Running head: LEARNABILITY OF SYNTAX

The learnability of abstract syntactic principles

Amy Perfors

Department of Psychology

University of Adelaide

Joshua B. Tenenbaum

Department of Brain & Cognitive Science

Massachusetts Institute of Technology

Terry Regier

Department of Linguistics, Cognitive Science Program

University of California, Berkeley

Word count: 21988

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 address a version of this argument and show that, given typical child-directed speech and certain innate domain-general capacities, an 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, may suggest that human learners innately know a deep organizing principle of natural language, that syntax is organized in terms of hierarchical phrase structures.

In outline form, this is one version of 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 and various alternatives. Second, the learner has access to a Bayesian engine for statistical inference that 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 –that syntactic representations are organized in terms of hierarchical phrase structure – can be inferred by a learner with these capabilities but without a language-specific innate bias favoring this conclusion.

There have been many different framings of stimulus poverty questions over the years, and ours differs from both Chomsky's original framing and recent alternatives in some subtle ways that we will clarify over the course of this article. Berwick & Chomsky (2008; under review) have argued that much recent work on the poverty of the stimulus misses the original intention of the argument in generative linguistic theory. This may be true; it is certainly not for us to debate with Chomsky the original intentions of generative theory. Yet the stimulus poverty debate has taken on a larger life of its own in cognitive science more generally, and our goal here is to explore what we see as a basic issue at the heart of language learning – the origins of hierarchical phrase structure in syntactic representation – as an instance of the more general question of what kinds of structure must be innate in cognitive development. In our view, the argument about innateness is primarily about the role of *domain-specificity* in the learner's innate endowment. 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.

Note also that our focus is on the issue of what kind of knowledge must be assumed as an innate constraint on the learner's inductive hypotheses, rather than on what kind of representational machinery must be available to the learner. We are not arguing that a learner lacking a potential to represent hierarchical phrase structures can somehow acquire this potential; we accept here for the sake of argument the traditional view that a learning system can only learn structures that it is capable of representing. The question is whether a learner who is capable of representing grammars based on hierarchical phrase structure, as well as other kinds of structure, can infer that hierarchical phrase structure is indeed the best way to describe natural language syntax – without requiring specific innate knowledge that language is structured in this way. Some traditional nativist arguments equate these ideas: prior knowledge concerning the hypotheses that the learner considers takes the form of limitations on the class of hypotheses that the learner is capable of representing. This assumption makes sense if the learning mechanism is very simple – if learners can only select hypotheses based on their consistency with the observed data. By positing a more powerful Bayesian learning engine, we are able to relax this assumption and study how learners can select from among multiple *a priori* possible representational frameworks the one that best describes the data they observe.

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; Reali & 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 logic will motivate the form of our Bayesian analysis, but our focus is on one of the abstract questions that emerged based on the original example: the learnability of hierarchical phrase structure.

Before moving into the argument itself, we should highlight and clarify 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, 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 explore how and whether learners can infer deeper and more abstract principles of linguistic structure, such as the hierarchical phrase-structure basis for syntax. This principle (in conjunction with many other aspects of linguistic knowledge) supports an entire class of specific generalizations that include the auxiliary-fronting rule but also many other phenomena surrounding agreement, movement, and extraction. We take as data a corpus of child-directed speech and evaluate hypotheses about candidate grammars that could account for the corpus as a whole. Our findings suggest that it is vital to consider the learnability of entire candidate grammars holistically. While crucial data that would independently support 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.

Second, our approach offers a way to tease apart two fundamental dimensions of linguistic knowledge that are often conflated in the language acquisition literature. 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 symbolic representations like generative phrase-structure grammars. Different approaches to language acquisition correspond to different answers to these questions, which we can visualize along a two-dimensional space of possibilities (Figure 1). However, the best known approaches have explored only two corners in this space: domain-general learning accounts in the emergentist tradition, which seeks to explain language as arising from non-linguistic cognitive bases, have been studied primarily using simple representations that avoid explicit symbolic structure, such as *n*-grams or recurrent neural networks (e.g., Elman et. al., 1996; Rumelhart & McClelland, 1986; Reali & Christiansen, 2005). By contrast, structured symbolic representations have been explored primarily in the context of accounts based on innate language-specific knowledge that largely eschew general-purpose learning mechanisms (e.g., Chomsky 1965, 1980; Pinker, 1984). Few cognitive scientists have explored the possibility that explicitly structured mental representations might be constructed or learned via domain-general learning mechanisms. Despite this, there are compelling reasons to believe that the human mind has available both powerful general-purpose learning abilities and powerful representational capacities. Our framework offers a way to explore this relatively uncharted territory in the context of language acquisition. We will argue that domain-general learning of structured symbolic representations provides a valuable way to think about aspects of language acquisition (and potentially other areas of cognitive development) where data are sparse but the learner's generalizations are rich.

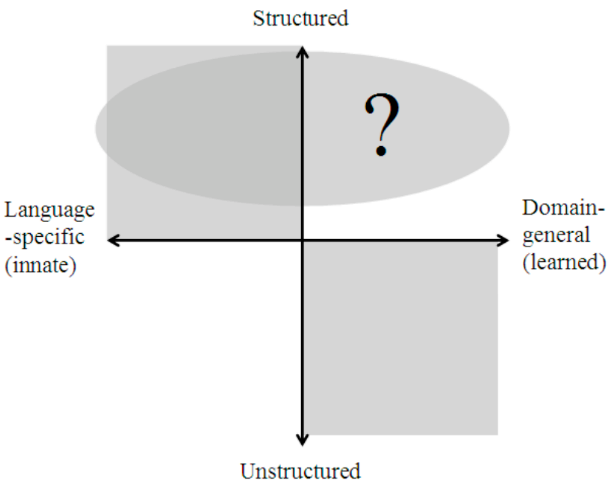

Figure 1: A schematic view of the theoretical landscape for language acquisition in cognitive science. The vertical axis reflects the nature of the representation. The horizontal axis reflects the source of inductive bias: "innate" and "learned" are in parentheses because they are often conflated with "language-specific" and "domain-general", which we suggest is closer to the real issue. The two most prominent approaches are represented by the two opposite shaded quadrants. We explore a different part of the landscape, represented by the shaded oval: assuming that mental representations of language are based on structured symbolic grammars (the upper half plane of the picture), we attempt to assess whether their form could be inferred based on domain-general learning mechanisms (the upper-right quadrant) or instead must be constrained by language-specific innate knowledge (the upper-left quadrant).

We hope that the position we lay out here may help to bridge the two more standard diagonally-opposed views in Figure 1. For an emergentist audience, we suggest that one may retain the core of emergentism – namely the focus on domain-general bases of language – while considering explicitly structured representations. Such a broadening of scope follows naturally from the observation that structured representations are themselves domain-general – family trees, organizational hierarchies, and plan hierarchies all rely on representations similar in some ways to those of language, but used for very different purposes. Moreover, we argue that domain-general learning mechanisms may suffice to determine what form of structured representation provides the best account of the child's linguistic input. Thus, our approach is fundamentally consistent in spirit with an emergentist view, even though the representations we consider are different from those traditionally adopted by many emergentists. At the same time, for a nativist audience, we hope that our adoption of structured representations helps to cast the claims of domain-generality in terms that are more recognizable and more obviously relevant to traditional analyses of what cognitive structures must be innate based on the linguistic input children receive and the final knowledge state they achieve.

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 or another noun 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 basic representations over which syntactic rules operate are defined in terms of abstract phrases, which may be nested hierarchically, and do not simply consist of linear sequences 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 thus a necessary feature of any possible hypothesis that the learner will consider?

This question has been the target of stimulus poverty arguments in the context of a number of different syntactic phenomena, but perhaps most famously auxiliary-fronted interrogatives in English (Chomsky, 1965, 1971, 1980; Crain, 1991; Crain and Nakayama, 1987; Laurence & Margolis, 2001; Lewis & Elman, 2001; Legate & Yang, 2002; Pullum & Scholz, 2002; Reali & Christiansen, 2005). Different authors have framed this challenge in different ways, so we first lay out the classic analysis of auxiliary fronting and then discuss variants, including ours.

There appears to be a strong structural regularity in English, relating simple declaratives like (1a) and (2a) with corresponding interrogative forms (1b) and (2b):

(1a) The man was hungry.
(1b) Was the man hungry?

(2a) The boy is smiling.
(2b) Is the boy smiling?

A traditional way to describe this regularity is in terms of “movement”: between corresponding declarative and interrogative forms, the auxiliary verbs *was* and *is* in (1a) and (2a) appear to move to the front of the sentences in (2a) and (2b). A language learner who grasps this regularity could extend it to produce and comprehend an infinite range of new utterances. These new cases may be more complex than the simple cases above, yet the extension of this syntactic pattern still appears straightforward:

(3a) The little girl in the red dress is smiling.
(3b) Is the little girl in the red dress smiling?

Consider (4a), however, in which the declarative form contains two identical auxiliary verbs.

(4a) The boy who is smiling is happy.
*(4b) Is the boy who smiling is happy?
(4c) Is the boy who is smiling happy?

Which is the correct way to form the corresponding interrogative: (4b), in which the first *is* from (4a) appears to move to the front, or (4c), in which the second *is* appears to move? Such cases suggest that there is not a unique logical way to characterize the relation between the simple declarative and interrogative forms in (1) and (2), which could be objectively deduced from these data treated only as sequences of words. Any regularity is an inductive inference, and there will be different ways of analyzing these sentences as linguistic objects that make different inductive hypotheses more or less natural.

This ambiguity presents a challenge for the language learner. A learner who analyzes sentences as linear structures of words, and who assumes that any linguistic rules would be consistent with that underlying structure, might characterize the patterns in (1) and (2) as something like (5), while one who analyzes sentences in terms of hierarchical structures of phrases might characterize these same patterns more like (6):

(5) Interrogatives can be formed by moving the leftmost auxiliary in the declarative to the beginning of the sentence.

(6) Interrogatives can be formed by moving the auxiliary in the main clause of the declarative to the beginning of the sentence.

These two ways of describing the observed patterns suggest different inductive generalizations for complex utterances with two or more occurrences of the same auxiliary: the inference in (5) suggests that (4b) would be correct, while the inference in (6) suggests that (4c) would be correct. We know that only (4c) is acceptable in English, and that the actual grammar of English follows rules that are more like (6) than (5), but how is a child to know which inference is correct?

One possibility is that simple observation could show the child that (5) is wrong and (6) is (more or less) right. If children learning language hear a sufficient sample of sentences like (4c) and few or no sentences like (4b), they might reasonably infer that English follows the pattern in (6) rather than the (5). The poverty of the stimulus argument focuses here. It has been argued that complex interrogative sentences such as (4c) do not exist in sufficient quantity in child-directed speech to make this inference. For instance, Chomsky (1971) suggests 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." In spite of this paucity of evidence, children three to five years old can form correct complex interrogative sentences like (4c) but appear not to produce incorrect forms such as (4b) (Crain & Nakayama, 1987; but see also Ambridge et al., 2008).

Another possibility is that the generalization expressed by (6) is somehow *a priori* simpler or better than that in (5). But it is hard to see how to justify such a preference, at least if one does not assume *a priori* that language has hierarchical phrase structure. If anything, a general-purpose learning agent who knows nothing specifically about human natural languages might take (5) to be the simpler induction, because it does not assume the existence of hidden objects (e.g., syntactic phrases) structured according to some unobservable relations (e.g., hierarchical phrase structures). If the correct generalization is not directly indicated by the data and is also not preferred on the grounds of a general inductive bias favoring simplicity, a natural conclusion is that children come equipped with some innate constraint or knowledge that biases them to induce the correct generalization rather than the incorrect one.

What is the nature of that innate bias? In the most famous framing of this argument, which has led to decades of intense controversy among a broad range of researchers, Chomsky (1980) appeared to suggest that it is an innate restriction on the kinds of representations that the language faculty can consider. We quote at some length from one of Chomsky's most accessible statements of this argument:

> The issue is, in this case, do we look at sentences in a linear or a hierarchical manner in order to carry out the induction? … There are cases in which people deal with properties like *leftmost* (they may regard an array of elements as linear and consider the physical arrangement of the elements), whereas there are other cases where people take into account all kinds of hierarchical structures in visual space or whatever. What we have to ask is what is the property in the initial state S_0 [of the language faculty] that forces us, in this specific linguistic case, always to go to the hierarchical abstract rule and always to neglect the more elementary linear physical rule? Several answers have been proposed to this question: the right one, I think, is the one which is implicit in the theory of transformational grammar, which in effect asserts that there is a notation available for describing linguistic rules that does not permit the formulation of the property *leftmost*… There is a very specific theory of representations in terms of which *follows the first noun-phrase* is a more elementary property than *leftmost*; but that happens to be a property of *this* specific concrete theory and not a consequence of any general theory of representations and structures. Of course this property has many consequences elsewhere; it has vast consequences for grammar, where, applied to other linguistic structures, the use of the category *leftmost*… should always be less accessible than properties like *follows the first noun-phrase*. This hypothesis is one that is rich in empirical consequences and to my knowledge true. (Chomsky, 1980, pp. 115-116)

Although this argument for innate language-specific constraints on syntactic rules is clearly stated, its interpretation is subtle and depends on what sort of rules we take as the basis for syntax, or the focus of interest for a cognitive theory of language. Generative linguistic theorists have long been interested in auxiliary-fronting as an example of syntactic movement. To explain these phenomena, they posit rules like those expressed informally in (5) and (6) that invoke explicit "fronting" or "raising" of words or phrases, as part of what a speaker knows when they grasp the structure of complex utterances such as (4a) or (4c). We, along with many cognitive scientists, are less sure about whether explicit movement rules provide the right framework for representing people's knowledge of language, and specifically whether they are the best way to account for how a child comes to understand and produce utterances like (4c) but not like (4b). We agree with the more general insight from linguistic theory, however, that only by defining syntactic rules (in whatever form they take) over hierarchical phrase structure representations, rather than linear representations of word order, is a child likely to be able understand that (4c) expresses a certain complex thought while (4b) expresses no well-formed thought. Hence our focus here is on the more basic question of how a learner can come to know that language should be represented in terms of hierarchical phrase structure rather than the more immediately obvious linear structure of words or sound categories

Our goal in this paper is to show that a disposition to represent syntax in terms of hierarchical phrase structure rather than linear structures need not be innately specified as part of the language faculty, but instead could be inferred using domain-general learning and representational capacities. The basis for this inference is implicitly anticipated by Chomsky's characterization of the phrase-structure hypothesis, in the lines quoted above, as "rich in empirical consequences" throughout language, not just for a single linguistic structure. While a child may not receive direct evidence about the correctness of a particular hierarchical phrase structure rule for analyzing some particular set of sentences such as the aux-fronting examples, there is vast indirect evidence for the general superiority of syntax with that structure throughout language. A learner who adopts a hierarchical phrase structure framework for describing the syntax of English will arrive at a much simpler, more explanatory account of her observations than a learner who adopts a linear framework.

We formalize this argument in Bayesian terms, where the "simpler, more explanatory" account becomes the more probable hypothesis. Linguists in the generative grammar tradition came to this inference early on. When one looks at the structure of natural language and considers the possibility of a framework with hierarchical phrase structure as opposed to a linear description, the superiority of the formal system quickly becomes apparent due to its "rich empirical consequences." Indeed, Berwick and Chomsky (2008; under review) have suggested recently that hierarchical phrase structure in syntactic representations was always taken for granted by generative theorists – but it is nonetheless a significant inductive leap. As Chomsky observes in the quotation above, many other domains of human activity besides language unfold sequentially. In some cases these activities are structured based on linear order properties, while in others they are not. A learner or a linguist at some point must decide to view language in one or the other of these ways, even if that decision occurs quickly, unconsciously and automatically. Our Bayesian analysis could apply just as well to formalizing this inductive leap inside the minds of human learners or linguists. Where our proposal differs from the standard view in generative linguistics is in the suggestion that children may receive sufficient language data to make this inference to hierarchical phrase structure, and hence need not have this assumption given as part of the innate state of a language acquisition device. We are not arguing that children necessarily *do* learn about the hierarchical phrase structure of syntax in this way, but rather that there exists a plausible learning framework which could allow them to do so from the data they observe.

A general formulation of the Poverty of the Stimulus argument

The PoS argument is, of course, not merely a point about auxiliary fronting in interrogative formation. We can formulate the general PoS argument in more precise and abstract terms as follows:

(7.i) Children show a specific pattern of behavior *B*.
(7.ii) A particular generalization *G* must be grasped in order to produce behavior *B*.
(7.iii) It is impossible to reasonably induce *G* simply on the basis of the data *D* that children receive.
(7.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 (6) 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 (6), 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 (5).

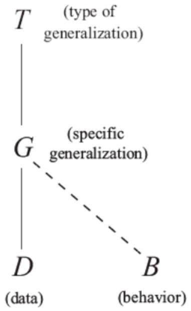

Figure 2. Graphical depiction of the standard Poverty of Stimulus argument. Abstract higher-level knowledge *T* is necessary to constrain the specific generalizations *G* that are learned from the data *D*, and that govern behavior *B*.

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:

(8.i) (Conclusion from above) Some abstract knowledge *T* is necessary.

(8.ii) *T* could not itself be learned, or could not be learned before the specific generalization *G* is known.

(8.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 (8.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 ability to correctly front the auxiliary in polar interrogatives 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:

(9.i) Children do not show the pattern of behavior *B*.

(9.ii) Behavior *B* is possible without having made the generalization *G*, through some other route from *D*.

(9.iii) It is possible to learn *G* on the basis of *D* alone, without the need for some more abstract knowledge *T*.

In the case of auxiliary fronting, one example of the first response (9.i) is the claim that children do not in fact always avoid errors that would be best explained under a linear rule rather than a hierarchical rule. Although Crain & Nakayama (1987) demonstrated that children do not spontaneously form incorrect complex interrogatives such as (4b), 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 phrase-structure grammar but might be consistent with a linear rule. Additionally, recent research by Ambridge et. al. (2008) suggests that 6 to 7 year-old children presented with auxiliaries other than *is* do indeed occasionally form incorrect sentences like (4b), such as "Can the boy who run fast can jump high?"

A different response (9.iii) accepts that children have inferred the correct hierarchical rule for auxiliary fronting (6), but maintains that the input data is sufficient to support this inference. If children observe sufficiently many complex interrogative sentences like (4c) while observing no sentences like (4b), then perhaps they could learn directly that the hierarchical rule (6) is correct, or at least better supported than simple linear alternatives. The force of this response depends on how many sentences like (4c) 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 corpora[1] 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 contexts (Landauer & Dumais, 1997; Regier & Gahl, 2004; Reali & Christiansen, 2005; Foraker et al., 2009).

Indirect evidence also plays a role in the second type of reply, (9.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

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

sentences like (4c) but not (4b) – even without having learned any grammatical rules like (5) or (6) at all. Perhaps the data, while poor with respect to complex interrogative forms, are rich in distributional and statistical regularities that would distinguish (4c) from (4b). If children pick up on these regularities, that could be sufficient to explain why they avoid incorrect complex interrogative sentences like (4b), 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. (2008) for a critique). They demonstrated that simple bigram and trigram models applied to a corpus of child-directed speech gave higher likelihood to correct complex interrogatives than to incorrect interrogatives, and that the *n*-gram models correctly classified the grammaticality of 96% of test sentences like (4b) and (4c). 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 – despite their interesting linguistic behavior once trained.

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 much of contemporary linguistics. PoS arguments begin with the assumption – taken by most linguists as self-evident – that language does have explicit hierarchical phrase 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 PoS arguments more directly on their own terms.

One place where our analysis does make a significant simplification (relative to the standard linguistic treatment of aux-fronting and related phenomena) is that we – like Reali & Christiansen (2005) and Lewis & Elman (2001) – do not attempt to explain these phenomena in terms of movement or transformation rules. Chomsky's (1980) formulation of "linear" and "hierarchical" hypotheses for forming complex interrogatives, (5) and (6) respectively, framed these as alternative rules for moving elements of a base declarative form, and the question was whether these rules should be defined over a hierarchical phrase-structure analysis or the linear sequence of words in the declarative form. Instead, as we explain above, we focus on the more basic question of how and whether a learner could infer that representations with linear or hierarchical phrase structure provide the best way to characterize the set of syntactic forms found in a language. We see this question as the simplest way to get at the essence of the core inductive problem of language acquisition posed in the generative tradition. It is also relevant to the original phenomenon we began with: the best hierarchically phrase structured grammars we find do indeed generate correct aux-fronted complex interrogative forms like 4(c) and not incorrect forms like 4(b), while the best linear[2] grammars do not. An important direction for future work would be to link our learnability analysis more tightly to standard syntactic analyses, by extending it to grammars based on explicit movement rules or other means to the same end. Even without doing so, it seems a reasonable premise that any such extension would naturally involve rules defined over the constituents of the grammar, and thus the identification of those constituents – the problem we address here – is important and relevant: if the hierarchical nature of phrase structure can be inferred, then any reasonable approach to inducing rules defined over constituent structure should result in appropriate structure-dependent rules.

Overview of results

We present two main results. First of all, we demonstrate that a learner equipped with the capacity to explicitly represent both linear and hierarchical phrase-structure grammars – but without any initial bias to prefer either in the domain of language – can infer that the hierarchical phrase-structure 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 may be possible to acquire domain-specific knowledge about the form of structured representations via domain-general learning mechanisms operating on data from that domain. Secondly, we show that the hierarchical phrase-structure grammar favored by the model – unlike the other grammars it considers – succeeds in one important auxiliary fronting task, 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

[2] The terminology here is potentially confusing. Throughout this paper we use the term “linear grammar” in the informal sense employed during much of this stimulus poverty debate, refer ring to grammars that do not incorporate hierarchical phrase structure. This usage should be distinguished from the technical concept of *linear language* in formal language theory. In that technical sense, the linear (non-hierarchical) grammars considered in this paper correspond to the class of *regular languages* rather than the class of linear languages.

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.

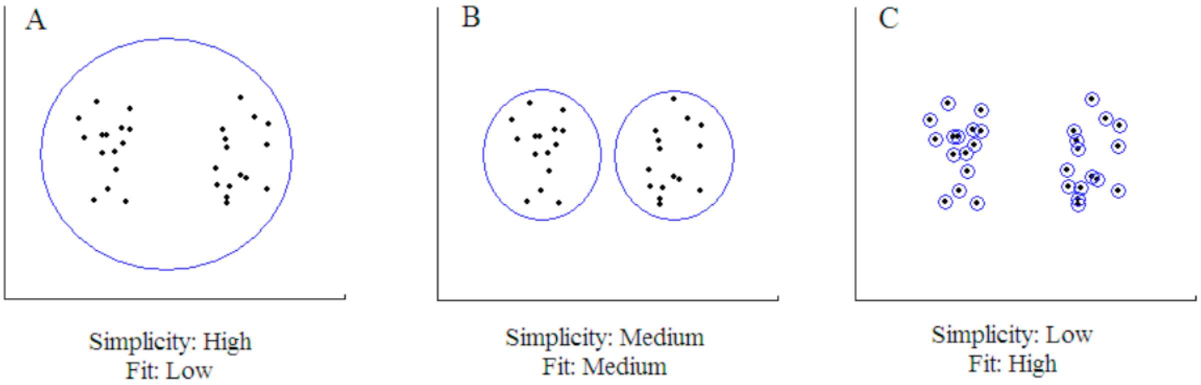

Figure 3. Fitting models of different complexity (represented by the circles) to a dataset (the points). The complexity of a model reflects the number of choices necessary to specify it: the model A can be fully specified by the location and size of only one circle, while model C is more complex because 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 phrase structure can be learned without an innate language-specific bias, we do not argue 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 & Tenenbaum, 2008; 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 framework[3] 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:

$$p(G,T|D) \propto p(D|G)p(G|T)p(T). \tag{1}$$

[3] 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) can be inferred from the data.

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).

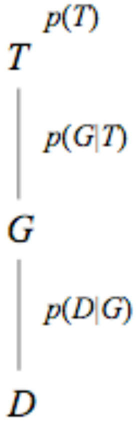

Figure 4. A hierarchical Bayesian model for assessing Poverty of Stimulus arguments. The model is organized around the same structure as Figure 2, but now each level of representation defines a probability distribution for the level below it. Bayesian inference can be used to make inferences at higher 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 & Vitànyi, 1997; Chater & Vitànyi, 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 it demonstrates that learning a grammar in a probabilistic sense is possible if the learner makes certain assumptions about the statistical distribution of the input sentences (Horning, 1969; Angluin, 1988). 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 & Vitànyi, 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 phrase 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 language learning problems (e.g., Eisner, 2002; Johnson & Riezler, 2002; Light & Grieff, 2002; Klein & Manning, 2004; Alishahi & Stevenson, 2005; Chater & Manning, 2006; Liang et al., 2007; Rose Finkel et al., 2007). 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 is based either on small fragments of (sometimes artificial) corpora (e.g., Dowman, 2000; Alishahi & Stevenson, 2005; Clark & Eyraud, 2006) or on corpora of adult-directed speech (e.g., Eisner, 2002; Klein & Manning, 2004). Neither is ideal for addressing learnability questions. 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 (at least not to the same extent). 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 search effectively over the joint space of *G* and *T* for grammars that maximize a 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. In particular we do not consider the computational tractability of searching for the best-scoring grammars, which could be seen as another side of learnability. Setting this challenge aside allows us to focus with more clarity on those aspects of learnability that classic PoS arguments address: claims about what data might be sufficient for learning, or what language-specific prior knowledge must be assumed in order to make learning possible. We consider the limitations and implications of this "ideal learnability" approach more fully in the discussion.

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 best captured by a grammar with hierarchical phrase-structure, if one is not innately biased to consider only such 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.[4] 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.[5] 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 context-free grammars will be more preferred as the input grows more

[4] 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 (*vi*), 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.

[5] 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).

complex. The final corpus contains 21671 individual sentence tokens corresponding to 2336 unique sentence types, out of 25755 tokens in the original corpus.[6]

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 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.[7]

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 ($\Sigma_G\, p(T,G|D)$) or by choosing the best G of that type (max_G $p(T,G|D)$). Ultimately it is the specific grammar G that governs how the learner understands

[6] 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.

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

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 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. For decades, they have often been treated as a first approximation to the structure of natural language. 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 standard 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 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 grammars without hierarchical phrase structure. The first is based on regular grammars, also known as finite-state grammars. Regular grammars were originally proposed by Chomsky (1957) as the "minimal linguistic theory that merits serious consideration." They are "the simplest type of grammar which, with a finite amount of apparatus, can generate an infinite number of sentences", although they do so 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 a 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 (and is very similar to a standard unigram model). Though these three approaches may not capture exactly what was originally envisioned as grammars without hierarchical phrase structure, 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. Among the context-free grammars, CFG-S is smaller but fits less precisely, while CFG-L is larger and fits the full corpus with more precision. The three regular grammars also span the range of simplicity and goodness-of-fit: REG-B is the smallest, with the least precise fit; REG-M occupies a middle ground; and REG-N is the largest and most precise. All of the grammars are described more precisely in Appendix A, and Table 1 contains sample productions from each.[8]

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

[8] All full grammars, corpora, maximum-likelihood parses, and perplexity values (corresponding to all likelihood calculations) may be found at http://www.psychology.adelaide.edu.au/personalpages/staff/amyperfors/research/cognitionpos/

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. This type of search for context-free grammars presents a difficult computational problem, and current unsupervised 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 describe the search over regular grammars in Appendix A.

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 probabilistic generative model under which each grammar is selected from the space of all grammars of that type. This generative model is itself a kind of grammar, but at a higher level of abstraction – a meta-grammar, or grammar for generating grammars for syntax (c.f., Feldman, Gips, Horning & Reder, 1969). More complex grammars are those that result from making more (and more specific) choices under this generating process. This method of assigning priors to models based on 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 – hence requiring more choices to be made. 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 under the meta-grammar. Appendix B contains specific details about the generative process that defines the simplicity metric over grammars. It is important to emphasize that this measure is not in general equivalent to a simple count of the number of free parameters or independent choices needed to specify a model. Our prior also takes into account how free each choice is. The less restrictive any choice is, the lower the probability of making that choice in any particular way, and hence the lower the prior probability of the resulting model.

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 5). 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.

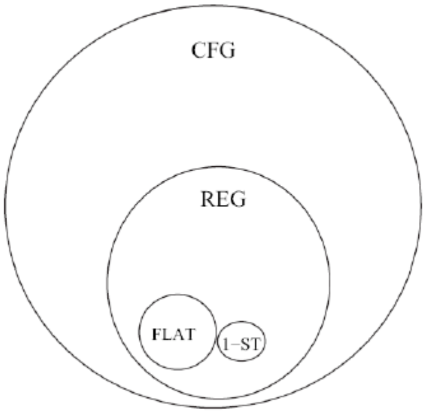

Figure 5. Venn diagram depicting the relation of the grammar types *T* to each other. The set of context-free grammars contains regular, flat, and one-state grammar types as special cases. Flat and one-state grammars are themselves special cases of regular grammars.

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 & Vitànyi, 1997; Chater & Vitànyi 2003, 2007). These have been useful for the induction of specific context-free grammars *G* (e.g., Dowman, 1998), and reflect a similar intuition of simplicity.

Scoring the grammars: likelihood.

The likelihood $p(D|G)$ can be defined straightforwardly for any probabilistic context-free grammar or regular grammar by assuming that each sentence in the corpus is generated independently from the grammar. The likelihood assigned to a grammar based on a corpus of sentences 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.

In the context of evaluating the fit of a grammar to a language corpus, an ideal learner must also solve the problem of parsing each sentence in the corpus as an "inner loop" in computing likelihood. Each possible parse of a sentence under a grammar can be assigned a probability that is the product of the probabilities of the production rules used to generate that parse. A grammar fits a sentence tightly if it the sentence results from one or more high-probability parses generated by the grammar, that is, if the sentence can be parsed using relatively few productions and relatively high probability productions.

We defer most technical details about how likelihood is calculated to Appendix B, but it is worth noting here two general features of how likelihood functions in our framework. First, the probabilistic preference for the most specific or tightest fitting grammar consistent with the observed data is related to 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.

Second, while the classical approach in probabilistic grammar induction (e.g., Feldman et al., 1969; Stolcke & Omohundro, 1994; Manning & Schütze, 1999) treats each individual sentence token in the corpus as a distinct sample from the grammar, we view it as an open question whether sentence tokens should be viewed in this way or not. One reason to evaluate the appropriateness of the classical approach is that context-free grammars with production probabilities based on sentence token frequency generate statistical distributions of sentences that differ systematically from the well-attested power law distributions characteristic of language at multiple scales (e.g., Zipf, 1932; see Briscoe, 2006, for a more recent overview[9]). Another reason is that it seems plausible that many common sentences – such as “How’s it going?” or “See you around” – are not generated directly from the grammar on each utterance. Instead, it is possible that once generated and uttered a few times, such sentences are cached away in memory as full-sentence exemplars, and may be produced again as unanalyzed wholes when context is appropriate. Thus, a model that presumes that each sentence token is emitted independently based only on production probabilities in a PCFG may be inappropriate.

We address this issue by making use of a version of the *adaptor grammar* framework of Goldwater et al. (2006) and Johnson et al. (2007). This framework captures the intuition that a sentence may be produced either by generating the sentence directly from the grammar, or by calling up a sentence exemplar that had earlier been generated from the grammar and stored in memory. Accordingly, the framework assumes a language model that is divided into two components. The first component, the generator, assigns a probability distribution over the potentially infinite set of syntactic forms that are accepted in the language. The generator can naturally take the form of a traditional probabilistic generative grammar, such as a PCFG. For simplicity we sometimes refer to this component as the “grammar.” The second component, the adaptor, produces a finite observed corpus through a nonparametric stochastic process that combines draws from the generating grammar with draws from a stored memory of previously produced sentence forms – thus interpolating between types and tokens. The adaptor component is primarily responsible for capturing the precise statistics of observed utterance tokens, and unlike simpler traditional probabilistic grammars, it can account naturally for the characteristic power-law distributions found in language.

We begin by evaluating performance in the one-component model (i.e. grammar without the adaptor) based on sentence types rather than tokens, and find that a hierarchical grammar is preferred. We then evaluate performance in the one-component model (again without the adaptor) based on sentence tokens, and find that a non-hierarchical grammar is

[9] Although the classic work exploring Zipfian distributions in language did not look at the distribution of sentence tokens of the form we evaluate here, we did observe a Zipfian distribution of such tokens in our own corpus.

preferred. Finally, we adjudicate between these two conflicting results by evaluating performance within the full two-component model, which includes the adaptor as well as the grammatical generator. Results show (1) that the posterior probability is higher for a type-based than for a token-based analysis, and (2) that a hierarchical phrase-structure grammar scores higher than any other grammar under the full two-part model. This provides statistical validation for both an analysis that tends toward being type-based, and for the overall finding that grammars with hierarchical phrase structure provide the best overall account of the corpus. We consider implications of this result in the discussion.

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 type-based 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 grammar with hierarchical phrase structure 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 context-free 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 context-free 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 finite-state (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.

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 grammars to grammars with hierarchical phrase structure? 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. Many linear grammars found by the local 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.

The results of the local search, shown in Table 3, 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 type-based corpus, compared to a variety of plausible linear grammars. Though the ability of the hierarchical phrase structure grammars to generate a higher variety of sentences from fewer productions typically results in a lower likelihood, this compression helps dramatically in the prior. This type of 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.

Sentence tokens vs sentence types

So far we have evaluated the grammars only on type-based corpora, but we conceived of the likelihood as emerging from 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. Defining the likelihood in this way is not standard in computational linguistics, but it is a principled choice motivated by the recent computational work of Goldwater et al. (2006) and the standard practice in theoretical linguistics, where grammars are evaluated based on how well they account for which sentences occur, rather than their frequency distribution. It is nevertheless useful to explore precisely what the effect of making this choice is. We therefore evaluated the best grammars of each class found so far on a corpus of sentence tokens rather than types.

Interestingly, the linear grammars were overwhelmingly preferred relative to the context-free grammars (*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). Thus the likelihood is weighted relatively more strongly relative to the prior (which does not change); this works against the context-free grammars, which overgeneralize more.

This result suggests that if the hierarchical phrase structure of syntax is to be inferred from observed data based on Bayesian inference with probabilistic grammars, the learner may need to have some sort of disposition to evaluate grammars with respect to type-based rather than token-based data. It is unlikely that such a disposition emerges from a general insensitivity to token frequencies: there is extensive psycholinguistic and developmental evidence demonstrating that people are sensitive to quantitative frequency variations in a wide variety of contexts. A more plausible interpretation – on both behavioral and computational grounds – is that grammar induction is based on more sophisticated grammar models (such as the adaptor framework) which do not treat each sentence as an independent statistical sample from the grammar. In such a framework, quantitative variation in token frequencies is not ignored, but is separated from the principles of the grammar that generate acceptable forms.

We can evaluate the reasonableness of this interpretation in more detail by explicitly calculating the probability of the corpus under the full two-component adaptor model, rather than the one-component model considered thus far. The two-component adaptor model requires summing over all possible interpolations between type-based and token-based input, and calculating the probability of the corpus given the grammar for each specific interpolation.

What does it mean to interpolate between type-based and token-based input? The central metaphor underlying the adaptor model conceives of the input as consisting of "tables" in a restaurant, and "customers" of the restaurant as corresponding to specific sentence tokens. In a fully type-based analysis, all of the sentence tokens of a given type are seated at the same table, resulting in 2336 tables total. For instance, all six sentences of the form *det n v n* would be on one table, meaning that the input to the model consisted only of one *det n v n* rather than six. By contrast, in a fully token-based analysis, each customer would be seated at their own table, resulting in six tables each corresponding to sentences of the form *det n v n*, and 21671 tables in total. Interpolating between these analysis would correspond to different ways of assigning the six sentence tokens of the same type to more than one table but less than six. In our example of the form *det n v n*, this might correspond to having two tables, each with three customers; three tables, one with four customers and two with one; or any other possible distribution of tokens to tables.

The essential idea is that when language is produced, sometimes a sentence is generated directly from a grammar, and sometimes it is generated from the memory cache of previous sentences that have been spoken. If a sentence is generated directly from a grammar, it corresponds to a table; it is relevant for a learner seeking to identify the particular grammar that did the generating. If it is generated from the memory cache, this corresponds to one of the customers (tokens) sitting at an existing table; since it was not generated from the grammar directly it would be sensible for a learner to disregard this sentence token when seeking to identify the grammar. Of course, sentences don't come clearly labeled as being generated from either the grammar or the memory cache, so the job for the learner is to figure out how to optimally distribute tokens among tables in such a way as to maximize the probability of the observed sentences given a grammar. We assume that the learner has a prior that favors an intermediate analysis in which there are more tables than a fully type-based analysis would imply, but fewer tables than a fully token-based analysis would.

It is clear that for a corpus with 21671 tokens and 2336 types, there are millions of ways of distributing tokens to tables; thus, interpolating between types and tokens by evaluating the probability of each possible assignment is computationally intractable, for much the same reason that it is computationally intractable to effectively search the space of all context-free grammars: the space is very large, and has many local maxima. However, as in our previous analysis, we can address this question by searching for an approximate best interpolation. As before, this may be accomplished by searching the space of possibilities from multiple starting points, with several questions in mind. First, is the single best interpolation closer to a type-based or a token-based corpus? Second, does the search – which tends to move in the direction of increasing overall probability – always tend to move toward more type-based analyses, regardless of its starting point? The logic is as follows: if, in maximizing the joint probability of the grammar, corpus, and level of interpolation between types and tokens, the overall highest-probability outcome is one that favors grammars with hierarchical phrase structure, this is evidence that a rational learner might be able to realize that these grammars offer a better explanation of the data than grammars without.

As detailed more fully in Appendix B, we performed multiple partial searches of the space of possibilities, interpolating between a fully type-based corpus and a fully token-based corpus. The results remain consistent with the one-component type-based results reported previously: the context-free grammars still had the highest posterior probability. Moreover, search steps that made the analysis more type-based were more likely to improve the overall probability than search steps that made it more token-based. Although these searches are not fully comprehensive, these results are coherent and suggestive. The best-performing grammars are still the ones with hierarchical phrase structure, and there is reason to believe that evaluating grammars on the basis of types rather than tokens is more appropriate. We therefore focus on type-based analyses for the remainder of the results.

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 context-free grammars are 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 context-free 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.

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

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

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 context-free 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 phrase structure grammars provide the best description of the language's syntax, given only minimal exposure to child-directed speech.

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.

A more stringent test of generalization is to evaluate performance with respect to completely novel corpora. To that end, we selected the final file of the Sarah corpus (Brown, 1973) of the CHILDES database (MacWhinney, 2000); we chose this because, since Sarah was 5;1 at the time, this presented a more stringent test of generalization than if she were younger. Processing the corpus in the same way as the Adam corpus (i.e., replacing lexical items with syntactic categories and removing the more complex sentence types) results in a dataset with 156 sentence types corresponding to 230 sentence tokens. The *Level 6* CFG-L parses the highest percentage of sentence types in that corpus (94.2%), followed closely by CFG-S (93.6%), with REG-M (91.7%), REG-B (91.0%), and REG-N (87.2%) trailing. Although the magnitude of these differences is not large, they follow the same pattern as we found on the Adam corpus.

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 *un*grammatical English sentences than the regular grammars? Of the 191 ungrammatical sentence types excluded from the full Adam 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. 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 phrase-structure 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. 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.[11]

The superior generalization ability of the context-free 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 context-free 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 phrase structure 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.

Although our best context-free grammars perform correctly on sentences like those in Table 7, as Berwick & Chomsky (under review) point out, they do not capture the entire auxiliary system in English. This is largely because of two main simplifications in our

[11] It is interesting that none of the best grammars found – either regular or context-free – can parse the incorrect complex interrogative forms (such as "Are eagles that alive can fly?", or (4b) above) that traditional PoS arguments imagine would be the first hypotheses of an unbiased learner. To understand why not, it is useful to compare the performance of the best grammars found to the same grammars supplemented with the minimal number of additional productions necessary to parse these incorrect forms. When we add these productions to the grammars (whether regular or context-free), they are not used to parse any of the sentences in the full training corpus, and the inside-outside algorithm consequently prunes them away (by setting their probabilities to zero). If these productions are forced to be present with nonzero probabilities, the resulting grammars have lower prior probability (due to the additional productions) as well as lower likelihood (due to predicting sentences not found in the corpus), and thus will be dispreferred relative to similar grammars that lack those productions.

modeling: 1) our selection of grammars based on fit to a small corpus of child-directed speech, which does not contain all of the many types of sentences necessary to support the entire auxiliary system; and 2) our choice of part-of-speech encoding, which collapses certain distinctions (for instance, modal auxiliaries, *do*, and *be* forms are all coded as *aux*) that would be needed to account for the full auxiliary system. These simplifications enabled a tractable exploration of the question we are most concerned with – the learnability of hierarchical phrase structure in language. Full mastery of the auxiliary system or any other specific aspect of syntax is not the goal. It would be of interest to conduct a similar study using more sophisticated grammars capable of better approximating the full syntax of English, but this is beyond the scope of our current techniques and so we leave it for future work.

Linguistic adequacy

Although we have shown that the grammars favored by the model also generalize more appropriately according to a variety of measures, it remains possible that the grammars with the highest posterior probability might assign linguistically implausible structures. We therefore compare the accuracy of each of our grammars on a "gold standard" parsed corpus, where accuracy is measured by F-score (which reflects the harmonic mean of precision and recall of the grammars; see Manning & Schütze, 1999). As our gold standard corpus we selected the child-directed speech from the final file of the Eve corpus (Brown, 1973) of the CHILDES database (MacWhinney, 2000), which contains one hour of speech directed to Eve at age 2;3. Processing the corpus in the same way as the Adam and Sarah corpora were processed results in a corpus of 224 sentence types corresponding to 415 sentence tokens.

We chose this corpus because it has previously been augmented with syntactic dependency annotations (Sagae et al., 2007) which were then converted to constituency annotations (Borensztajn et al., 2008), and thus provides an objective standard of comparison for our grammars. Upon inspection of the parses, however, many of them seemed incorrect; we therefore also had an independent rater with a PhD in linguistics, blind to the nature of any of our grammars, provide hand-annotated parses. We report the accuracy of the maximum-likelihood parse yielded by each of our *Level 6* grammars according to both of these standards.

The results, shown in Table 8, demonstrate that the most accurate grammars according to both standards are the two context-free grammars, and the one favored by the model (CFG-

L) is the most accurate of all. All of the regular grammars yield the same parses as the right-branching baseline, but have lower accuracy because they fail to parse the entire corpus. As with the Sarah corpus, the CFG-L parses the highest percentage of sentence types (96.5%), followed closely by CFG-S (95.5%), with REG-M (95.1%), REG-B (95.1%), and REG-N (92.4%) trailing. The improvement in accuracy on CFG-L and CFG-S is not solely due to the fact that they successfully parse more sentences, however; when accuracy is calculated only including those sentences that each grammar can parse, the relative performance of each grammar remains the same (F-scores on hand-parsed corpus: CFG-L = 91.8; CFG-S = 90.9; REG-B = 89.0; REG-M = 89.0; REG-N = 89.2; on automatically-parsed corpus: CFG-L = 68.1; CFG-S = 67.2; REG-B = 67.4; REG-M = 67.4; REG-N = 67.4).[12]

These results suggest that our grammars are not implausible, and that the grammars which score best under our Bayesian scoring criterion also perform best in precision and recall. However, because our corpora are child-directed speech, the sentence structure is fairly simple and precision/recall may not be the most diagnostic analysis: all of the grammars have fairly high accuracy scores, and all of them are fairly similar. This includes the right-branching baseline grammar, since many of the sentences in the corpus are in fact right-branching. For the cases where the correct analysis is not, the context-free grammars perform better; this suggests that our analysis converges with more traditional measures of successful learning. That said, because of their overall simplicity, we do not propose any of our grammars as a serious model of English; the purpose of this analysis was simply to demonstrate that the grammars which were favored by the Bayesian scoring criterion are also more linguistically adequate by other measures as well.

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 form

[12] The reason the F-scores for the regular grammars are not identical to the F-scores for the right-branching baseline, despite the fact that all of the regular grammars impose parses equivalent to that of the right-branching baseline, is because some of the sentences that the regular grammars failed to parse had a lower-than-average precision/recall for the right-branching baseline. They therefore brought the average accuracy for the baseline down relative to that of the regular grammars.

grammatically correct English interrogatives, 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 phrase 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 extensive means performed less well than the hierarchical phrase-structure grammars we examined. 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 (e.g., Kearns & Valiant, 1989).[13] 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 capacity to represent various types of grammars, including grammars with linear structure and grammars with hierarchical phrase structure. 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 in sequential structures – linguistic or non-linguistic – is necessary to even ask the questions we consider here. Our analysis also assumes that the learner represents the different grammar classes *as* different grammar classes, 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 linear grammars more heavily than they would be favored if we treated all grammars as instances of a single general type. (See Appendix B for details).

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).

[13] 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.

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 the centrality of explicit symbolic structure in formal linguistics and computational 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 to grammar induction and parsing 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 set of assumptions concerns the way we represent the input to learning. We have given our model a corpus consisting of sequences of syntactic categories rather than sequences of lexical items.[14] Working with syntactic categories rather than individual lexical items 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, 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.,

[14] 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.

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 phrase-structure grammars continue to score better than linear grammars if the input to learning consists of automatically labeled part-of-speech tags rather than hand-labeled syntactic categories.

Is there a reasonable psychological interpretation for the two-component adaptor grammar framework? The framework corresponds to assuming that language users can generate the syntactic forms of sentence tokens either by drawing on a memory store of familiar sentence types, or by consulting a deeper level of grammatical knowledge about how to generate the infinite variety of acceptable syntactic forms in the language. Any sentence type generated by the former system would originally have been generated by the latter, but this framework suggests that speakers may not need to consult their rule-based grammatical knowledge for every sentence they utter or comprehend. For example, a common greeting such as “How’s it going?” might be cached away as a unit such that it need not be generated afresh from first grammatical principles on every utterance. One might imagine other factors that might affect the frequency of what sentences are spoken that have nothing to do with the grammar itself, including the conversational context, the nature of the interlocutor, or salience in memory. A sensible learner might want to separate the factors that affect the frequency of observed sentence tokens from the factors that guide which sentences are grammatical in the first place. The adaptor grammar framework provides one way to do that. Our work suggests that if human learners, like our model, are capable of evaluating whether type-based or token-based analyses are *themselves* more appropriate for a given problem, they might rationally decide to favor a more type-based analysis when deciding among grammars (not necessarily for other aspects of language acquisition). Token frequencies might still be quite useful for driving aspects of the acquisition problem that we have not considered here, such as the formation of syntactic categories (e.g., Borovsky & Elman, 2006).

Would a disposition to evaluate grammars within a two-component adaptor-grammar-like framework, or based on type data only, constitute a language-specific or domain-general disposition? It is difficult to say, but the conceptual underpinnings of the adaptor grammar framework are consistent with a domain-general interpretation, emerging due to memory constraints or other cognitive factors. Indeed, one novel prediction of this work, which may be empirically evaluated in artificial grammar learning experiments, suggests that people should evaluate artificial grammars with respect to sentence types rather than tokens. Determining whether a disposition to do so exists – and, if so, whether it is language-specific or not – is a question for future work.

While the preference for linear grammars given token-based input does not change our overall positive conclusion about the learnability of hierarchical structure, it does highlight one set of assumptions that would lead to a different conclusion. Our adaptor grammar simulations decide in favor of the results that were closer to type-based rather than token-based – and thus in favor of learnability rather than unlearnability of hierarchical structure – but since they were based on approximations rather than an exhaustive search of the entire space, they should be revisited as technology improves. Further probing of the assumptions implicit in the modeling framework about the roles of grammar and memory in production are also questions for future work.

In any case, all of the assumptions made in our analysis involve either abilities that are plausibly domain-general, or language-specific representations that are distinct from the knowledge of hierarchical phrase structure. 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 phrase 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 phrase-structure grammars in. The learned preference for grammars with hierarchical phrase structure 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 linear rule. 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 phrase 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 asserts that it is implausible to expect a reasonable learner to arrive at the correct forms for very rare sentence types such as complex aux-fronted interrogatives, given realistic language input but no language-specific innate bias towards hierarchical syntactic structure. All that is required to respond to such a claim is to demonstrate that some reasonable learner could in fact do 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 linguistic rules are defined over hierarchical phrase structures.

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 the best 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 likelihood of the context-free grammar relative to the others (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 specifically linguistic bias towards syntactic systems organized around hierarchical phrase structures: our point is that the classic PoS argument is not a good reason to believe that they do. Another answer is that children may have an innate *cognitive* bias towards hierarchical phrase structure: 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 phrase 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, or to view grammar generation as a two-stage process as in the adaptor grammar framework, is necessary to explain children's acquisition patterns. Finally, it is also still possible that there are no biases in this direction at all – cognitive or linguistic – 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.

Recent characterizations of an innate language faculty have concentrated on recursion in particular (Hauser et. al., 2002; Pinker & Jackendoff, 2005). 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).[15] 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 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. Many 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

[15] See Perfors et al. (2010) for a more detailed exploration of this issue.

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; Foraker et al., 2009). This point is also 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 missing from some of the 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 *T* 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 suggests 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. If an ideal learner can make this inference from actual child-directed speech, it is 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 learn more about these two fundamental issues. There are also good reasons to believe that the acquisition of hierarchical phrase structure in syntax ultimately depends on broader cognitive capacities we have not considered here, such as the hierarchical structure of thought. Still, we have offered a positive and plausible "in principle" alternative to the classic negative "in principle" poverty-of-stimulus arguments for innate knowledge of hierarchical phrase structure in syntax.

More generally, we see this work as an example of a new and productive approach to old questions of innateness in cognitive science. By working with sophisticated statistical inference mechanisms that can operate over structured representations of knowledge such as generative grammars, and by evaluating these models on real data representative of children's experience in the world, we can more rigorously explore a relatively uncharted region of the theoretical landscape: the possibility that genuinely structured knowledge can be genuinely learned, as opposed to the classic positions of nativism (structured but unlearned knowledge) or empiricism (learned but unstructured knowledge, where apparent structure is merely implicit or emergent). 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, Lila Gleitman, 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, Charles Kemp, and Susanne Gahl. Special thanks to Virginia Savova for hand-parsing the Eve corpus and Gideon Borensztajn for sharing the constituency-annotated parses for that corpus. 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. (2008). Is structure dependence an innate constraint? New experimental evidence from children's complex-question production. *Cognitive Science 32*, 222-255.

Angluin, D. (1988). Identifying languages from stochastic examples. Tech report, Yale University. DCS/RR-614

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. (1985) *The Acquisition of Syntactic Knowledge*. Cambridge, MA: MIT Press.

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

Berwick, R., & Chomsky, N. (2008). 'Poverty of the stimulus' revisited: Recent challenges reconsidered. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

Berwick, R., & Chomsky, N. (under review). Poverty of the stimulus revisited: Recent challenges reconsidered.

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

Borensztajn, G., Zuidema, W., & Bod, R. (2008). Children's grammars grow more abstract with age - evidence from an automatic procedure for identifying the productive units of language. *Proceedings of the 30th Annual Conference of the Cognitive Science Society.*

Borovsky, A., & Elman, J. (2006). Language input and semantic categories: A relation between cognition and early word learning. *Journal of Child Language, 33*, 759-790

Briscoe, E. (2006). Language learning, power laws, and sexual selection. *6th International Conference on the Evolution of Language.*

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., & Vitànyi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*, 19-22

Chater, N., & Vitànyi, 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

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences, 14*: 597-650.

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

Feldman, J., Gips, J., Horning, J., & Reder, S. (1969) *Grammatical complexity and inference* (Tech. Rep. CS-TR-69-125). Stanford Univ.

Foraker, S., Regier, T., Khetarpal, N., Perfors, A., & Tenenbaum, J. (2009). Indirect evidence and the poverty of the stimulus: The case of anaphoric one. *Cognitive Science, 33,* 287-300.

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. (2006). 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., & Tenenbuam, 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.

Ishwaran, H., & James, L. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica, 13*: 1211-1235.

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

Johnson, M. (2006). *Inside-outside algorithm.* Brown University.

Johnson, M. (2007). Adaptor Grammars: a Framework for Specifying Compositional Nonparametric Bayesian Models. *Neural Information Processing Systems*, *19*

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.

Jurafsky, D. (2002). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. *Probabilistic linguistics.* Bod, R., Hay, J., & Jannedy, S. (eds). MIT Press

Kam, X., Stoyneshka, I., Tornyova, L., Fodor, J., Sakas, W. (2008). Bigrams and the richness of the stimulus. *Cognitive Science 32*, 771-787.

Kearns, M., & Valiant, L. (1989). Cryptographic limitations on learning (Boolean) formulae and finite automata. *Proceedings of the 21st Annual ACM Symposium on Theory of Computing.* 433-444

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2004). Learning domain structures. *Proceedings of the 26th annual conference of the Cognitive Science Society*.

Kemp, C. and Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687-10692.

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

Liang, P., Petrov, S., Jordan, M., & Klein, D. (2007). The infinite PCFG using hierarchical Dirichlet processes. *Proceedings of EMNLP*.

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* (3^{rd} 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. (2010). How recursive is language? A Bayesian exploration. *Linguistic Review.*

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

Pinker, S. & Jackendoff, R. (2005). The faculty of language: what's special about it? *Cognition*, *95*: 201-236.

Pitman, J., & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, *25*: 855-900.

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

Rose Finkel, J., Grenager, T., & Manning, C. (2007). The infinite tree. *Proceedings of ACL.*

Rumelhart, D., & McClelland, J. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge, MA: MIT Press.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2007). High-accuracy annotation and parsing of CHILDES transcripts. *Proceedings of ACL-2007 workshop on cognitive aspects of computational language acquisition.*

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

Generating the grammars

Generating grammars for each specific grammar type is accomplished in three ways: first, by designing the best grammars possible by hand; second, by using those grammars as the starting point of a local search through the space of all possible grammars; and, finally, by generating regular grammars in a completely automated fashion. Here we describe each of these processes in more detail.

Hand-designed grammars

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 [$V_{inf} \rightarrow V_{inf}$ PP] and [$V_{inf} \rightarrow$ vi 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.

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 → a] or [A → 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 → 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 → NP PP] expands to [NP → pro PP], [NP → 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 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. 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.[16]

The second search problem focuses on 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.

[16] 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 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.

Appendix B

The probabilistic model

Prior probability

Prior probability is a measure of the simplicity of a grammar, which can be captured by evaluating the number of choices required to generate it using a meta-grammar or "grammar grammar" (c.f., Feldman, Gips, Horning & Reder, 1969). If one were generating a grammar from scratch, one would have to make the series of choices depicted in Figure B1, 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-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:

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

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).[17] 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 will move to the "off" state and stop generating non-terminals. This process creates a distribution over non-terminals described by Equation 3.

[17] 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.

$$p(1-p)^{n-1} \quad (3)$$

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.

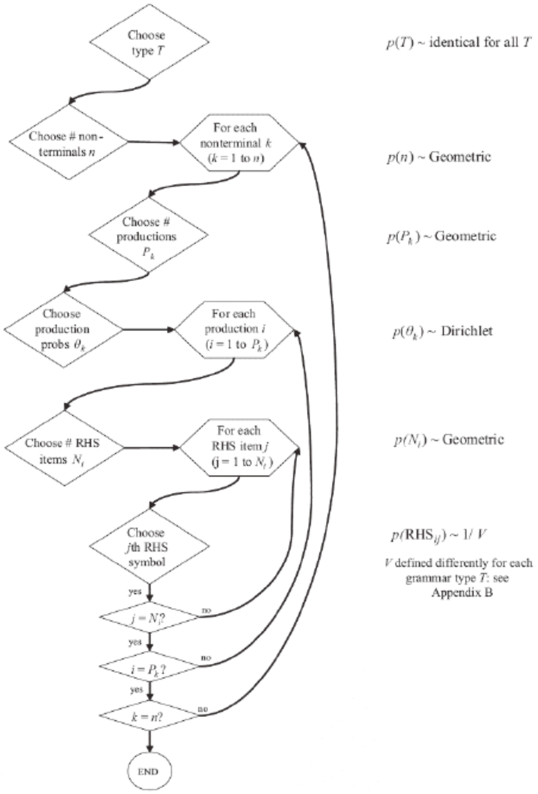
Choose type T
$p(T)$ ~ identical for all T
Choose # non-terminals n
For each nonterminal k (k = 1 to n)
$p(n)$ ~ Geometric
Choose # productions P_k
$p(P_k)$ ~ Geometric
Choose production probs θ_k
For each production i (i = 1 to P_k)
$p(\theta_k)$ ~ Dirichlet
Choose # RHS items N_i
For each RHS item j (j = 1 to N_i)
$p(N_i)$ ~ Geometric
Choose jth RHS symbol
$p(\text{RHS}_{ij})$ ~ 1/ V
V defined differently for each grammar type T: see Appendix B
yes
$j = N_i$?
no
yes
$i = P_k$?
no
yes
$k = n$?
no
END

Figure B1. Flowchart depicting the series of choices required to generate a grammar. More subtle differences between grammar types are discussed in the text.

Production-probability parameters.

Because each θ_k corresponds to the production-probability parameters for non-terminal k, the individual parameters $\theta_1,\ldots,\theta_m$ 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 $\theta_1,\ldots,\theta_m$, given that each production has been used α-1 times. We set α =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 α =1is 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.

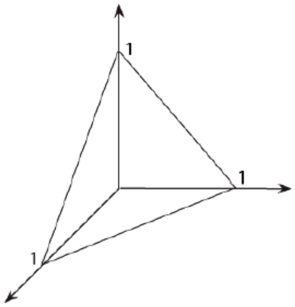

Figure B2. The unit simplex for m_k =3 (a triangle), corresponding to the Dirichlet distribution with α =1 on a θ_k vector of production-probability parameters with three productions.

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 $\theta_1,\ldots,\theta_m$, even though the Dirichlet distribution is continuous.

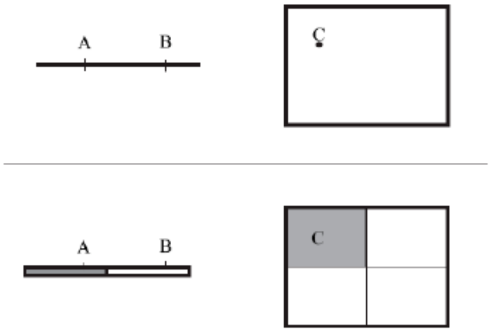

Figure B3. Top: one cannot compare the probability of continuous points A and C with different dimensionality. Bottom: when A and C correspond to discrete points, comparison is possible.

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 B3, 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 $\theta_1,\ldots,\theta_m$ 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 α =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 context-free 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 (right-)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.

Likelihood

The likelihood, or goodness of fit, of a grammar is calculated by comparing the effective set of sentences that the grammar *can* produce with the actual sentences in the corpus. The effective 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 set of distinct sentences 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:

$$p(D|G) = \prod_{l=1}^{M} p(S_l|G). \tag{4}$$

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 sum 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 4 in the same way, using the maximum-likelihood parameter values but integrating over all possible parses of the corpus.[18] Sentences with longer derivations will tend to be less probable, because each production used contributes a factor less than one to the product in Equation 4. 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.

Two-component adaptor grammar analysis

This analysis instantiates the "restaurant" metaphor described in the main paper, in which input corresponds to tables in a restaurant, and customers at each table are individual sentence tokens. Interpolating between types and tokens corresponds to assigning the *n* exemplars of a given sentence token to more than 1 but fewer than *n* tables. The adaptor component calculates the log probability of any given "seating assignment" of sentences to tokens, given by Equation 5 below (corresponding to Equation 4 in Goldwater et al., 2006). The probability of the entire two-component adaptor model is calculated by adding this to the

[18] 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). Doing so results in lower likelihoods but qualitatively identical outcomes in all cases.

log posterior probability of the grammars under one-component model (i.e., the results reported at the beginning of the results section).

$$P(\mathbf{w} \mid \theta) = \sum_{\mathbf{z},\ell} \left(\prod_{k=1}^{K(\mathbf{z})} \theta_{\ell_k} \right) \cdot \frac{\Gamma(K(\mathbf{z}))}{\Gamma(N)} \cdot a^{K(\mathbf{z})-1} \cdot \left(\prod_{k=1}^{K(\mathbf{z})} \frac{\Gamma(n_k^{(\mathbf{z})} - a)}{\Gamma(1-a)} \right) \tag{5}$$

Here, **w** corresponds to the data, θ captures the parameters of the multinomial distribution used as generator (i.e., the base measure in a Dirichlet Process mixture model), *N* corresponds to the number of tokens in the dataset, and **z** is the "seating assignment" in the Pitman-Yor process assigning tokens to tables (Pitman & Yor, 1997; Ishwaran & James, 2003). When α→1, *K*(**z**)=*N*, meaning that each table corresponds to one sentence token: this corresponds to a prior favoring a completely token-based analysis. As α→0, the sum over **z** is dominated by the arrangement that minimizes the total number of tables, corresponding to a prior favoring a completely type-based analysis. Intermediate values of α correspond to priors favoring an intermediate analysis.

The full two-component adaptor model requires summing over all possible assignments of sentences to tables for a reasonable intermediate[19] value of α and calculating the probability of each grammar and corpus for that table assignment. As explained in the main text, this sum is computationally intractable; we therefore search for an approximate best assignment given a corpus and grammar, with two questions in mind. First, is the single best assignment of sentences to tables closer to a type-based or a token-based analysis? Second, does the search, as it tends to move toward assignments with greater overall probability, always tend toward more type-based analyses, regardless of its starting point?

We performed such a search of table assignments using a simple Metropolis-Hastings algorithm with two types of proposal steps, one of which (the *condense* step) would move the search algorithm toward a more type-based analysis, and one of which (the *separate* step) would move it toward a more token-based analysis. At each step in the search, we calculated the probability of each grammar on the corpus given the current table assignment, as the sum of the log prior (Equation 2), the log likelihood (Equation 4), and the log probability of the adaptor component (Equation 5). We performed eight searches of 1000 steps each, and each

[19] We set this to an intermediate value, 0.5. An exhaustive search over all possible values of α would add even more to the intractability of the search, and in any case is not necessary; our goal is simply to evaluate what is learnable given reasonable assumptions, of which this is one.

with a different starting point (one starting from the fully type-based corpus, with 2336 sentences; one starting from the fully token-based corpus, with 21671; six starting at random intermediate corpora).

Results of this analysis indicate that the best-performing grammar out of all of these searches is hierarchical (*Level 6* posterior: CFG-L: -133659; REG-B: -133958; REG-M: -134242; CFG-S: -134566; other grammars were not analyzed because of the time required for the analysis, and because they did so much more poorly previously, making them unlikely to be serious contenders). Moreover, the search algorithm – which tends to accept proposal steps if they improve the overall probability, and not if they don't – accepted far more *condense* proposals (34.9%) than *separate* proposals (8.9%). In other words, proposals that made the corpus more type-based were more likely to improve the overall probability than proposals that made it more token-based. As another indication of the general preference for a more type-based analysis, every one of our searches except for the one that started on the fully type-based corpus[20] ended on a corpus that was smaller (i.e., more type-based) than where it started; that is, the gradient of search was always in the type-based direction. Although this search is by no means fully comprehensive – the computational problem is, as yet, too difficult for that – these results are coherent and suggestive. The results remain consistent with the one-component analysis reported previously: the best-performing grammars are still hierarchical, and there is good reason to believe that evaluating grammars on the basis of types rather than tokens is more appropriate.

[20] Obviously, this did not because it is impossible for a fully type-based corpus to become more fully-type based than it is. However, the search on this corpus did not move very far; after 1000 samples, the best-performing analyses used between 2340 and 2345 sentences for all of the grammars; this is not far from the smallest possible (fully type-based corpus) of 2336.

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 phrase-structure grammar CFG-L is* e^{101} *(~* 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.

	REG-AUTO			Other best grammars (posterior)						
Corpus	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 phrase-structure 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	Prior	-3968	-1915	-1349	-1166	-244	-698	-864
(2;3)	Likelihood	-881	-1265	-1321	-1322	-2199	-1489	-1448
	Posterior	-4849	-3180	-2670	-2488	-2433	**-2187**	-2312
Epoch 1	Prior	-22832	-3791	-1974	-1728	-257	-838	-1055
(2;3-2;8)	Likelihood	-5945	-7811	-8223	-8164	-13123	-8834	-8467
	Posterior	-28777	-11602	-10197	-9892	-13380	-9672	**-9522**
Epoch 2	Prior	-34908	-4193	-2162	-1836	-257	-865	-1096
(2;3-3;1)	Likelihood	-9250	-12164	-12815	-12724	-20334	-13675	-13099
	Posterior	-44158	-16357	-14977	-14560	-20591	-14540	**-14195**
Epoch 3	Prior	-48459	-4621	-2202	=1862	-257	-876	-1111
(2;3-3;5)	Likelihood	-12909	-17153	-17975	-17918	-28487	-19232	-18417
	Posterior	-61368	-21774	-20177	-19780	-28744	-20108	**-19528**
Epoch 4	Prior	-59625	-4881	-2242	-1903	-257	-876	-1111
(2;3-4;2)	Likelihood	-15945	-21317	-22273	-22293	-35284	-23830	-22793
	Posterior	-75570	-26198	-24515	-24196	-35541	-24706	**-23904**
Epoch 5	Prior	-67612	-5169	-2283	-1943	-257	-876	-1111
(2;3-5;2)	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.

			Can parse?						
Type	In input?	Example	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 8

Accuracy of each grammar (in terms of precision, recall, and F-score) on the final file from the Eve corpus. The table on the left compares the grammars to the automatic parses yielded by Sagae et al (2004) and Borensztajn et al. (2008); the table on the right compares them to hand-annotated parses. All grammars are also compared to right-branching (RB) and left-branching (LB) baselines. In all cases, the grammars favored by our model also have the highest accuracy (F-score).

Automatically parsed					Hand-parsed			
Grammar	Precision	Recall	F-score		Grammar	Precision	Recall	F-score
CFG-L	51.8	94.1	66.8		CFG-L	89.6	90.4	90.0
CFG-S	50.8	92.3	65.6		CFG-S	88.3	89.1	88.7
REG-B	50.1	90.8	64.6		REG-B	85.1	85.6	85.3
REG-M	50.1	90.8	64.6		REG-M	85.1	85.6	85.3
REG-N	49.4	89.6	63.7		REG-N	84.1	84.5	84.3
RB	50.7	92.4	65.5		RB	86.9	87.3	87.1
LB	32.0	62.9	42.4		LB	33.1	33.6	33.4

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 → B C	A → B F	A → b C	A → b F
A → B D	F → C	A → b D	F → d
A → B E	F → D	A → b E	F → g E
	F → E	C → g E	F → e D
		D → d	
		E → e D	