corpus) than either context-free grammar, but parses many more “incorrect” (ungrammatical) sentences than the others.

The hierarchical grammars also generalize more appropriately than the linear grammars in the specific case of auxiliary-fronting for interrogative sentences. As Table 7 shows, both context-free grammars can parse aux-fronted interrogatives containing subject NPs that have relative clauses with auxiliaries – Chomsky's critical forms – despite never having seen an example of these forms in the input. They can do so because the input does contain simple declaratives and interrogatives, which license interrogative productions that do not contain an auxiliary in the main clause. The input also contains relative clauses, which are parsed as part of the noun phrase using the production $[NP \rightarrow NP CP]$. Both context-free grammars can therefore parse an interrogative with a subject NP containing a relative clause, despite never having seen that form in the input.

Unlike the context-free grammars, neither regular grammar can correctly parse complex aux-fronted interrogatives. The larger regular grammar REG-N cannot because, although its NP_{CP} productions can parse a relative clause in an NP, it does not have productions that can parse input in which a verb phrase without a main clause auxiliary follows an NP_{CP} production. This is because there was no input in which such a verb phrase *did* occur, so the only NP_{CP} productions occur either at the end of a sentence in the object NP, or followed by a normal verb phrase. Complex interrogative sentences – exactly the input that Chomsky argued are necessary – would be required to drive this grammar to the correct generalization.

The other regular grammars, REG-M and REG-B, cannot parse complex interrogatives for a different reason. Because they do not create a separate non-terminal like NP_{CP} for NPs containing relative clauses, they do have productions that can parse input in which such a subject NP is followed by a verb phrase without a main clause auxiliary. However, since they do not represent phrases *as phrases*, successful parsing of the complex interrogative “Can eagles that are alive fly?” (*aux n comp aux adj vi*) would require that the sentence have an expansion in which the non-terminal *adj* is followed by a *vi*. Because no sentences in the input follow this pattern, the grammars cannot parse it, and therefore cannot parse the complex interrogative sentence in which it occurs.

The superior generalization ability of the hierarchical grammars, though it hurts their likelihood scores, is of critical importance. Chomsky's original suggestion that structure-independent (linear) rules might be taken as more natural accounts of the data may have rested on the intuition that a grammar that sticks as closely as possible to the observed data is simpler without any *a priori* biases to the contrary. Such grammars do indeed predict the data better; they receive higher likelihood than the hierarchical grammars, which overgeneralize and thus waste some probability mass on sentence types that are never observed. However, a grammar that overgeneralizes – not too far, and just in the right ways – is necessary in order to parse the potentially infinite number of novel sentences faced by a learner of natural language. Of all the grammars explored, only the hierarchical grammars generalize in the same way humans do. While in a sense this should not be a surprise, it is noteworthy that a rational learner given child-directed language input prefers these

grammars over those that do not generalize appropriately, without direct evidence pointing either way.

Discussion

Our model of language learning suggests that there may be sufficient evidence in the input for an ideal rational learner to conclude that language has hierarchical phrase structure without having an innate language-specific bias to do so. The best-performing grammars correctly form interrogatives by fronting the main clause auxiliary, even though the input contains none of the crucial data Chomsky identified. In this discussion, we consider the implications of these results for more general questions of innateness, and for the nature of language acquisition in human children.

The question of innateness

In debates about innateness, there are often tradeoffs between the power of the learning mechanism, the expressive potential of the representation, and the amount of built-in domain-specific knowledge. Our modeling framework enables us to make assumptions about each of these factors explicit, and thereby analyze whether these assumptions are fair as well as to what extent the conclusions depend upon them. The issue is also more complicated than is captured by making the distinction between representational structure and the nature of the cognitive biases necessary (as in Figure 1). There is the additional question of which of the many capacities underlying successful language use are innate, as well as to what extent *each* capacity is domain-general or domain-specific. The PoS argument we consider here is concerned with whether a particular feature of linguistic syntax – hierarchical structure – must be innately specified as part of a language-specific learning module in children’s minds. Our analysis incorporates several assumptions about the cognitive resources available to children, but these resources are plausibly domain-general.

Probably the strongest assumption in the analysis is a powerful learning mechanism. We assume both that the learner can effectively search over the space of all possible grammars to arrive at optimal or near-optimal hypotheses, and that the grammars we have

analyzed are sufficiently close to the optimal ones. Advances in computational linguistics and the development of more powerful models of unsupervised grammar induction will do much to address the latter assumption, and until then, our conclusions are of necessity preliminary. In the meantime, we can have some confidence based on the fact that every linear grammar we were able to construct through various and exhaustive means performed less well than the best hierarchical grammars we found. Moreover, the poor performance of linear grammars appears to occur for a principled reason: they require more productions in order to match the degree of fit attained by context-free grammars, and therefore fail to maximize the complexity-fit tradeoff.

Even if our approach succeeds in identifying (near-)optimal grammars, the assumption that child learners can effectively search the space of all possible grammars is a strong one. Especially for context-free grammars, where the space is much larger than for regular grammars, it may be that learners will need some built-in biases in order to search effectively.¹¹ In general, one must assume either a powerful domain-general learning mechanism with only a few general innate biases that guide the search, or a weaker learning mechanism with stronger innate biases, or some compromise position. Our results do not suggest that any of these possibilities is more likely than the others. Our core argument concerns only the specific need for a bias to *a priori* prefer analyses of syntax that incorporate hierarchical phrase structure. We are arguing that a rational learner may not require such a bias, not that other biases are also unnecessary.

In addition to assumptions about the learning mechanism, our model incorporates some assumptions about the representational abilities of the child. First of all, we assume that children have the (domain-general) capacity to represent various types of grammars, including both hierarchical and linear grammars. We are not claiming that the specific grammars we analyze are exactly the ones children represent; clearly all of the grammars we have worked with are oversimplified in many ways. But an assumption that children in some sense have the capacity to represent both linear and hierarchical patterns of sequential structures – linguistic or non-linguistic – is necessary to even ask the questions we consider here. If children were not born with the capacity to represent the thoughts they later grow to

¹¹ Of course, because linear grammars are a subset of context-free grammars, biases for searching the space of context-free grammars could work for linear grammars as well. Furthermore, such biases need not be domain-specific.

have, no learning in that direction could possibly occur. Our analysis also assumes that the learner represents the different grammar types *as* different grammar types, choosing between context-free, regular, flat, and one-state grammars. This stratification is not critical to the results, however. If anything, it is a conservative assumption, because it favors the non-hierarchical grammars more heavily than they would be favored if we treated all grammars as instances of a single general type.¹²

Perhaps the most basic representational assumption is that learners are evaluating grammars with explicit symbolic structure. Although this assumption is not particularly controversial in linguistics, it has been expressly denied in other recent analyses of PoS arguments by cognitive modelers (e.g., Lewis & Elman, 2001; Reali & Christiansen, 2005). We are not arguing that the assumption of explicit structure is the only viable route to understanding language acquisition as a kind of inductive learning. It is important and useful to explore the alternative possibility that generalizations about grammatical structure are not innate because such structure either does not exist at all or is present only implicitly in some kind of sub-symbolic representation. But it is also important to consider the possibility that these generalizations about grammatical structure exist explicitly and can still be learned. One motivation is simply thoroughness: any possibility that cannot be ruled out on *a priori* grounds should be investigated. In other words, we should not artificially restrict ourselves from exploring the upper right quadrant of Figure 1. Another reason is that the reality of linguistic structure is widely accepted in standard linguistics. There are many linguistic phenomena whose only satisfying explanations (to date, not in principle) have been framed in structured symbolic terms. Explicitly structured representations also provide the basis of most state-of-the-art approaches in computational linguistics (e.g., Charniak, 1993; Manning & Schütze, 1999; Collins, 1999; Eisner, 2002; Klein & Manning, 2004). Given how useful structured grammatical representations have been both for explaining linguistic phenomena and behavior and for building effective computer systems for natural language processing, it seems worthwhile to take seriously the possibility that they might be the substrate over which children represent and learn language.

A final assumption concerns the way we represent the input to learning. We have given our model a corpus consisting of sequences of syntactic categories, corresponding to

¹² See Appendix B for details.

types of sentences, rather than sequences of lexical items which would correspond to actual sentence tokens.¹³ The use of syntactic categories may not be necessary in principle, but it greatly simplifies the analysis. It allows us to focus on learning grammars from the syntactic-category data they immediately generate rather than having to infer this intermediate layer of representation from raw sequences of individual words. We make no claims about how children might initially acquire these syntactic categories, and our analysis would not change if they themselves were shown to be innate (whatever that would mean). There is some evidence that aspects of this knowledge may be in place even in children below the age of two (Booth & Waxman, 2003), and that syntactic categories may be learnable from simple distributional information without reference to the underlying grammar (Schütze, 1995; Redington et. al., 1998; Mintz et. al., 2002; Gerken et. al., 2005; Griffiths et. al., 2005). Thus we think it is plausible to assume that children have access to something like the input we have used here as they approach problems of grammar acquisition. However, it would still be desirable for future work to move beyond the assumption of given syntactic categories. It is possible that the best linear grammars might use entirely different syntactic categories than those we assumed here. It would be valuable to explore whether hierarchical grammars continue to score better than linear grammars if the input to learning consisted of automatically labeled part-of-speech tags rather than hand-labeled syntactic categories.

All of these assumptions involve either domain-general representational or learning abilities, or language-specific knowledge (about syntactic categories) that is distinct from the knowledge being debated: we critically did *not* assume that learners must know in advance that language specifically has hierarchical phrase structure. Rather, we showed that this knowledge can be acquired by an ideal learner equipped with sophisticated domain-general statistical inference mechanisms and a domain-general ability to represent hierarchical structure in sequences – a type of structure found in many domains outside of natural language. The model contains no *a priori* bias to prefer hierarchical grammars in language – a bias that classic PoS arguments have argued was necessary. The learned

¹³ Tomasello (2000) and others suggest that children initially restrict their syntactic frames to be used with particular verbs (the so-called “verb island” hypothesis). Our model treats all members of a given syntactic category the same, and therefore does not capture this hypothesis. However, this aspect of the model reflects a merely pragmatic decision based on ease of computation. An extension of the model that took lexical items rather than syntactic categories as input could incorporate interesting item-specific dependencies.

preference for hierarchical grammars is data-driven, and different data could have resulted in a different outcome. Indeed, we find different outcomes when we restrict attention to only part of the data available to the child.

Relevance to human language acquisition

What conclusions, if any, may we draw from this work about the nature of grammatical acquisition in human learners? Our analysis focuses on an ideal learner, in the spirit of Marr's level of computational theory. Just as Chomsky's original argument focused on what was in principle impossible for humans to learn without some innate knowledge, our response looks at what is in principle possible. While this ideal learning analysis helps recalibrate the bounds of what is possible, it may not necessarily describe the actual learning processes of human children.

One concern is that it is unclear to what extent humans actually approximate rational learners. On the positive side, rational models of learning and inference based on Bayesian statistical principles have recently developed into a useful framework for understanding many aspects of human cognition (Anderson, 1991; Chater & Oaksford, 1999; Chater et. al., 2006). Chomsky himself appealed to the notion of an objective neutral scientist studying the structure of natural language, who rationally should first consider the linear rule for auxiliary-fronting because it is *a priori* less complex (Chomsky, 1971). Although there is some debate about how best to formalize rational scientific inference, Bayesian approaches offer what is arguably the most promising general approach (Howson & Urbach, 1993; Jaynes, 2003). A more deductive or falsificationist approach (Popper, 1959) to scientific inference might underlie Chomsky's view: an objective neutral scientist should maintain belief in the simplest rule – e.g., the linear rule for auxiliary-fronting – until counterevidence is observed, and because such counterevidence is never observed in the auxiliary-fronting case, that scientist would incorrectly stay with the linear rule. But under the view that scientific discovery is a kind of inference to the best explanation – which is naturally captured in a Bayesian framework such as ours – the hierarchical rule could be preferred even without direct counterevidence eliminating the simpler alternative. This is particularly true when we consider the discovery problem as learning the grammar of a language as a whole, where the rules for parsing a particular kind of sentence (such as complex auxiliary-

fronted interrogatives) may emerge as a byproduct of learning how to parse many other kinds of sentences. The rational Bayesian learning framework we have adopted here certainly bears more resemblance to the practice of actual linguists – who after all are mostly convinced that language does indeed have hierarchical structure! – than does a falsificationist neutral scientist.

Defining the prior probability unavoidably requires making particular assumptions. A simplicity metric defined over a very different representation would probably yield different results, but this does not pose a problem for our analysis. The classic PoS argument claims that it would be impossible for a reasonable learner to learn a structure-dependent rule like aux-fronting. All that is required to respond to such a claim is to demonstrate that such a reasonable learner *could* learn this. Indeed, our prior is reasonable: consistent with intuition, it assigns higher probability to shorter and simpler grammars, and it is defined over a sensible space of grammars that is capable of representing linguistically realistic abstractions like noun and verb phrases. Even if a radically different simplicity metric were to yield different results, this would not change our conclusion that *some* reasonable learner could learn that language is structure-dependent.

Another issue for cognitive plausibility is the question of scalability: the largest corpus presented to our model contains only 2336 sentence types, many less than the average human learner is exposed to in a lifetime. Since our results are driven by the simplicity advantage of the context-free grammars (as reflected in their prior probabilities), it might be possible that increasing quantities of data would eventually drown out this advantage in favor of advantages in the likelihood. We think this is unlikely for two reasons. First, the number of sentence types grows far less rapidly than the number of distinct sentence tokens, and the likelihoods in our analysis are defined over the former rather than the latter. Secondly, as we have shown, additional (grammatical) sentence types are more likely to be already parseable by a context-free grammar than by a regular grammar. This means that the appearance of those types will actually *improve* the relative likelihood of the context-free grammar (because they will no longer constitute an overgeneralization) while not changing the prior probability at all; by contrast, the regular grammar may more often need to add productions in order to account for an additional sentence type, resulting in a lower prior probability and thus a lower relative posterior score.

If the knowledge that language has hierarchical phrase structure is not in fact innate, why do all known human languages appear to have hierarchical phrase structure? This is a good question, and we can only offer speculation here. One answer is that nothing in our analysis precludes the possibility that children have a cognitive bias towards syntactic systems organized around hierarchical phrase structures: our point is that the classic PoS argument may not be a good reason to believe that they do, or that the bias need be specifically linguistic. Another answer is that even if a rational learner could in principle infer hierarchical structure from typical data, that does not mean that actual children necessarily do so: such knowledge might still be innate in some way, either language-specifically or emergent from cognitive, memory-based, or perceptual biases. For instance, if human thoughts are fundamentally structured in a hierarchical fashion, and if children have an initial bias to treat syntax as a system of rules for mapping between thoughts and sequences of sounds, then this could effectively amount to an implicit bias for hierarchical structure in syntax. In fact, our finding that hierarchical phrase structure is only preferred for corpora of sentence types (rather than tokens) may suggest that a bias to attend to types is necessary to explain children's acquisition patterns. It is also still possible that there are no biases in this direction at all – cognitive or otherwise – in which case one might expect to see languages without hierarchical phrase structure. There have recently been claims to that effect (e.g., Everett, 2005), although much work remains to verify them.

Although Chomsky's original formulation of the PoS argument focused on the innateness of the hierarchical structure of language, recent characterizations of an innate language faculty have concentrated on recursion in particular (Hauser et. al., 2002). An interesting aspect of our results is that although all of the best context-free grammars we found contained recursive productions, the model prefers grammars (CFG-L) that also contain non-recursive counterparts for complex NPs (noun phrases with embedded relative clauses).¹⁴ It is difficult to know how to interpret these results, but one possibility is that perhaps syntax, while fundamentally recursive, could also usefully employ non-recursive rules to parse simpler sentences that recursive productions could parse in principle. These non-recursive productions do not alter the range of sentence types the grammar can parse, but they are useful in more precisely matching the linguistic input. In general, our paradigm

¹⁴ See Perfors et al. (under review) for a more detailed exploration of this issue.

provides a method for the quantitative treatment of recursion and other contemporary questions about the innate core of language. Using it, we can address questions about how much recursion an optimal grammar for a language should have, and where it should have it.

More general implications

Our analysis makes 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 that children appear to master and conclude that because there is not enough evidence for that phenomenon in isolation, it must be innate. We have suggested here that even when the data does not appear to explain an isolated inference, there may be enough evidence to learn a larger system of linguistic knowledge – a whole grammar – of which the isolated inference is a part. A similar intuition underlies other arguments about the important role that indirect evidence might play in language acquisition (Landauer & Dumais, 1997; Regier & Gahl, 2004; Reali & Christiansen, 2005). This point is broadly consistent with the generative tradition in linguistics (Chomsky, 1957), one of whose original goals was to unify apparently disparate aspects of syntax (such as phenomena surrounding wh-fronting, auxiliary-fronting, and extraction) as resulting from the same underlying linguistic system. However, this insight has been largely missing from more recent discussions of PoS arguments (e.g., Chomsky, 1980; Laurence and Margolis, 2001; Pullum and Scholz, 2002), which have set the agenda for ongoing debates about language-specific innate knowledge primarily by arguing about the learnability of individual syntactic phenomena.

In general, our work suggests a paradigm for investigating some of the unexplored regions of Figure 1: the possibility that structured representations of specific domains may be learnable by largely domain-general mechanisms. Bayesian modeling is appropriate and useful in contexts like this for several reasons. It offers a normative framework for rational inference and for quantitatively exploring the domain-general tradeoff between simplicity and fit to data. The computations can be defined over structured representations, not just simple kinds of input statistics or correlations as in other paradigms for statistical learning. In addition to domain-general inferential principles, the framework can incorporate domain-

specific information, either by specifying unique details of the representation, incorporating biases into priors, or calculating likelihoods in some domain-specific way. Thus, the framework lets us investigate the role of both domain-general and domain-specific factors in learning, as well as the role of different kinds of representational structure.

In virtue of how it integrates statistical learning with structured representations, the Bayesian approach can apply to questions of learnability for many different aspects of linguistic knowledge, not just the specific question of hierarchical phrase structure addressed here. It can also be extended to the more general version of the Poverty of the Stimulus argument explicated by Laurence & Margolis (2001), in which the hypothesis space of possible grammars is infinite in size. Even under such conditions, Bayesian model selection can identify the best grammar. Laurence & Margolis (2001) argue that strong innate knowledge would be needed to rule out many logically possible but unnatural grammars, such as those that incorporate disjunctive hypotheses, which face no direct counterevidence in the observed data. But many of these “unnatural” alternatives – in particular, needlessly disjunctive hypotheses – would naturally be disfavored by a Bayesian learner, due to the automatic Bayesian Occam's razor, without the need for language-specific innate biases against them. Grammars that posit unnecessary complexity that does not result in improved fit to the data, including some of the “unnatural” cases that they worry about, would receive lower posterior scores than simpler grammars which fit the data just as well. There may still be unnatural alternative grammars that cannot be ruled out in this way: we are not trying to claim that all PoS arguments will lose their force. Rather, we now have tools to more clearly identify which PoS arguments for innate domain-specific knowledge are compelling, if any, and to sharpen their points by showing exactly when and why powerful domain-general learning principles might fail to account for them.

One implication of our work is that it may be possible to learn a higher-order abstraction T even before identifying all of the correct lower-level generalizations G that T supports. Therefore, it may be possible for T to operate to constrain G even if it itself is learned. Though our model here did not explicitly use inferences about T to constrain inferences about G , it could have done so, since T was learned at lower levels of evidence than were necessary to acquire the full specific grammar or to parse complex interrogative sentences.

In a sense, this finding reconstructs the key intuition behind linguistic nativism, preserving what is almost certainly right about it while eliminating some of its less justifiable aspects. The basic motivation for positing innate knowledge of grammar, or more generally innate constraints on cognitive development, is that without these constraints, children would be unable to infer the specific knowledge that they seem to come to from the limited data available to them. What is critical to the argument is that some constraints are present prior to learning specific grammatical rules, not that those constraints must be innate. Approaches to cognitive development that emphasize learning from data typically view the course of development as a progressive layering of increasingly abstract knowledge on top of more concrete representations; under such a view, learned abstract knowledge would tend to come in after more specific concrete knowledge is learned, so the former could not usefully constrain the latter. This view is sensible in the absence of learning mechanisms that can explain how abstract constraints could be learned together with (or before) the more specific knowledge they are needed to constrain. However, our work offers an alternative, by providing just such a learning mechanism in the form of hierarchical Bayesian models. If an abstract generalization can be acquired very early and can function as a constraint on later development of specific rules of grammar, it may function effectively as if it were an innate domain-specific constraint, even if it is in fact not innate and instead is acquired by domain-general induction from data.

How is it possible to learn a higher-order generalization before a lower-order one? Although it may seem counterintuitive, there are conditions under which higher-order generalizations should be easier to acquire for a Bayesian learner, and these conditions apply to the case we study here. While there are infinitely many possible specific grammars G , there are only a small number of possible grammar types T . It may thus require less evidence to identify the correct T than to identify the correct G . More deeply, because the higher level of T affects the grammar of the language as a whole while any component of G affects only a small subset of the language produced, there is in a sense much more data available about T than there is about any particular component of G . For instance, the sentence *adj adj n aux part* contributes evidence about certain aspects of the specific grammar G – that it is necessary to have productions that can generate such a sequence of words – but the evidence is irrelevant to other aspects of G – for instance, productions

involving non-auxiliary verbs. In general any sentence is going to be irrelevant (except for indirectly, insofar as it constitutes negative evidence) to inferences about most parts of the grammar: in particular, to all of the productions that are not needed to parse that sentence. By contrast, every sentence offers at least some evidence about the grammar type T – about whether language has hierarchical or linear phrase structure – based on whether rules generated from a hierarchical or linear grammar tend to provide a better account of that sentence. Higher-order generalizations may thus be learned faster simply because there is much more evidence relevant to them.

Conclusion

We have demonstrated that an ideal learner equipped with the resources to represent a range of symbolic grammars that differ qualitatively in structure, as well as the ability to find the best fitting grammars of various types according to a Bayesian score, can in principle infer the appropriateness of hierarchical phrase-structure grammars without the need for innate language-specific biases to that effect. By showing that an ideal learner can make this inference from actual child-directed speech, it becomes possible that human children could make this inference as well. Two important open questions remain: how well an ideal learnability analysis corresponds to the actual learning behavior of children, and how well our computational model approximates this ideal. Our specific conclusions are therefore preliminary and may need to be revised as we begin to learn more about these two fundamental issues. Regardless, we have offered a positive and plausible “in principle” response to the classic negative “in principle” poverty-of-stimulus arguments for innate language-specific knowledge.

More generally, we have suggested a new approach to classic questions of innateness. By working with sophisticated statistical inference mechanisms that can operate over structured representations of knowledge such as generative grammars, we can rigorously explore a relatively uncharted region of the theoretical landscape: the possibility that genuinely structured knowledge is genuinely learned, as opposed to the classic positions that focus on innate structured knowledge or learned unstructured knowledge, where apparent structure is merely implicit. Some general lessons can be drawn. It does not make sense to ask whether a specific generalization is based on innate knowledge when that

generalization is part of a much larger system of knowledge that is acquired as a whole. Abstract organizational principles can be induced based on evidence from one part of the system and effectively transferred to constrain learning of other parts of the system, as we saw for the auxiliary-fronting rule. These principles may also be learned prior to more concrete generalizations, or may be learnable from much less data than is required to identify most of the specific rules in a complex system of knowledge. We expect that these ideas could be usefully applied to explore learnability issues in other aspects of language, as well as for other areas of cognitive development, such as the development of children's intuitive theories of physical, biological, psychological or social domains.

Acknowledgements

For many helpful discussions, we thank Virginia Savova, Jeff Elman, Jay McClelland, Steven Pinker, Tim O'Donnell, Adam Albright, Robert Berwick, Cedric Boeckx, Edward Gibson, Ken Wexler, Danny Fox, Fei Xu, Michael Frank, Morten Christiansen, Daniel Everett, Noah Goodman, Vikash Mansinghka, and Charles Kemp. This work was supported by an NDSEG graduate fellowship (AP), an NSF graduate fellowship (AP), the Paul E. Newton Career Development Chair (JBT) and the James S. McDonnell Foundation Causal Learning Collaborative Initiative (JBT).

References

- Alishahi, A., & Stevenson, S. (2005). A probabilistic model of early argument structure acquisition. *Proceedings of the 27th annual meeting of the Cognitive Science Society*.
- Ambridge, B., Rowland, C., & Pine, J. (2005). Structure dependence: An innate constraint? *Poster presented at 30th annual Boston University Conference on Language Development*.
- Anderson, J. (1991). The adaptive nature of human categorization. *Psychology Review*, 98(3), 409-429
- Berwick, R. (1982) Locality principles and the acquisition of syntactic knowledge. PhD dissertation, MIT. Cambridge, MA
- Berwick, R. (1986). Learning from positive-only examples: The subset principle and three case studies. *Machine Learning*, 2, 625-645
- Berwick, R., & Weinberg, A. (1986) *The grammatical basis of linguistic performance: Language use and acquisition*. Cambridge, MA: MIT Press
- Booth, A., & Waxman, S. (2003). Mapping words to the world in infancy: Infants' expectations for count nouns and adjectives. *Journal of Cognition and Development*, 4(3), 357-381
- Brown, R. (1973). *A first language: The early stages*. Harvard University Press.
- Charniak, E. (1993). *Statistical language learning*. Cambridge, MA: MIT Press.
- Chater, N., & Manning, C. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10 (7), 335-344
- Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, 3, 57-65
- Chater, N., Tenenbaum, J., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10(7), 292-293
- Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7, 19-22
- Chater, N., & Vitányi, P. (2007). 'Ideal learning' of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology*, 51(3), 135-163
- Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2, 113-123

- Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.
- Chomsky, N. (1959). On certain formal properties of grammars. *Information and Control*, 2, 137-167
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
- Chomsky, N. (1971). *Problems of knowledge and freedom*. London: Fontana.
- Chomsky, N. (1980). In M. Piatelli-Palmarini (Ed.), *Language and learning: The debate between Jean Piaget and Noam Chomsky*. Cambridge, MA: Harvard University Press.
- Clark, A., & Eyraud, R. (2006). Learning auxiliary fronting with grammatical inference. *Proceedings of the 28th annual meeting of the Cognitive Science Society*.
- Collins, M. (1999). *Head-driven statistical models for natural language parsing*. Unpublished doctoral dissertation, University of Pennsylvania.
- Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language*, 24, 139-186
- Dowman, M. (1998). A cross-linguistic computational investigation of the learnability of syntactic, morphosyntactic, and phonological structure. *EUCCS-RP-1998-6*
- Dowman, M. (2000). Addressing the learnability of verb subcategorizations with Bayesian inference. *Proceedings of the 22nd annual conference of the Cognitive Science Society*.
- Edelsbrunner, H., & Grayson, D. (2000). Edgewise subdivision of a simplex. *Discrete Computational Geometry*, 24, 707-719
- Eisner, J. (2002). Discovering deep structure via Bayesian statistics. *Cognitive Science*, 26, 255-268
- Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.
- Everett, D. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current Anthropology*, 46(4), 621-646
- Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis* (2nd ed.). Chapman & Hall.
- Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, 32, 249-268

- Gold, E. M. (1967). Language identification in the limit. *Information and Control*, 10(5), 447-474
- Goldwater, S., & Griffiths, T. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. *45th annual meeting of the Association for Computational Linguistics*.
- Goldwater, S., Griffiths, T., & Johnson, M. (2005). Interpolating between types and tokens by estimating power law generators. *Neural Information Processing Systems*, 18.
- Goodman, N. (1954) *The new riddle of induction*. London, The Athlone Press
- Griffiths, T., Baraff, E., & Tenenbaum, J. (2004). Using physical theories to infer hidden causal structure. *26th annual conference of the Cognitive Science Society*.
- Griffiths, T., Steyvers, M., Blei, D., & Tenenbaum, J. (2005). Integrating topics and syntax. *Neural Information Processing Systems*, 17.
- Hauser, M., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298 (5598), 1569-1579
- Horning, J. J. (1969). *A study of grammatical inference* (Tech. Rep. #139). Stanford Univ.
- Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach* (2nd ed.). Open Court Publishing Company.
- Jaynes, E. (2003). *Probability theory: The logic of science*. Cambridge University Press.
- Johnson, M. (2006). *Inside-outside algorithm*. Brown University.
- Johnson, M., & Riezler, S. (2002). Statistical models of syntax learning and use. *Cognitive Science*, 239-253
- Jurafsky, D., & Martin, J. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition*. Prentice Hall.
- Kam, X., Stonyeshka, I., Tornyova, L., Sakas, W., Fodor, J.D. (2005) Statistics vs. UG in language acquisition: Does a bigram analysis predict auxiliary inversion?
- Kemp, C., Perfors, A., & Tenenbaum, J. B. (2004). Learning domain structures. *Proceedings of the 26th annual conference of the Cognitive Science Society*.
- Klein, D., & Manning, C. (2004). Corpus-based induction of syntactic structure: Models of dependency and constituency. *Proceedings of the 42nd annual meeting of the Association for Computational Linguistics*, 479-486

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104 (2), 211-240

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

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

Lewis, J., & Elman, J. (2001). Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. *Proceedings of the 26th Boston University Conference on Language Development*.

Li, M., & Vitányi, P. (1997). *An introduction to Kolmogorov complexity and its applications*. NY: Springer Verlag.

Light, M., & Greiff, W. (2002). Statistical models for the induction and use of selectional preferences. *Cognitive Science*, 26, 269-281

MacKay, D. (2003). *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.) Lawrence Erlbaum Associates.

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

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt & Company.

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

Perfors, A., Tenenbaum, J., Gibson, E., Regier, T. (under review) How recursive is language? A Bayesian exploration. *Linguistic Review*.

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

Popper, K. (1959). *The logic of scientific discovery*. Routledge.

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

- Reali, F., & Christiansen, M. (2004). Structure dependence in language acquisition: Uncovering the statistical richness of the stimulus. *Proceedings of the 26th Conference of the Cognitive Science Society*.
- Reali, F., & Christiansen, M. (2005). Uncovering the statistical richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, 29, 1007-1028
- Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22, 425-469
- Regier, T., & Gahl, S. (2004). Learning the unlearnable: The role of missing evidence. *Cognition*, 93, 147-155
- Rumelhart, D., & McClelland, J. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge, MA: MIT Press.
- Schütze, H. (1995). Distributional part-of-speech tagging. *Proceedings of the 7th conference of the European Chapter of the Association for Computational Linguistics*.
- Solomonoff, R. (1964). A formal theory of inductive inference. *Information and Control*, 7(1-22), 224-254
- Solomonoff, R. (1978). Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory*, 24, 422-432
- Stolcke, A., & Omohundro, S. (1994). Introducing probabilistic grammars by Bayesian model merging. *2nd International Colloquium on Grammatical Inference*.
- Tenenbaum, J., & Griffiths, T. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24, 629-641
- Tomasello, M. (2000). The item based nature of children's early syntactic development. *Trends in Cognitive Sciences*, 4, 156-163
- Wexler, K., & Culicover, P. (1980). *Formal principles of language acquisition*. Cambridge, MA: MIT Press.
- Xu, F., & Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review*. (in press)
- Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, 10(7)
- Zipf, G. (1932). *Selective studies and the principle of relative frequency in language*. Harvard University Press.

Appendix A

Searching the space of grammars

There are two search problems, corresponding to the two ways of building or improving upon our initial hand-designed grammars. The first is to perform a fully automated search over the space of regular grammars, and was described in the main text and in detail in Goldwater & Griffiths (2007). The second, described here, is the problem of performing local search using the best hand-designed grammar as a starting point.

Our search was inspired by work by Stolcke & Omohundro (1994), in which a space of grammars is searched via successive merging of productions; some sample merges are shown in Table A1. Merge rules are different for context-free and regular grammars; this prevents a search of regular grammars from resulting in a grammar with context-free productions.

At each stage in the search, all grammars one merge step away from the previous grammar are created. If the new grammar has a higher posterior probability than the current grammar, it is retained, and search continues until no grammars with higher posterior probability can be found within one merge step away.

Appendix B

Prior probabilities

Non-terminals, productions, and items

We model the probabilities of the number of non-terminals $p(n)$, productions $p(P)$, and items $p(N_i)$ as selections from a geometric distribution. One can motivate this distribution by imagining that non-terminals are generated by a simple automaton with two states (on or off).¹⁵ Beginning in the “on” state, the automaton generates a series of non-terminals; for each non-terminal generated, there is some probability p that the automaton

¹⁵ Productions and items can be generated in the same way, but for clarity of exposition we restrict ourselves to explaining the process in terms of non-terminals.

will move to the “off” state and stop generating non-terminals. This process creates a distribution over non-terminals described by Equation 4 and illustrated in Figure B1.

$$p(1-p)^{n-1} \quad (4)$$

No matter the value of the parameter p , this distribution favors smaller sets: larger values – i.e., those corresponding to more productions, non-terminals, or items – are less probable. All reported results use $p=0.5$, but the qualitative outcome is identical for a wide variety of values.

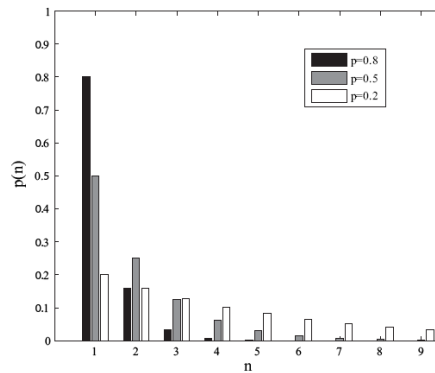


Figure B1. A geometric distribution, with three possible values of p .

Production-probability parameters

Because each θ_k corresponds to the production-probability parameters for non-terminal k , the individual parameters $\theta_1, \dots, \theta_m$ in each vector θ_k should sum to one. As is standard in such cases, we sample each θ_k from the Dirichlet distribution. Intuitively, this distribution returns the probability that the m_k production-probability parameters for non-terminal k are $\theta_1, \dots, \theta_m$, given that each production has been used $\alpha-1$ times. We set $\alpha = 1$, which is equivalent to having never observed any sentences and not assuming *a priori* that any one sentence or derivation is more likely than another. This therefore puts a uniform distribution on production-probability parameters and captures the assumption that any set of parameters is as likely as any other set. In general, drawing samples from a Dirichlet distribution with $\alpha = 1$ is equivalent to drawing samples uniformly at random from the $m_k - 1$ unit simplex; the simplex (distribution) for $m_k = 3$ is shown in Figure B2.

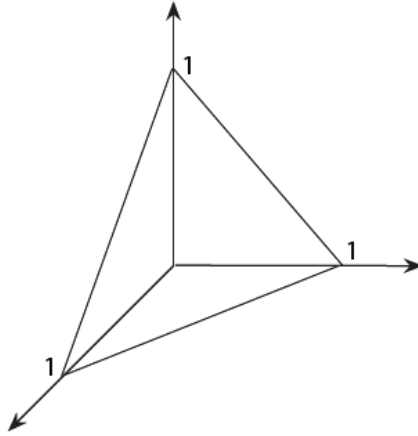


Figure B2. The unit simplex for $m_k = 3$ (a triangle), corresponding to the Dirichlet distribution with $\alpha = 1$ on a θ_k vector of production-probability parameters with three productions.

The Dirichlet distribution is continuous, which means that the probability of any specific θ_k is zero; this may seem paradoxical, but no more so than the fact that a line of length one inch contains an infinite number of zero-length points. Even though the distribution is continuous, one can still compare the relative probability of choosing the points from the line. For instance, consider the line in the upper part of Figure B3. If the probability of choosing any particular point is normally distributed about the center of the line, point A is more likely than point B. In much the same way, it is possible to calculate the relative probability of specific $\theta_1, \dots, \theta_m$, even though the Dirichlet distribution is continuous.

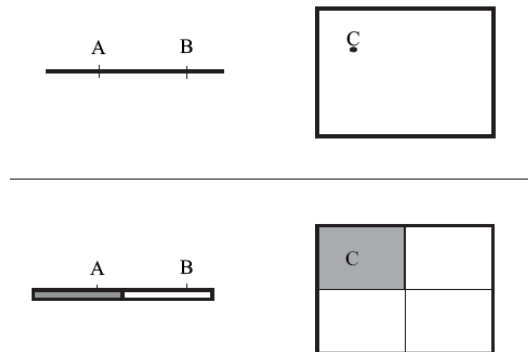


Figure B3. Top: one cannot compare the probability of continuous points A and C with different dimensionality. Bottom: when A and C correspond to discrete points, comparison is possible.

However, one cannot validly compare the relative probability of choosing points from sets with different dimensions, as in A and C in Figure B3. Because they are continuous, the probability of each is zero, but – unlike the previous instance – they are not normalized by the same factor. In an analogous way, it is also invalid to compare the probability of two specific θ_k of different dimensionalities.

This poses a difficulty for our analysis, because our grammars have different numbers of productions with the same left-hand sides, and therefore the θ_k are defined over different dimensionalities. We resolve this difficulty by using a discrete approximation of the continuous Dirichlet distribution. This is conceptually equivalent to comparing the probability of selecting point A and point C by dividing each dimension into g discrete segments. If we split each dimension into $g=2$ equally-sized discrete segments or grids, as in the lower half of Figure 3B, it becomes clear that the grid corresponding to point A contains half of the mass of the line, while the grid corresponding to C contains approximately one quarter the mass of the square. Thus, the probability of drawing C is 25%, while A is 50%. As g approaches infinity, the relative probabilities approach the true (continuous) value.

Since drawing samples $\theta_1, \dots, \theta_m$ from a Dirichlet distribution is equivalent to drawing samples from the $m_k - 1$ unit simplex, we calculate their probability by dividing the simplex into identically-sized pieces. Any $m-1$ simplex can be subdivided into g^{m-1} simplices of the same volume, where g is the number of subdivisions (grids) along each dimension (Edelsbrunner & Grayson, 2000). If $\alpha = 1$, all grids are *a priori* equally probable; thus, $p(\theta_k)$ is given by the volume of one grid divided by the volume of the entire simplex, that is, $1 / g^{m-1}$. Production-probability parameters are then set to the center-of-mass point of the corresponding grid.

As in the main analysis, there is a simplicity/goodness-of-fit tradeoff with size of grid g . If $g=1$, then vectors with many production-probability parameters have high prior probability (each is 1.0). However, they fit the data poorly: the parameters are automatically set to the center-of-mass point of the entire simplex, which corresponds to the case in which each production is equally likely. As g increases, the likelihood approaches the maximum likelihood value, but the prior probability goes down. We can capture this tradeoff by scoring g as we do other choices. We assign a possible distribution of grid sizes over g by

assuming that $\ln(g)$ is distributed geometrically with parameter $p=0.5$. Thus, smaller g has higher prior probability, and we can select the grid size that best maximizes the tradeoff between simplicity and goodness-of-fit. We evaluated each grammar with $g=1, 10, 100, 1000$, and 10000 . The results reported use $g=1000$ because that is the value that maximizes the posterior probability for all grammars; the hierarchical grammar type was preferred for all values of g .

Additional complexities involved in scoring prior probability

Depending on the type of the grammar, some specific probabilities vary. The flat grammar has no non-terminals (aside from S) and thus its $p(n)$ is always equal to 1.0. Both the regular and context-free grammars, written in Chomsky Normal Form to conform to standard linguistic usage, are constrained to either have one or two items on the right hand side. The regular grammars have further type-specific restrictions on what kind of item (terminal or non-terminal) may appear where, which effectively increase their prior probability relative to context-free grammars. These restrictions affect $p(N_i)$ as well as the effective vocabulary size V for specific items. For example, the first item on the right-hand side of productions in a regular grammar is constrained to be a terminal item; the effective V at that location is therefore smaller. A context-free grammar has no such restrictions.

Table 1

Sample productions from each of the hand-designed grammars. These are chosen to illustrate the differences between each grammar, and may not be an exhaustive list of all of the expansions of any given non-terminal.

Context-free grammar CFG-S	
NP → NP PP NP CP NP C N det N adj N pro prop	
N → n adj N	
Context-free grammar CFG-L	
NP → NP PP NP CP NP C N PP N CP N C pro PP pro CP pro C prop PP prop CP prop C N det N adj N pro prop	
N → n adj N	
Flat grammar	
S → pro aux part	S → det n v n
S → adj n aux n prep det n	S → pro aux adj n comp pro v
Regular grammar REG-N	
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
Regular grammar REG-M	
NP → pro prop n det N adj N pro PP prop PP n PP pro CP prop CP n CP pro C prop C n C	
N → n adj N n PP n CP n C	
Regular grammar REG-B	
HP → pro prop n det N adj N pro HP prop HP n HP pro CP prop CP n CP pro C prop C n C prep HP prep adj adj HP to V _{inf}	
N → n adj N n HP n CP n C	

Table 2

Log prior, likelihood, and posterior probabilities of each hand-designed grammar for each level of evidence. Because numbers are negative, smaller absolute values correspond to higher probability. If two grammars have log probabilities that differ by n , their actual probabilities differ by e^n ; thus, the best hierarchical grammar CFG-L is e^{101} ($\sim 10^{43}$) times more probable than the best linear grammar REG-M.

Corpus	Probability	FLAT	REG-N	REG-M	REG-B	1-ST	CFG-S	CFG-L
Level 1	Prior	-99	-148	-124	-117	-94	-155	-192
	Likelihood	-17	-20	-19	-21	-36	-27	-27
	Posterior	-116	-168	-143	-138	-130	-182	-219
Level 2	Prior	-630	-456	-442	-411	-201	-357	-440
	Likelihood	-134	-147	-157	-162	-275	-194	-177
	Posterior	-764	-603	-599	-573	-476	-551	-617
Level 3	Prior	-1198	-663	-614	-529	-211	-454	-593
	Likelihood	-282	-323	-333	-346	-553	-402	-377
	Posterior	-1480	-986	-947	-875	-764	-856	-970
Level 4	Prior	-5839	-1550	-1134	-850	-234	-652	-1011
	Likelihood	-1498	-1761	-1918	-2042	-3104	-2078	-1956
	Posterior	-7337	-3311	-3052	-2892	-3338	-2730	-2967
Level 5	Prior	-10610	-1962	-1321	-956	-244	-732	-1228
	Likelihood	-2856	-3376	-3584	-3816	-5790	-3917	-3703
	Posterior	-13466	-5338	-4905	-4772	-6034	-4649	-4931
Level 6	Prior	-67612	-5231	-2083	-1390	-257	-827	-1567
	Likelihood	-18118	-24454	-25696	-27123	-40108	-27312	-26111
	Posterior	-85730	-29685	-27779	-28513	-40365	-28139	-27678

Table 3

Log prior, likelihood, and posterior probabilities of grammars resulting from local search. Because numbers are negative, smaller absolute values correspond to higher probability.

Corpus	Probability	FLAT	REG-N	REG-M	REG-B	1-ST	CFG-S	CFG-L
Level 1	Prior	-99	-99	-99	-99	-94	-133	-148
	Likelihood	-17	-19	-20	-19	-36	-26	-25
	Posterior	-116	-118	-119	-118	-130	-159	-173
Level 2	Prior	-630	-385	-423	-384	-201	-355	-404
	Likelihood	-134	-151	-158	-155	-275	-189	-188
	Posterior	-764	-536	-581	-539	-476	-544	-592
Level 3	Prior	-1198	-653	-569	-529	-211	-433	-521
	Likelihood	-282	-320	-339	-346	-553	-402	-380
	Posterior	-1480	-973	-908	-875	-764	-835	-901
Level 4	Prior	-5839	-1514	-1099	-837	-234	-566	-798
	Likelihood	-1498	-1770	-1868	-2008	-3104	-2088	-1991
	Posterior	-7337	-3284	-2967	-2845	-3338	-2654	-2789
Level 5	Prior	-10610	-1771	-1279	-956	-244	-615	-817
	Likelihood	-2856	-3514	-3618	-3816	-5790	-3931	-3781
	Posterior	-13466	-5285	-4897	-4772	-6034	-4546	-4598
Level 6	Prior	-67612	-5169	-2283	-1943	-257	-876	-1111
	Likelihood	-18118	-24299	-25303	-25368	-40108	-27032	-25889
	Posterior	-85730	-29468	-27586	-27311	-40365	-27908	-27000

Table 4

Log probabilities of the regular grammar constructed from scratch. As a comparison, the probabilities for the best other grammars are shown.

Corpus	REG-AUTO			Other best grammars (posterior)						
	Prior	Likelihood	Posterior	FLAT	REG-N	REG-M	REG-B	1-ST	CFG-S	CFG-L
Level 1	-105	-18	-123	-116	-118	-119	-118	-130	-159	-173
Level 2	-302	-193	-495	-764	-536	-581	-539	-476	-544	-592
Level 3	-356	-505	-841	-1480	-973	-908	-875	-764	-835	-901
Level 4	-762	-2204	-2966	-7337	-3284	-2967	-2845	-3338	-2654	-2789
Level 5	-1165	-3886	-5051	-13466	-5285	-4897	-4772	-6034	-4546	-4598
Level 6	-3162	-25252	-28414	-85730	-29468	-27586	-27311	-40365	-27908	-27000

Table 5

Log prior, likelihood, and posterior probabilities of each grammar type on the Epoch corpora, which reflect an age split. A hierarchical grammar is favored for all epochs, even on the first corpus (Epoch 0), corresponding to one hour of conversation at age 2;3.

Corpus	Probability	FLAT	REG-N	REG-M	REG-B	1-ST	CFG-S	CFG-L
Epoch 0 (2;3)	Prior	-3968	-1915	-1349	-1166	-244	-698	-864
	Likelihood	-881	-1265	-1321	-1322	-2199	-1489	-1448
	Posterior	-4849	-3180	-2670	-2488	-2433	-2187	-2312
Epoch 1 (2;3-2;8)	Prior	-22832	-3791	-1974	-1728	-257	-838	-1055
	Likelihood	-5945	-7811	-8223	-8164	-13123	-8834	-8467
	Posterior	-28777	-11602	-10197	-9892	-13380	-9672	-9522
Epoch 2 (2;3-3;1)	Prior	-34908	-4193	-2162	-1836	-257	-865	-1096
	Likelihood	-9250	-12164	-12815	-12724	-20334	-13675	-13099
	Posterior	-44158	-16357	-14977	-14560	-20591	-14540	-14195
Epoch 3 (2;3-3;5)	Prior	-48459	-4621	-2202	-1862	-257	-876	-1111
	Likelihood	-12909	-17153	-17975	-17918	-28487	-19232	-18417
	Posterior	-61368	-21774	-20177	-19780	-28744	-20108	-19528
Epoch 4 (2;3-4;2)	Prior	-59625	-4881	-2242	-1903	-257	-876	-1111
	Likelihood	-15945	-21317	-22273	-22293	-35284	-23830	-22793
	Posterior	-75570	-26198	-24515	-24196	-35541	-24706	-23904
Epoch 5 (2;3-5;2)	Prior	-67612	-5169	-2283	-1943	-257	-876	-1111
	Likelihood	-18118	-24299	-25303	-25368	-40108	-27032	-25889
	Posterior	-85730	-29468	-27586	-27311	-40365	-27908	-27000

Table 6

Proportion of sentences in the full corpus that are parsed by smaller grammars. 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.

Grammar	FLAT	REG-N	REG-M	REG-B	1-ST	CFG-S	CFG-L	
			% types					
Level 1	0.3%	0.7%	0.7%	0.7%	100%	2.4%	2.4%	
Level 2	1.4%	3.7%	5.1%	5.5%	100%	31.5%	16.4%	
Level 3	2.6%	9.1%	9.1%	32.2%	100%	53.1%	46.8%	
Level 4	10.9%	50.7%	61.2%	75.2%	100%	87.6%	82.7%	
Level 5	18.7%	68.8%	80.3%	88.0%	100%	91.8%	88.7%	
			% tokens					
Level 1	9.9%	32.6%	32.6%	32.6%	100%	40.2%	40.2%	
Level 2	21.4%	58.8%	61.7%	60.7%	100%	76.4%	69.7%	
Level 3	25.4%	72.5%	70.9%	79.6%	100%	87.8%	85.8%	
Level 4	34.2%	92.5%	94.3%	96.4%	100%	98.3%	97.5%	
Level 5	36.9%	95.9%	97.6%	98.5%	100%	99.0%	98.6%	

Table 7

Ability of each grammar to parse specific sentences. The complex declarative sentence “Eagles that are alive can fly” occurs in the Adam corpus. Only the context-free grammars can parse the corresponding complex interrogative sentence.

Type	In input?	Example	Can parse?						
			FLAT	REG-N	REG-M	REG-B	1-ST	CFG-S	CFG-L
Decl Simple	Y	Eagles can fly. (n aux vi)	Y	Y	Y	Y	Y	Y	Y
Int Simple	Y	Can eagles fly? (aux n vi)	Y	Y	Y	Y	Y	Y	Y
Decl Complex	Y	Eagles that are alive can fly. (n comp aux adj aux vi)	Y	Y	Y	Y	Y	Y	Y
Int Complex	N	Can eagles that are alive fly? (aux n comp aux adj vi)	N	N	N	N	Y	Y	Y
Int Complex	N	* Are eagles that alive can fly? (aux n comp adj aux vi)	N	N	N	N	Y	N	N

Table A1

Sample merges for context-free and regular grammars. Identical merges for right-hand side items were also used.

CFG merge example		REG merge example	
Old	New	Old	New
$A \rightarrow B C$	$A \rightarrow B F$	$A \rightarrow b C$	$A \rightarrow b F$
$A \rightarrow B D$	$F \rightarrow C$	$A \rightarrow b D$	$F \rightarrow d$
$A \rightarrow B E$	$F \rightarrow D$	$A \rightarrow b E$	$F \rightarrow g E$
	$F \rightarrow E$	$C \rightarrow g E$	$F \rightarrow e D$
		$D \rightarrow d$	
		$E \rightarrow e D$	