CSLI Lecture Notes Number 201

LANGUAGE FROM A COGNITIVE PERSPECTIVE

Grammar, Usage, and Processing

Studies in Honor of Thomas Wasow

edited by

Emily M. Bender & Jennifer E. Arnold

CSLI Publications
Stanford

Copyright © 2011
CSLI Publications
Center for the Study of Language and Information
Leland Stanford Junior University
Printed in the United States
15 14 13 12 11 1 2 3 4 5

Library of Congress Cataloging-in-Publication Data

Language from a cognitive perspective : grammar, usage, and processing / edited by Emily M. Bender and Jennifer E. Arnold.
p. cm. – (CSLI lecture notes number 201)
"This book is a collection of papers on language processing, usage, and grammar, written in honor of Thomas Wasow to commemorate his career on the occasion of his 65th birthday."
Includes bibliographical references.
ISBN 978-1-57586-611-6 (alk. paper) –
ISBN 978-1-57586-610-9 (pbk. : alk. paper)
1. Cognitive grammar. 2. Grammar, Comparative and general–Syntax. I. Bender, Emily M., 1973- II. Arnold, Jennifer E. III. Wasow, Thomas.

P165.L38 2011
415—dc22

2011002689

CIP

∞ The acid-free paper used in this book meets the minimum requirements of the American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

CSLI was founded in 1983 by researchers from Stanford University, SRI International, and Xerox PARC to further the research and development of integrated theories of language, information, and computation. CSLI headquarters and CSLI Publications are located on the campus of Stanford University.

CSLI Publications reports new developments in the study of language, information, and computation. Please visit our web site at
http://cslipublications.stanford.edu/
for comments on this and other titles, as well as for changes and corrections by the author and publisher.

Contents

6

Simplicity and Fit in Grammatical Theory

Amy Perfors

Copyright 2011 CSLI Publications

Preface

This paper reflects two, somewhat independent, ways that Tom Wasow has influenced my work. The first concerns the motivating question: how should we evaluate different grammatical theories or formalisms? Which best account for human language, and how do we decide? This issue was emphasized beginning with the first syntax course I ever took (taught, of course, by Tom), and was an undercurrent in many of the discussions I had with him henceforth. The second way that Tom influenced my thinking was in his interest in using computational techniques to study aspects of language. In work with him, David Beaver, and several others, we used a genetic programming framework to explore a question in language evolution. Although the approach and question were different than the one discussed here, the general insight that computational methods can be a useful tool for investigating aspects of language dates back to my time spent working with Tom.

1 Overview

Much current research in linguistics and cognitive science is focused on the question of innateness: whether the cognitive capacities that enable humans to learn language are language-specific, or whether our linguistic skill is the result of more domain-general abilities and biases. Interest in this topic grew especially strong after Chomsky's (1965) claim that language learning is only explicable on the basis of an innate

Language from a Cognitive Perspective.
Emily M. Bender and Jennifer Arnold, Editors
Copyright © 2011, CSLI Publications.

language faculty, or Universal Grammar. Since then, much inquiry has focused on the dual questions of (a) to what extent this claim is true; and (b) to the extent that it is, what is the nature of this Universal Grammar, or UG?

Addressing these questions requires both the ability to accurately describe the data that need to be explained, as well as the capacity to evaluate the different theories that aim to explain that data. In this paper I describe a paradigm that meets both of these requirements. The paradigm is based on Bayesian computational modeling of grammars and complements existing linguistic methodologies.

The structure of the paper is as follows: In the first section, the utility of such an approach is motivated. Subsequent sections describe its basic underlying principles, and briefly consider two case studies that illustrate how it might provide additional insight. The paper concludes with a discussion of some of the issues and limitations associated with the paradigm.

2 A Motivation of the Bayesian Approach

Advances in understanding the bases of language learning in the brain require, among other things, an accurate characterization of the data (i.e., the language) being learned, as well as a means to evaluate theories explaining that data. In this section I discuss some of the issues that surround these two requirements. This motivates the ways in which the Bayesian paradigm fills some existing gaps.

2.1 Characterizing the Data as a Whole

As a field, linguistics relies on several different kinds of methodologies in order to properly characterize the data that need to be explained. One common approach, at least in the subfield of generative syntax (which we will be focusing on in this article), is a reliance on grammatical intuitions—introspective judgments as to an expressions' grammaticality or well-formedness. Though these intuitions can be a useful tool in guiding the formation of theories, Wasow and Arnold (2005) argue that using them as the primary or only source of empirical support for a theory can sometimes be problematic, since individual speakers may often disagree, and intuitions may be rather marginal even for a single speaker. Other sources of empirical evidence, emerging from subfields such as psycholinguistics, experimental psychology, and cognitive science, include reaction-time experiments (e.g., Spivey & Tanenhaus, 1998), eye-tracking paradigms (e.g., Just & Carpenter, 1980; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Altmann & Kamide, 1999), corpus analyses (e.g., Nunberg, Sag, & Wasow,

1994; Lohse, Hawkins, & Wasow, 2004; Levy, 2008), and survey data (e.g., Langendoen, Kalish-Landon, & Dore, 1973; Wasow & Arnold, 2005), all of which result in a statistically valid and nuanced picture of grammatical acceptability.

However, all of these methods often yield data only regarding the particular constructions or phenomena in question. Though this may be interesting in its own right, because syntacticians are often focused on the question of which grammatical formalism or theory best describes an entire language, it is, of necessity, limited in scope: every theory includes some phenomena that it can explain easily and some that are only accounted for via more *ad hoc* measures. What may often be desirable is some method or mechanism that can objectively decide between entire theories on the basis of how well they account for observed natural language usage. But how is this achievable even in theory, much less in practice?

2.2 Evaluating Theories in a Principled Way

Intuitively, there are two main criteria that are essential when deciding between different linguistic theories. One is the issue of goodness-of-fit: what is the coverage of the theory to the observed linguistic data? Does it account for the important phenomena in a wide variety of languages, while not predicting phenomena that are unobserved or unattested? It is this criterion that empirical data are relevant to, and the evaluation of theories with respect to this criterion underlies the importance of acquiring accurate data. Another, equally important, criterion is that of *simplicity*: any theory can achieve complete coverage simply by exhaustively listing all of the phenomena in question, but we would quite clearly like to rule out those theories. Philosophers of science have long emphasized the importance of simplicity (e.g., Wrinch & Jeffreys, 1921; Good, 1968; Jaynes, 2003). Within linguistics, many people (including Chomsky) have called attention to the critical role that simplicity plays in the evaluation of a linguistic theory (e.g., Chomsky, 1956, 1957, 1965; Wolff, 1982; Chater & Vitányi, 2007, among others).

Despite the fact that almost everyone accepts the importance of simplicity in evaluating theories, there are few widely agreed-upon criteria for measuring it. One might argue that simpler formalisms or theories are those that are less expressive—those that license the fewest phenomena.[1] An important branch of research has focused on the expressive-

[1]This is a view of simplicity that overlaps somewhat with the traditional view of coverage or goodness-of-fit; one reason that more expressive theories are dispreferred is that they are so powerful they license grammars or phenomena that do not fall within the purview of natural language.

Copyrighted Material Publications

ness of different syntactic theories, including GB, HPSG, minimalism, and others: for instance, every recursively enumerable set of strings is a transformational language (Peters & Ritchie Jr., 1973), suggesting that transformational grammars are more powerful than necessary to account for natural language. (See also, e.g., Kornai & Pullum, 1990 and Rogers, 1998 for other formal analyses of the expressive power of different linguistic theories, and Immerman, 1999 for a mathematical overview).

However, equally important is the question of simplicity of explanation *within* a theory. Are the central linguistic phenomena accounted for easily by the theory, or must it contain a multitude of *ad hoc* exceptions, or increasingly complicated rules, in order to account for it? This sort of simplicity argument is the basis for Chomsky's famous conclusion that regular grammars (i.e., Markov models, also known as word-chain grammars) are inadequate to capture natural language (1956): because any actual corpus or set of data is finite, regular grammars are in principle capable of capturing them completely, but the presence of long-distance dependencies means that these grammars will be, in his words, "so complex as to be of little use or interest." (p. 115) Under this notion, then, the expressiveness of a linguistic theory or grammar is a good thing, because more expressive grammars will (as a general rule) find it easier to capture any given linguistic phenomenon in a parsimonious way.

Another measure of the simplicity of a formalism or theory focuses on the number of primitives or basic operations defined within that theory. By this measure, minimalism is very simple, since it focuses on the importance of economy of derivation and economy of representation when defining linguistic theories, and explains phrase structure in terms of only two operations, Merge and Move (Chomsky, 1995). In fact, minimalism is often justified on the basis of simplicity, or based on the related notion of "perfection" in a theory. However, it has also been criticized on the grounds that these notions of simplicity and perfection are too vague to be useful (e.g., Lappin, Levine, & Johnson, 2000).

It is possible to view much of the debate in formal syntax about different theories as actually a debate between different views of simplicity, and how it should be balanced with goodness-of-fit—both areas in which there is no substantial agreement. This is partly because, as already discussed, there is not agreement about what aspects of simplicity matter or how it should be measured; but it is also because it is unclear how to properly implement the tradeoff between the two. Is a theory with few primitives and operations, like minimalism, "better" than a theory that builds more in but produces more precise explana-

Copyright Material - Cambridge Publications

tions? And how should one evaluate theories that explicitly shovel part of the explanation into another component of the language faculty—a component that is not fully fleshed out within the theory itself? For instance, it may be accurate to assume that lexical semantics and syntax are so intertwined that any good syntactic theory should displace much of the explanation onto word-specific knowledge, as does HPSG—but that still leaves us with the problem of how to balance HPSG's simplicity with its explanatory coverage. On some measures it appears quite parsimonious, but how much of that is because it has succeeded in offloading most of the explanation onto the lexical semantics of each word?

The tradeoff between simplicity and goodness-of-fit is a perennial issue in all of the sciences, and a central topic in philosophy of science as well. A common heuristic is that of Occam's Razor—*entia non sunt multiplicanda praeter necessitatem*—that an explanatory hypothesis or theory should not make assumptions (or "postulate entities") unless absolutely necessary. Although this heuristic is generally regarded as little more than a rule of thumb, it has deep connections with Bayesian probability theory and information theory (e.g., Jeffreys, 1931, 1939; de Finetti, 1974; Vitànyi & Li, 2000; Jaynes, 2003; MacKay, 2003). In the next sections, I will illuminate those connections, and demonstrate how the Bayesian framework may be used to compare and evaluate grammatical rules and grammatical theories in linguistics. In so doing, it can provide a means to address some of the difficulties that current linguistic approaches wrestle with.

3 Principles behind Bayesian Grammar Induction

In Bayesian probability, one's degree of belief in some hypothesis or theory is represented by a real number between 0 and 1. The mathematics of probability theory provides rules for "proper reasoning"—for how to validly combine different premises and beliefs in such a way as to be sure that you have arrived at the correct conclusion (e.g., Jaynes, 2003). In essence, it is an extension of deductive logic to the case where propositions, or hypotheses, have degrees of truth or falsity (and is identical to deductive logic if we know all of the hypotheses with 100% certainty). Thus, just as formal logic describes a deductively correct way of thinking, Bayesian probability theory describes an inductively correct way of thinking. As Laplace said, "probability theory is nothing but common sense reduced to calculation."

What does this mean? If we were to try to come up with a set of desiderata that a system of proper reasoning should meet, they

Copyright © 2007 CSLI Publications

might include things like consistency and qualitative correspondence with common sense—if you see some data supporting a new theory A, you should conclude that A is more plausible than it was, rather than less; the more you think A is true, the less you should think it false; if a conclusion can be reasoned in multiple ways, its probability should be the same regardless of how you got there; etc. The basic axioms and theorems of probability theory, including Bayes' Rule, emerge when these desiderata are formalized mathematically (Cox, 1946, 1961), and correspond to common-sense reasoning and the scientific method (Jeffreys, 1931, 1939; de Finetti, 1974; Jaynes, 2003).

This means that optimal inductive inference—of the sort we hope to achieve in scientific reasoning—should follow the Bayes' Rule, in which the probability of a hypothesis given some data $p(H|D)$ is proportional to the probability of the data given that hypothesis $p(D|H)$, or likelihood, times the prior probability of that hypothesis $p(H)$:

$$p(H|D) \propto p(D|H)p(H) \tag{6.1}$$

Hypotheses (and data) are defined within the Bayesian framework as the outgrowth of a generative process: for instance, data (such as spoken sentences) may be generated from some sort of underlying grammar, and grammars themselves are generated from a hypothesis space of candidate grammars. The job of the learner is to choose among different hypotheses—grammars—on the basis of which ones best account for the observed data. This choosing is done according to the laws of Bayesian probability theory, including Bayes' Rule.

Simplicity is naturally accounted for via the prior probability $p(H)$. The definition of simplicity and the corresponding calculation of $p(H)$ are not generally the result of some externally-imposed *ad hoc* mechanism; rather, they emerge naturally from the assumption that hypotheses (grammars) themselves are generated from a space of candidate hypotheses. For instance, the hypotheses in Figure 1 correspond to different sets of rectangular regions within a two-dimensional space. Simpler hypotheses require fewer "choice points" during the generation process: Hypothesis A can be fully captured by making only four choices, two for the coordinates of the lower-left-hand corner of the rectangle (x and y), one for its length (l), and one for its width (w). By contrast, hypothesis C contains thirty distinct rectangular regions, and therefore requires 120 separate choices to specify, four for each region. This notion of calculating complexity as a function of the number of choice points is a reflection of the idea that the more complicated

Copyright © 2011 John Benjamins

something is, the more likely it becomes that it will be messed up at some point in the generation process. The more choices a hypothesis resulted from, the more likely it is that those choices could have been made in a different way, resulting in a different hypothesis.

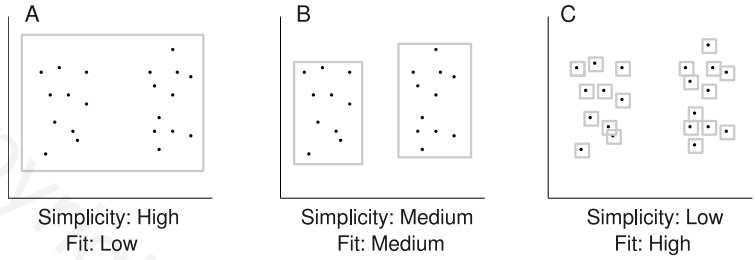

FIGURE 1 Hypothesis A is too simple, C is too complex, and B is "just right." Hypothesis A is quite simple, but fits the observed data poorly: C fits closely but is highly complicated. The best description of the data should optimize a tradeoff between complexity and fit, as in B.

The precise prior probability of a hypothesis is therefore not arbitrarily assigned, but rather falls out in a principled way from how the hypotheses are generated. The generative model for the hypotheses in Figure 1 is one that can result in any possible combination of rectangular regions within the space. A different generative model would result in a different—but no less principled—assignment of prior probabilities. For instance, if we assumed that the regions could be circles rather than rectangles, then each region would require three choice points rather than four (the x and y coordinates of the center of the circle, plus its radius). The logic favoring simple hypotheses would be the same: multiple regions will still be *a priori* less likely than a few. The precise generative model therefore matters for determining exactly what the relative probability of a hypothesis would be, but most reasonable models would give qualitatively similar relative probabilities to qualitatively similar hypotheses.

The Bayesian framework, then, offers a natural way to both calculate the simplicity of different hypotheses or theories, and also evaluate those theories on the basis of how well they account for the observed data.[2] Bayes' Rule offers a principled way to evaluate the tradeoff between simplicity (prior probability) and goodness-of-fit (likelihood).

[2] A quite similar insight is offered by a minimum description length (MDL) approach, which suggests that coding length (a measure of simplicity) can be an important tool for choosing between different linguistic analyses (e.g., Chater & Vitányi, 2007; Goldsmith, 2007). Indeed, such an approach has been used effec-

Copyright © 2011 Linguistic Publications

Thus, as in Figure 1, it will naturally tend to prefer hypotheses (like Hypothesis B) that—like Goldilocks in the famous story—are neither too weak nor too strong, but are "just right." Hypothesis C, for instance, clearly has a high degree of goodness-of-fit (likelihood): if the hypothesis is true—that is, if the data is truly generated by thirty distinct underlying processes corresponding to the thirty rectangles of C—the datapoints could hardly be anywhere else. In other words, it fits (or predicts) the data well. By contrast, hypothesis A has relatively low likelihood: it does not explain why the datapoints are where they are, rather than elsewhere within the rectangle. However, hypothesis A is simple, while C is quite complex. The best description of the data would be a hypothesis that optimizes the tradeoff between complexity and fit, as in hypothesis B.

This framework is applicable to hypotheses of much greater complexity than rectangles in a two-dimensional space. It is possible to define generative models for grammars in which specific grammars, G, are generated from a larger class of grammar types T (see Horning, 1969; and Feldman, Gips, Horning, & Reder, 1969 for other examples of this idea). Consider, for instance, context-free grammars, which naturally capture hierarchical phrase structure by being able to generate sentences in which clauses can be located inside other clauses. Context-free grammars, or CFGs, have productions of the form $X \rightarrow y$; X is a single non-terminal production (meaning that it can appear on the left-hand side of a production) and y is a string of non-terminals or terminals (terminals are constrained to appear on the right-hand-side, and consist of the output symbols of the grammar). Within the class of context-free grammars, one could generate a specific grammar by going through the following steps: (a) to choose the number of non-terminals n; (b) for each non-terminal k to generate P_k productions; (c) for each ${P_k}^{th}$ production i, to generate N_i right-hand-side items (either one or two, if we constrain the grammars to be in Chomsky normal form), each of which are drawn from the grammars vocabulary V (the set of all non-terminals and terminals). If one wanted the grammars to be probabilistic, one would also have to assign a vector of production-probability parameters θ_k for each non-terminal k. This process imposes

tively for the acquisition of morphology (Goldsmith, 2001), word segmentation (de Marcken, 1995), and other aspects of grammar (e.g., Dowman, 1998; Grünwald, 1996). One of the main differences between the MDL and the Bayesian frameworks lies in how simplicity is measured: in the MDL approach, simplicity is captured by short encoding lengths, while in the Bayesian approach, it is captured by higher prior probability. However, there are deep similarities between the two approaches (see, e.g., Vitànyi & Li, 2000, for a discussion).

Copyright Material, Oxford University Press

a prior probability, as in Equation 6.2, in which simpler grammars—those with fewer non-terminals, productions, and items—have higher prior probability (see Perfors, Tenenbaum, and Regier (under review) for a more thorough explanation of this process). This is captured by the equation below:

$$p(G|T) = p(n) \prod_{k=1}^{n} p(P_k)p(\theta_k) \prod_{i=1}^{p_k} p(N_i) \prod_{j=1}^{N_i} \frac{1}{V} \tag{6.2}$$

Not only does this process naturally impose a prior probability metric in which shorter grammars with fewer non-terminals are simpler, the generative framework also naturally operates so that more expressive—i.e., more complex—grammar *types* will be effectively penalized. For instance, the generative model for regular grammars would be analogous to that of the process for context-free grammars, except that the form of the right-hand side of productions would be more constrained. Permissible productions for a (right-branching) regular grammar include only those of the form A $\rightarrow$ a B or A $\rightarrow$ a (where capital letters indicate non-terminals and lower-case letters indicate terminals), whereas context-free grammars may additionally include productions of the form A $\rightarrow$ B C, A $\rightarrow$ B, or A $\rightarrow$ B a. As a result of this, regular grammars are a subset of context-free grammars, and if a particular grammar could be generated as an example of more than one grammar type, it would receive higher prior probability when generated from the less expressive type. All other things being equal, one would have 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.

In essence, then, prior probability can be defined over grammars in such a way as to naturally capture our intuitive notion of simplicity, in such a way that simpler grammars *within* a theory will be favored, and simpler (less expressive) theories will also be favored, all else being equal. The Bayesian framework also provides a way to compare different grammars in terms of how well they fit the observed linguistic data in the world. Consider, for instance, data consisting of a corpus of sentences spoken by native English speakers. A grammar's degree of fit to that data—its likelihood—reflects the probability that the data would be generated by that grammar. Assuming that each sentence is generated independently from the grammar, this would be given by the

Copyright © 2011 Taylor & Francis

product of the likelihoods of each sentences S_i in the corpus; with M unique sentences in the corpus, this would be:

$$p(D|G) = \prod_{l=1}^{M} p(S_l|G) \tag{6.3}$$

Likelihood reflects the goodness-of-fit of a corpus of data to an underlying grammar in the same way that it reflects the goodness-of-fit of the dataset of dot points to an underlying rectangular "theory" in Figure 1. In that example, it seems intuitively that hypothesis B fits the data more closely than hypothesis A, but why? If A were the correct model, it would be quite a coincidence that all of the datapoints fall only in the regions covered by B. Similarly, if we were comparing two grammars X and Y, and X could generate all and only the sentences observed in the corpus but Y generated many others that were never observed, then X has better fit: if Y were the correct grammar, it would be an amazing coincidence that all of the sentences just happened to be the ones that X could generate. Likelihood is thus dependent on the quantity of data observed: while it would not be much of a surprise to see just one or a few sentences consistent with X if Y were in fact the correct grammar, seeing 1000 sentences—and none that could *not* be generated by X—would be very surprising indeed, if Y were correct.

The effective set of sentences that a probabilistic grammar can produce depends on several factors. All other things being equal, a grammar with more productions will produce more distinct sentence types. But the number of distinct sentences generated also depends on how those productions relate to each other: how many of the same left-hand side terms there are (and thus how much flexibility there is in expanding any one non-terminal), whether the productions can be combined recursively, and other factors. A penalty for overly expressive or flexible grammars exists here, too, because likelihood is assigned by considering all possible ways of generating a sentence under a given grammar and assigning probabilities to each derivation. The total probability that a grammar assigns over all possible sentences must sum to one, and thus the more flexible or expressive the grammar, the lower probability it will tend to assign to any one sentence.

So far I have demonstrated how the Bayesian framework can be used in theory to compare entire grammars in terms of their simplicity and their goodness-of-fit to actual corpora of real, naturalistic data. This approach is consistent with Chomsky's formulation of the problem of language learning, which presumes both a hypothesis space of grammars and the existence of an evaluation metric based on simplicity

Copyright © 2010 John Benjamins

(Chomsky, 1965). Prior probability produces an objective measure of a grammar's simplicity, while likelihood captures the degree of fit of a grammar to the data, and penalizes grammars or grammar types that are too expressive—that overgeneralize too much beyond the data.

Bayes' Rule and the mathematics of probability theory provides a principled way to combine these two factors in such a way to guarantee optimal inductive reasoning ability. Indeed, it has been formally proven that an ideal learner incorporating a simplicity 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; Chater & Vitányi, 2007). It is therefore reasonable to think that the Bayesian approach may be well-suited to providing an objective way to compare different grammatical theories and formalisms within linguistics—and is thus another method for addressing many of the questions that have occupied linguists for years. In the next section, I will give some examples of how this method has been applied, and I will end with a discussion of the limitations and complexities inherent in applying it further.

4 Bayesian Grammar Comparison in Practice

4.1 Learning Abstract Syntactic Information

One issue that has been the focus of much work in linguistics for years is the question of abstract syntactic structure, and to what extent human learners are born with innate language-specific knowledge about that structure. It is widely accepted that natural language incorporates hierarchical phrase structure: that is, that the rules of syntax are defined over linguistic elements corresponding to phrases that can be represented hierarchically with respect to one another (Chomsky, 1965, 1980). 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.

Why do linguists believe that language has hierarchical phrase structure? We have already discussed one of the main arguments, originally proposed by Chomsky in (1956). His conclusion that regular languages are inadequate to capture natural language centered around their inability to capture hierarchical phrase structure (and therefore long-distance dependencies based on that structure). The reasoning is, at its essence, a simplicity-based argument: because regular languages have so much less expressivity than language classes that incorporate hierarchical phrase structure, a regular grammar sufficient to capture natural language would have to be unrealistically complex. This ar-

gument, though intuitively compelling and reasonable, is still based on intuition; would an objective learner capable of trading off the complexity of regular grammars and how well they explained natural language data arrive at the same conclusion as Chomsky? And what implication might that have for the question of whether children learning language might also be able to arrive at the same conclusion?

To explore these questions, Perfors et al. (under review) presented a Bayesian learner capable of representing both regular and context-free grammars with a corpus of naturalistic child-directed speech. The learner was unbiased with respect to grammar type, meaning that it initially favored neither regular nor context-free grammars as being *a priori* more or less likely. Its prior probability and likelihood were defined as in Equations (6.2) and (6.3), so that it favored grammars that balance simplicity (containing fewer productions and terminals) with fit (overgeneralizing less). The data consisted 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, individual words were replaced by their syntactic categories.[3] Furthermore, each grammar was evaluated based on the probabilities it assigned to the set of sentence types occurring in the corpus, independent of the frequencies with which those types occurred (i.e., the sentence token frequencies). This choice was based on the adaptor grammar framework of grammar induction introduced by Goldwater, Griffiths, and Johnson (2006). This parallels—and gives a principled justification for—the standard linguistic practice of assessing grammars based on the forms they produce rather than the precise frequencies of those forms.

A range of grammars was evaluated on this corpus, based on how well they optimized the tradeoff between simplicity and goodness-of-fit. Multiple context-free and regular grammars were identified and compared. Because the computational problem of searching the "grammar space" to identify the optimal one is intractable given current technology, we cannot be certain that the grammars considered represent the "best" of each type. However, every available method for searching the space as thoroughly as possible was implemented: some grammars were designed by hand; others were found via local search of the space using

[3] Although learning a grammar and learning a lexicon are probably tightly linked, this may be a fair assumption for several reasons: first, because grammars are defined over these categories, and second, because there is some evidence that aspects of syntactic-category knowledge may be in place even in young children (e.g. Louann, Rachel, & William, 2005).

Copyright © Oxford University Press. All Rights Reserved.

the hand-designed grammars as a starting point; and other grammars were generated via an automatic search of the space.

Results indicated that the Bayesian learner preferred grammars that incorporated hierarchical phrase structure over grammars that did not: the model attributed the highest overall (posterior) probability to context-free grammars, and less to regular grammars or a simple list of memorized sentences. Interestingly, this remained true even if the data consisted of a tiny subset of the corpus, equivalent to just over an hour's worth of conversation at age 2;3. The reason for this is that although the regular grammars for the most part achieved a closer "fit" to the corpus by overgeneralizing less (i.e., producing fewer unobserved sentences) they accomplished this by sacrificing simplicity: just as Chomsky hypothesized, these grammars were unwieldy and long, containing many extra productions and non-terminals relative to the simpler context-free grammars. Regular grammars constructed to be as simple as the context-free grammars, on the other hand, lacked the expressivity to closely capture the sentences of natural language, and had a lower likelihood than equivalently simple context-free grammars. In essence, the grammars without hierarchical phrase structure were like hypotheses *A* and *C* in Figure 1, whereas the grammars with it were more like hypothesis *B*.

In addition to evaluating entire grammars, this framework can also be used to explore particular phenomena in linguistics. We will see an example of this in the next subsection.

4.2 Exploring Recursion

One of the most notable features of human language is its capacity to generate a potentially infinite number of possible sentences. Because such a capacity must result from an underlying generative mechanism (a grammar) that is recursive in some way, many linguists have concluded that recursion must be a fundamental, possibly innate, part of the language faculty (Chomsky, 1957). Some have gone further and claimed that the core mechanism underlying recursion is the only part of language that is specific to humans (Hauser, Chomsky, & Fitch, 2002). While the latter, stronger claim is contested (Pinker & Jackendoff, 2005), the former has been largely accepted for decades. However, recent work on Pirahã, a language spoken in the Amazon basin, suggests that there may be a language that does not contain any recursion in its phrase structure whatsoever (Everett, 2005).

The empirical claim about Pirahã is the subject of much debate (Nevins, Pesetsky, & Rodrigues, 2007; Everett, 2007), and an essential key to resolving the debate is to be able to objectively determine

Computational Linguistics Publications

whether Pirahã is better described by a grammar with recursive elements or by one without. Lacking an objective and principled mechanism for comparing grammars with respect to linguistic corpora, it can be difficult to ascertain whether one grammar constitutes a "better description" than another. However, as before, the Bayesian framework provides a way of resolving this difficulty.

Here again we see the roles that simplicity and goodness-of-fit play. The standard reason for thinking that grammar is recursive is because of its property of discrete infinity: it is composed of discrete basic elements (words) which can be combined to produce apparently infinitely many sentences. An infinite set can be generated from a finite grammar only if the grammar contains some form of recursion; but is it true that natural language is infinite? After all, there are no infinitely long sentences, and only a finite number of sentences have been uttered. It is therefore possible to believe that the true grammar is one without any recursive rules. However, most linguists reject this possibility, on the grounds of simplicity: a non-recursive grammar capable of generating natural language would be very large, since it would require additional sets of rules for each additional depth of recursive expansion.

This simplicity-based argument is reasonable, but is not airtight, and is based on our intuitions about how much more complex a grammar with non-recursive instead of recursive rules would be. The complexity of a grammar would increase with each additional rule, and how many non-recursive rules would necessarily depend on the precise sentences being explained. Unfortunately, recursive productions hurt the fit of a grammar on any finite dataset, since they will always predict sentences that are not observed. The fewer sentences there are in the dataset that result from multiple expansions of recursive rules, the more a grammar with recursive rules is favored relative to one without.

Thus, recursion involves an inherent tradeoff between simplicity and goodness-of-fit, and we cannot conclude on *a priori* grounds that any grammar for natural language must contain recursion. At the very least, it may not be true in all cases, whether for a specific language (e.g., Pirahã), or for a specific rule or set of rules (e.g., center-embedded relative clauses in English). Perfors, Tenenbaum, Gibson, and Regier (2010) addressed this issue by comparing grammars (using the definitions of prior and likelihood given in Equations (6.2) and (6.3) based on how well they accounted for corpora of natural language data. Instead of comparing grammars of different types—those with hierarchical phrase structure and those without—they compared grammars with different levels of recursion.

Copyright © Oxford University Press

All grammars were context-free, since CFGs are often adopted as a first approximation to the structure of natural language (Chomsky, 1959) and are standard tools in computational linguistics (e.g., Jurafsky & Martin, 2000; Manning & Schütze, 1999). Three main grammars were evaluated that differed from one another only in whether certain rules were recursive or not. The fully recursive grammar contained fully recursive noun phrases (e.g., NP → NP CP); another grammar contained no recursive noun phrases at all, but rather multiple-embedded non-recursive productions involving an additional new non-terminal, N2, which permitted parses of up to a depth of two. There was also a "middle ground" grammar containing both recursive and non-recursive "shadow" rules (which decrease the weight assigned to the recursive rules by accounting for the many non-recursive instances of noun phrases).

The grammar with both recursive and non-recursive rules was favored by the Bayesian learner, largely because it achieved an expressiveness similar to the fully-recursive grammar, but without an equivalent loss in goodness-of-fit. This suggests that syntax, while fundamentally recursive, might usefully employ non-recursive rules to parse simpler sentences that recursive rules could parse in principle. This would not change the expressive capability of the grammar, but might dramatically decrease the cost of recursion. This may also suggest how a learner could infer that language is recursive, despite never having heard sentences that go beyond only a few levels of embedding in the input: as long as the recursive rules have low enough weight, the penalty for "overgeneralizing" beyond a few levels of embedding would be minimal.

5 Further Issues and Concerns

The analysis discussed in this paper is potentially relevant to both of the two issues raised in section 2. It demonstrates how the Bayesian framework can provide a useful means for evaluating which specific grammars and grammar types best capture or explain natural language, going beyond qualitative, intuitive arguments to provide an objective criterion for grammar comparison. The analysis does incorporate certain assumptions—e.g., that sentence types rather than sentence tokens are the relevant data, or that grammars with fewer productions and non-terminals were simpler in the relevant sense—but the Bayesian framework forces those assumptions to be made explicit and provides a means to evaluate the extent to which the conclusions depended on them. Furthermore, this work may have implications for questions of innateness: if it is possible for an unbiased Bayesian learner to realize

Copyright © Cognitive Science Publications

that language has hierarchical phrase structure on the basis of a limited amount of child-directed speech, what does this imply about whether such knowledge is (or need be) innate to children? By their nature, demonstrations of effective learning by Bayesian models cannot *necessarily* imply anything positive about the learning abilities of children, but they do serve as a proof-of-concept that something is learnable, given the assumptions built into the model. As such, they provide another path toward understanding the learning abilities children actually have.

Other examples of Bayesian methods—or, more generally, computational methods that combine structured representations with statistical inference machinery—abound in the computational linguistics and cognitive science literature. Some may have implications for human learning even if that was not the primary original purpose of the research. One example of work like this would include the adaptor grammar framework we briefly discussed earlier (Goldwater et al., 2006; Johnson, 2008). It was originally developed in order to create an adequate model for the unsupervised learning of morphology, but the general framework—including the notion of a two-part generative process for language, which separates the generation of allowable types from the process that explains their frequencies—may have much broader implications.

Bayesian methods are also increasingly common as a means to characterize the nature of the learning problem confronting the child, including a way to solve it. To list just a few examples, word segmentation may be accomplished by a learner sensitive to the transition probabilities between words, as well as further contextual dependencies among words (Goldwater, Griffiths, & Johnson, 2007); the problem of identifying the referents of nouns may be addressed by a learner who attends to the statistics of word use across multiple situations, and is attentive to social cues (Frank, Goodman, & Tenenbaum, 2009); and the acquisition of verb argument constructions, even without negative evidence, may be achieved by a learner sensitive to the statistics of what does not appear in the input (Alishahi & Stevenson, 2008; Perfors, Tenenbaum, & Wonnacott, 2010). This sort of research differs from the examples considered in this paper in that it does not involve the explicit comparison of specific grammars or grammar rules with the goal of identifying which theories best describe natural language, but the questions are similar.

One of my goals with this paper was to convince readers of the utility of applying the Bayesian framework for grammar comparison as a means of addressing two of the most important issues to linguists

Copyright © Cambridge University Press

today: addressing the question of innateness, and deciding which formalisms or theories best capture natural language. The examples given here illustrate both how this framework can be useful for addressing these issues, but they also illustrate some of the potential limitations. Two of those are especially salient for our purposes here.

First, the utility of this method is limited by the extent to which it is possible to define and generate all of the grammars or grammatical rules in question, as a part of a coherent framework. Both of the examples given analyzed context-free and regular grammars, but none with greater sophistication: e.g., dependency grammars, grammars with explicit transformational rules, or minimalist grammars. In part this was because the simpler grammars were all that was necessary to address the questions under consideration—but in part it was because the simpler grammars were significantly easier to define within a generative framework, and to successfully calculate the likelihood for. Likelihood calculations require accurate and quick parsing of all sentences in the corpus, and the ability to assign a probability to each of those sentences. The technology for accomplishing this for context-free and regular grammars exists (Jurafsky & Martin, 2000; Manning & Schütze, 1999) but is less well-established for other types of grammars. This does not mean that the Bayesian framework for grammar comparison is not *in principle* extendible to higher-complexity grammars—but it does mean that such an implementation would need to co-occur, or build on, technical advances in these areas.

Second, the extent that one can draw strong conclusions from the performance of a Bayesian learner on a corpus of natural language data to the abilities of actual human learners may be somewhat limited. Something similar could probably be said for any single method, of course, but this is nevertheless good to keep in mind. Exploring what sort of grammars or theories a Bayesian learner favors, given some input, can shed light on (a) abstract learnability issues of what it may be *possible* to acquire, given certain assumptions about the learner and the data; (b) what different assumptions about the learner, the data, or the representation buys you in terms of how it changes the abstract learnability; and (c) in some objective sense, which theories better describe the observed data. But in order to draw stronger conclusions about the *actual* nature of human learners, the predictions of theoretical models (including Bayesian ones) crucially need to be compared against empirical evidence about language learning and language use.

In sum, then, this paper has described a computational framework for comparing grammars and grammatical rules, and given several examples intended to illustrate how it can be of utility when addressing

Copyright © Cambridge University Press

two of the major questions of concern to linguists. The framework is not intended to supplant other methods in cognitive science or linguistics, but I suggest that it is a useful tool in the toolbox as we move toward constructing a full and accurate picture of the human mind.

Acknowledgements

The work discussed in this paper was done jointly with Josh Tenenbaum, Terry Regier, and Ted Gibson. Special thanks to David Beaver and James Gee for helpful comments on this manuscript.

References

Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science*, *32*(5), 789–834.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

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

Chater, N., & Vitányi, P. (2007). 'Ideal learning' of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology*, *51*(3), 135–163.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, *2*(3), 113–124.

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. (1980). *On cognitive structures and their development: A reply to Piaget* (M. Piattelli-Palmarini, Ed.). Cambridge, MA: Routledge & Kegan Paul.

Chomsky, N. (1995). *The minimalist program*. Cambridge, MA: MIT press.

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, *14*(1), 1–13.

Cox, R. T. (1961). *The algebra of productive inference*. Baltimore, MD: Johns Hopkins University Press.

de Finetti, B. (1974). *Theory of probability*. New York: J. Wiley & Sons.

Copyright © John Benjamins Publications

de Marcken, C. (1995). *Unsupervised language acquisition.* Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Dowman, M. (1998). *A cross-linguistic computational investigation of the learnability of syntactic, morphosyntactic, and phonological structure* (Tech. Rep. No. EUCCS-RP-1998-6). Edinburgh, UK: Edinburgh University, Centre for Cognitive Science.

Everett, D. L. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current Anthropology*, *46*(4), 621–646.

Everett, D. L. (2007). *Cultural constraints on grammar in Pirahã: A reply to Nevins, Pesetsky, and Rodrigues (2007).* Available from `http://ling.auf.net/lingBuzz/000427`

Feldman, J. A., Gips, J., Horning, J. J., & Reder, S. (1969). *Grammatical complexity and inference* (Tech. Rep. No. CS-TR-69-125). Stanford, CA: Stanford University.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, *27*(2), 153–198.

Goldsmith, J. (2007). Morphological analogy: Only a beginning. In J. P. Blevins & J. Blevins (Eds.), *Analogy in grammar: Form and acquisition.* Oxford: Oxford University Press.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2006). Interpolating between types and tokens by estimating power law generators. In *Advances in Neural Information Processing Systems* (Vol. 18). Vancouver, BC, Canada.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2007). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21-54.

Good, I. J. (1968). Corroboration, explanation, evolving probability, simplicity and a sharpened razor. *The British Journal for the Philosophy of Science*, *19*(2), 123–143.

Grünwald, P. (1996). A minimum description length approach to grammar inference. *Symbolic, Connectionist, Statistical Approaches to Learning for Natural Language Processing. Lecture Notes in Computer Science*, *1040*, 203–216.

Hauser, M. D., 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.

Copyright © 2010 John Benjamins Publications

Horning, J. J. (1969). *A study of grammatical inference* (Tech. Rep. No. 139). Stanford, CA: Stanford University.

Immerman, N. (1999). *Descriptive complexity.* Springer Verlag.

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

Jeffreys, H. (1931). *Scientific inference.* Cambridge: Cambridge University Press.

Jeffreys, H. (1939). *Theory of probability.* Oxford: Clarendon Press.

Johnson, M. (2008). Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *46th annual meeting of the Association for Computational Linguistics* (pp. 398–406). Columbus, OH.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition.* New York: Prentice Hall.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354.

Kornai, A., & Pullum, G. K. (1990). The X-bar theory of phrase structure. *Language*, *66*(1), 24–50.

Langendoen, D. T., Kalish-Landon, N., & Dore, J. (1973). Dative questions: A study in the relation of acceptability to grammaticality of an English sentence type. *Cognition*, *2*, 451–477.

Lappin, S., Levine, R. D., & Johnson, D. E. (2000). The structure of unscientific revolutions. *Natural Language and Linguistic Theory*, *18*, 665–671.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Lohse, B., Hawkins, J., & Wasow, T. (2004). Processing domains in English verb-particle constructions. *Language*, *80*(2), 238–261.

Louann, G., Rachel, W., & William, L. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, *32*(02), 249–268.

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

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

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

Nevins, A., Pesetsky, D., & Rodrigues, C. (2007). *Pirahã exceptionality: A reassessment.* Available from `http://ling.auf.net/lingBuzz/000411`

Copyright 2011 John Benjamins Publications

Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. *Language*, *70*(3), 491–538.

Perfors, A., Tenenbaum, J., Gibson, E., & Regier, T. (2010). How recursive is language? A Bayesian exploration. *The Linguistic Review*.

Perfors, A., Tenenbaum, J., & Regier, T. (under review). The learnability of abstract syntactic principles. *Cognition*.

Perfors, A., Tenenbaum, J., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, *37*, 607-642.

Peters, P. S., & Ritchie Jr., R. W. (1973). On the generative power of transformational grammars. *Information Sciences*, *6*, 49–83.

Pinker, S., & Jackendoff, R. (2005). The faculty of language: What's special about it? *Cognition*, *95*(2), 201–236.

Rogers, J. (1998). *A descriptive approach to language-theoretic complexity*. Stanford, CA: CSLI Publications.

Solomonoff, R. (1978). Complexity-based induction systems: Comparisons and convergence theorems. *IEEE transactions on Information Theory*, *24*(4), 422–432.

Spivey, M. J., & Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning Memory and Cognition*, *24*, 1521–1543.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Vitànyi, P., & Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE transactions on Information Theory*, *IT–46*, 446–464.

Wasow, T., & Arnold, J. E. (2005). Intuitions in linguistic argumentation. *Lingua*, *115*(11), 1481–1496.

Wolff, J. G. (1982). Language acquisition, data compression and generalization. *Language and Communication*, *2*(1), 57–89.

Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *42*(249), 369–390.

Copyright © CSLI Publications

Language from a Cognitive Perspective:
Grammar, Usage, and Processing
edited by Emily M. Bender and Jennifer E. Arnold

Published by CSLI Publications, Stanford,
and distributed by University of Chicago Press.

Information on buying the full book can be found at

http://cslipublications.stanford.edu/
or at
http://www.press.uchicago.edu/