Learnability, representation, and language: A Bayesian approach

by

Amy Perfors

B.S., Stanford University, 1999
M.A., Stanford University, 2000

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2008

Author .....
Department of Brain and Cognitive Sciences
September 17, 2008

Certified by .....
Joshua Tenenbaum
Associate Professor of Cognitive Science
Thesis Supervisor

Accepted by .....
Matthew Wilson
Professor of Neurobiology
Chairman, Committee for Graduate Students

Learnability, representation, and language: A Bayesian approach

by

Amy Perfors

Submitted to the Department of Brain and Cognitive Sciences
on September 17, 2008, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Within the metaphor of the "mind as a computation device" that dominates cognitive science, understanding human cognition means understanding learnability – not only what (and how) the brain learns, but also what data is available to it from the world. Ideal learnability arguments seek to characterize what knowledge is in theory possible for an ideal reasoner to acquire, which illuminates the path towards understanding what human reasoners actually do acquire. The goal of this thesis is to exploit recent advances in machine learning to revisit three common learnability arguments in language acquisition. By formalizing them in Bayesian terms and evaluating them given realistic, real-world datasets, we achieve insight about what must be assumed about a child's representational capacity, learning mechanism, and cognitive biases.

Exploring learnability in the context of an ideal learner but realistic (rather than ideal) datasets enables us to investigate what could be learned in practice rather than noting what is impossible in theory. Understanding how higher-order inductive constraints can themselves be learned permits us to reconsider inferences about innate inductive constraints in a new light. And realizing how a learner who evaluates theories based on a simplicity/goodness-of-fit tradeoff can handle sparse evidence may lead to a new perspective on how humans reason based on the noisy and impoverished data in the world. The learnability arguments I consider all ultimately stem from the impoverishment of the input – either because it lacks negative evidence, it lacks a certain essential kind of positive evidence, or it lacks sufficient quantity of evidence necessary for choosing from an infinite set of possible generalizations. I focus on these learnability arguments in the context of three major topics in language acquisition: the acquisition of abstract linguistic knowledge about hierarchical phrase structure, the acquisition of verb argument structures, and the acquisition of word leaning biases.

Thesis Supervisor: Joshua Tenenbaum
Title: Associate Professor of Cognitive Science

Acknowledgements

In some ways, this section is the one that I have the most trepidation about writing, both because I feel some pressure to be witty, and because I lack the eloquence to capture how much some of these people have helped me over the years. I'm also afraid of forgetting somebody: there is truly quite a list.

First and foremost, I would like to thank my advisor, Josh Tenenbaum. I entered his lab with a passionate if somewhat inchoate interest in many issues in language acquisition and cognitive development, and he helped me turn that into a well-defined set of questions and an interesting research program. When I arrived I knew relatively little about abstract computational theory, and even less about Bayesian modelling: only enough to think that they seemed to be promising and important tools for understanding induction and higher cognition. I have learned a truly tremendous amount from him about these topics. More profoundly, he has taught me, by example, how to be a scientist: how to communicate your ideas to others; how to strike the difficult balance between safe (but necessary) research and less-certain, more risky, questions; how to recognize a bad or fruitless idea before spending weeks or months investing time in developing it; and how to nurture the good ideas, even if they aren't yet fully worked out. In the process, he also taught me a little something about being a good person as well. I remember once I was talking to him about many of the structural difficulties inherent in sometimes doing good research – the necessity for funding, the tendency of fields to get sucked into myopic ways of thinking, etc – and finally he said, "I think you have to be the change you want to see." That's a cliché, but it's a *good* cliché, and Josh lives it: if you don't like something, try to change it: slowly, organically, from the inside. I try to emulate that about him.

The members of my committee – Ted Gibson, Laura Schulz, and Lila Gleitman – each affected me in their own way as well. Ted has an unparalleled (and extremely refreshing) ability to see through bullshit[1], coupled with the courage to call it out in no uncertain terms. I have benefited over the years from his thoughts about the

[1]Is cursing allowed in dissertations? I hope so, because no other word seems appropriate here.

intersection of linguistics and cognitive science, as well as his broader perspective about academia and how all of our research fits into the developing narrative of the field. Laura has been a consistent source of insight about the relevance of my work to the issues in developmental psychology, and has (very usefully) prodded me to think about these ideas in a broader, more big picture context. And I count myself to have been extremely lucky to have benefited from the ideas and insights of Lila Gleitman, who has been doing groundbreaking work on these issues for longer than I have been alive. When I first met her at a conference I was somewhat intimidated, but it only took a few thought-provoking, wide-ranging, and immensely engaging conversations for the intimidation to be replaced by something far more valuable. In her incisive intellect, irreverent attitude, and sharp focus on the interesting questions, she also is somebody I seek to emulate.

I have had many other academic influences, too many to list in their entirety, but some deserve unique mention. Several pre-college teachers (Bonnie Barnes, Marge Morgenstern, Jon Dickerson, Kathy Heavers, among others) fed a love for learning that has lasted me my whole life. Anne Fernald, my undergraduate advisor, is probably the main person responsible for taking my initial interest in cognitive science questions and showing me how one could study them scientifically. She was an inspirational mentor and source of support at a time when I greatly needed both. Fei Xu taught me a great deal about not only the important details of doing experimental work on infants, but also imbued me with a much richer sense of the issues and questions as they appear from an empirical developmental perspective. Due to her generosity in inviting me to spend several months in the UBC Baby Cognition lab, among colleagues with an interest in similar issues as me but a very different approach to solving them, my intellectual boundaries were greatly expanded. Thanks in part to her influence, I now have a much deeper fascination with issues about early conceptual structure and essentialism, and I hope to explore them in the future. Another source of intellectual growth has been my coauthors, Terry Regier and Liz Wonnacott, each in their own way: Terry for his penetrative ability to hone in on and identify the real issues and interesting questions, Liz for her ability to ground my

theoretical approach in the important empirical questions, as well as her friendship. I have also been lucky enough to benefit from interesting, provocative conversations with a wide range of people, whose serious consideration of my ideas and the issues has greatly enhanced my thinking: Robert Berwick, who has kept me honest, and from whom I have learned a great deal of mathematical learnability theory and the history of the field of linguistics; Jay McClelland and Jeff Elman, both of whom have caused me to think much more deeply about the relations between and contributions of Bayesian methods, PDP models, and dynamical systems; and many, many others, including Adam Albright, Tom Griffiths, Steven Pinker, Ken Wexler, Danny Fox, Morten Christiansen, Jesse Snedeker, Susan Carey, Linda Smith, and Dan Everett.

Many of my colleagues associated with the Computational Cognitive Science Lab have also had a profound effect on my thinking. Charles Kemp, a coauthor, has been a tremendous source of instruction in everything from the nitty-gritty details of computational modelling to the deepest, most intractable questions in cognitive science. He has been one of my favorite sparring partners over the years, and is also a source of some of my favorite memories – road-tripping to Vermont, stumbling around Italy, listening to Bach cantatas in church. Mike Frank has inspired me with his ability to come up with simple and elegant ways to investigate deep and interesting problems, and several conversations with him have helped to crystallize issues in my mind. Vikash Mansinghka, despite having a last name that I can never remember how to spell, has also been an inspiration for his energy, deeply penetrating intellect, and profound moral sensibility (though he would probably laugh to know I said that about him). Noah Goodman and Tim O'Donnell have been great sources of insight about computation and linguistics. Liz Baraff-Bonawitz is not only a friend but also a powerful example to me of what sort of good science can emerge from somebody with wide-ranging interests, an incredible work ethic, and the passion and intensity to make it happen. And finally, I really couldn't have done this without Lauren Schmidt: her support when things went wrong, her encouragement when things went well, her scientific insights, her capacity for fun, her personal friendship. I will miss her tremendously. Many others, too many to list (but including Sara Paculdo, Virginia

Savova, Ed Vul, Steve Piantadosi, Ev Fedorenko, Konrad Koerding, and Mara Breen), have helped in other ways.

On a more personal level, no acknowledgements section would be complete without mentioning my partner, Toby Elmhirst. He has helped me in pretty much every way a person can be helped. Intellectually, he has given me the perspective of an intelligent analysis from the perspective of another field (in this case, mathematical biology), which has provided me with a much richer sense of how the deep issues in one area of human endeavor are closely related to the issues in other. He is my favorite person to talk to about any topic, scientific and otherwise, and I consider myself profoundly lucky to have found somebody with whom conversation never gets old. Emotionally, he has been my strength when things were tough, and has shared my joy when they were not. If I were to dedicate this thesis to anybody, it would be to him.

I also could not finish this without thanking my family. My parents, George and Diane Perfors, have influenced me in all the ways good parents should. I first realized how to think and how to be intellectually honest from them; how to take joy in the small things; to treasure the outdoors and small towns; and to realize that everybody, from the most famous scientist to the lowest beggar, can have something to teach. Because of their influence, I learned to trust in myself and not be swayed by fashion or superficial pressures (or, at least, to try not to be). It has been a great source of emotional strength to know that their love and support are not conditioned on earning a PhD, being a scientist, or doing anything in particular. My siblings (Dave, Tracy, Steve, and Julie) have not only been lifelong friends, but also a consistent source of perspective and joy. This PhD and these academic pursuits, which I truly love and I do think are very interesting, are still much less important than family and friends and the rest of life. And I thank them for continuously reminding me of that; ironically, I think it may even have made the science a bit better.

I know every acknowledgements section since the dawn of time says this, but that is because it is so true: in a very real sense, this research is not the product of one person's toil, but the result of many, many people's efforts. This is where I get to say thank you to them.

Contents

Chapter 1

The problem of learnability

Cognitive science views the mind from a computational perspective: as a thinking machine that, given some input, performs computations on it and produces a behavior. Within this paradigm, which is broadly accepted by almost every cognitive scientist nowadays, the fundamental questions are centered on the nature of the data in the world and the nature of the computations the mind can perform.

As a result, the subject of learnability is of fundamental importance. If we want to understand what computations the mind performs, we must understand its dependence on data – which means understanding both *what* data exists and *how* the mind might be able to use and learn from it. This is difficult in part because both of these unknowns are deeply enmeshed with each other: what data exists depends in part on how the mind operates, and how the mind performs may depend to some extent on what data it receives. Data depends on the mind because it is only available to the extent that the brain is capable of perceiving and interpreting it: a newborn child would not receive the same *effective* data as a newborn even if we could ensure that every aspect of their environment was identical, since a three-year-old can understand and use the input (e.g., the ambient language) in a way that a newborn cannot. Conversely, the mind depends on data in some ways. Every normal brain has the same potential capacity, but exercising that capacity may require the correct input; a child never exposed to English will not be able to produce or understand it.

These may look like trivial observations, but that is because they are focused on

the extremes. Unfortunately for us, most of cognition does not operate at the extremes – and that hazy middle ground is where it becomes difficult to tease apart which aspect of behavior is due to the brain and which is due to the data. In some ways, in fact, such a question verges on incoherence, since everything is cumulative: what the mind is like at birth influences the effective data the brain receives, which in turn affects what is learned, which then further shapes the effective data, and so on. The fact that human cognition doesn't appear to be highly sensitive to initial conditions in the way that many feedback loops are doesn't mean that cognitive scientists need not worry about cumulative effects; it might simply imply that everyone's brain and/or environment are broadly comparable enough in the relevant respects that the feedback loop operates quite similarly for all.

Another complication lies in the fact that learnability and representation are intimately intertwined. Whether something is learnable depends a great deal on what, precisely, is thought to be learned: do children avoid certain ungrammatical constructions because they have abstracted some form of structured syntactic knowledge, or because they have implicitly learned associations involving unstructured representations? In general, as representational system grows in expressivity, it should be more difficult to learn within that system, so more may need to be built in from the start. But if what is built in is a powerful learning mechanism, does this count as "nature" or "nurture"? And if the representation is quite expressive and powerful, does this count as building in a lot (because it is so complicated) or a little (because the space of unknowns is larger)?

Thus arises the perennial debates about nature and nurture. The simplistic view that both nature and nurture are important is – however true – not very constructive or explanatory from the scientific point of view. A major challenge in cognitive science, therefore, is to move beyond both extremes as well as this glib middle ground. We must work to elucidate exactly *how* innate biases and domain-general learning might interact to guide development in different domains of knowledge.

The study of ideal learnability is a valuable means towards that end. Ideal learnability arguments focus on the question of whether something is learnable in principle,

given the data: that is, they are more focused on the nature of the input than on the nature of the mind. The two cannot be completely disentangled for all of the reasons we have already considered – and those issues often prove to be the rocks against which the waves of disagreement break – but ideal learnability analyses at least provide a path towards answering what looks like an otherwise nearly intractable question.

What does it mean to be learnable in principle? After all, everything is learnable in the sense that it is possible to make some machine that can "learn" it by simply building it in. Learnability analysis attempt to avoid this intellectual vacuity by describing the characteristics of an ideal learner and then evaluating whether it could acquire the knowledge under consideration. This approach guides inquiry in a fruitful direction by suggesting a series of precise, testable questions: do humans really have the hypothesized characteristics? Are the assumptions made about the data accurate? Do the ways in which the analysis departs from reality prove critically important? If so, why? If not, why not? One value of learnability analyses is that they provide a way of converting intractable, vague questions like "is it nature or nurture?" into tractable, empirically resolvable ones.

That means, of course, that a learnability analysis is only as good as the assumptions it makes about the data and the learner. Formal mathematical proofs are valuable starting points but often have limited applicability because the assumptions they make in order to be mathematically tractable can be quite distant from the characteristics of human learners. Non-mathematical analyses, by contrast, have historically been somewhat constrained by technical limitations on their ability to evaluate the behavior of a powerful learner on a complex real-world dataset. Because of the difficulty in gathering and manipulating large datasets of realistic input, they frequently involve educated but intuitive guesses about what data would be relevant as well as how rich the data really is. Because they are often not formalized they may either contain implicit suppositions about the nature of the postulated learning mechanism, or draw vague or unjustified conclusions about what precisely must be innate. And because they are seldom implemented and tested, it can be difficult to

objectively evaluate what might *actually* be learned by given their assumptions.

Recent advances in machine learning and computer science provide a way to surpass some of these limitations. In this thesis I exploit these advances to revisit three of the most common learnability arguments in linguistics and cognitive science. The Bayesian framework I use produces novel answers for some of the specific questions to which these arguments have applied. More broadly, it yields insight about the types of conclusions that can be drawn from this sort of argumentation, and why. And in each case, it fulfills one of the primary goals of an ideal learnability analysis – providing a set of precise, empirically testable questions and predictions.

In the subsequent sections I review each of the three common learnability arguments this thesis is focused on. I describe the abstract logic of each with the goal of clarifying what assumptions are made and what conclusions may be validly drawn. I then discuss three areas in language acquisition, corresponding to Chapters 3, 5, and 4, where we see these learnability arguments play out. These chapters analyze the learnability claims with respect to the specific examples and evaluate the implications for other arguments with the same logical structure.

Three learnability problems

Problem #1: The Poverty of the Stimulus

Plato's dialogue *Meno* introduces us to the sophistic paradox, also known as the problem of knowledge: "man cannot enquire either about that which he knows, or about that which he does not know; for if he knows, he has no need to enquire; and if not, he cannot; for he does not know the very subject about which he is to enquire." Socrates' solution to this dilemma is to suggest that all knowledge is in the soul from eternity and simply forgotten at birth: learning is simply remembering what was already innately present. His conclusion is based on one of the first Poverty of the Stimulus (PoS) arguments, in which he demonstrates that a slave-boy who had never been taught the fundamentals of geometry nevertheless grasps them.

PoS arguments like this are used quite generally to infer the existence of some innate knowledge, based on the apparent absence of data from which the knowledge could have been learned. This style of reasoning is as old as the Western philosophical tradition. 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 mental capacities based on an apparent absence of support for acquiring them through learning. Thus, the Poverty of the Stimulus is both a problem for the learner – how to generalize based on limited or impoverished data – as well as a method for research: identifying where data is impoverished can be a useful step toward identifying what knowledge must be innate. The logical structure of the PoS argument is as follows:

1.1. (i) Children show a specific pattern of behavior B.

(ii) A particular generalization G must be grasped to produce behavior B.

(iii) It is impossible to reasonably induce G simply on the basis of the data D that children receive.

(iv) $\therefore$ Some abstract knowledge T, limiting which specific generalizations G are possible, is necessary.

This form of the PoS argument is applicable to a variety of domains and datasets both within and across linguistics. Unlike other standard treatments (Laurence & Margolis, 2001; Pullum & Scholz, 2002), it makes explicit the distinction between multiple levels of knowledge – a distinction which we will see again and again throughout this thesis. 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:

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

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

(iii) $\therefore$ T must be innate.

The problem of the Poverty of the Stimulus is the most general of the three learnability problems discussed in this thesis; in fact, it would not be inaccurate to say that the other two are simply special cases. Saying that the stimulus is impoverished with respect to some generalization is synonymous with saying that the generalization is not learnable without presuming that the learner has access to some specific abstract knowledge T. Learnability arguments differ from one another based on how they answer the question of precisely *why* or *in what way* the data is impoverished. Many PoS arguments – including the classic one about hierarchical phrase structure in language, which I discuss in Chapter 3 – focus on a lack of a certain kind of positive evidence, which is otherwise perceived as necessary for ruling out incorrect generalizations or hypotheses. This sort of PoS argument could apply even if only a finite amount of data would otherwise be necessary: it is about the *nature* of the data, rather than the *amount* of it.

By contrast, the No Negative Evidence and the Feature Relevance problems – the other two problems addressed in this thesis – gain most of their logical force because no human learner sees an infinite amount of data. In short, they are more about quantity of data than quality (especially in the case of the Feature Relevance problem). A lack of negative evidence would not be a problem, at least in the limit, if a dataset were infinite in size,[1] because a learner could simply assume that if certain input is unattested, that is because it is not allowed. For a similar reason, the Feature Relevance problem is only unresolvable in principle if there are an infinite number of features that could possibly *be* relevant – otherwise, in theory at least, a learner could simply eliminate the irrelevant features one-by-one.

Problem #2: No Negative Evidence

In a seminal 1967 paper, mathematician E. Mark Gold provided a mathematical analysis of learnability in the limit. His question was whether a learner with an infinite amount of linguistic data would be able to converge on the correct language (that is,

[1]This is true assuming a hypothesis space that is infinite in size, which (as we will see later) is the case with many interesting acquisition problems.

eventually produce only grammatical strings of that language). Gold considered several variations on the basic paradigm, but for our purposes, the relevant ones concern differences how linguistic information is presented. In one variation, an informant presents the learner with a list of strings and labels each as grammatical (positive evidence) or ungrammatical (negative evidence). In the other variation, the learner is simply shown a list of strings under the assumption that all are grammatical. Both versions incorporate the constraint that each string occurs at least once, but no other assumptions about the presentation are made.

The learning algorithm involves successively testing and eliminating candidate grammars. After each string, the learner decides whether that string could have been generated by its current grammar. If so, the grammar is retained; if not, it is discarded, and a new one is chosen. Because the listener is assumed to have a perfect memory, the new grammar will always be one that can generate not only the current string, but also all of the previous strings that have been. In many ways, this learner has more advantages than any human being: in addition to its perfect memory and ability to match new grammars on the fly to all of the strings it has already seen, its input contains no errors.

In spite of this, it turns out that only languages of finite cardinality are learnable from positive evidence alone. The reason is that for any infinite language it is impossible to determine how to generalize beyond the strings that have been seen. If the learner were to select a grammar that produced all and only those strings, it might miss a string that is in the language but has not yet appeared; but if it were to choose a grammar that could produce some string(s) that had not been seen, they might not be in the language after all. Only if the language is finite can the learner converge on a grammar that produces all and only the strings in the language.

This poses a problem from the point of view of language acquisition in humans, since human languages are not finite (Chomsky, 1959). If even a learner with a perfect memory and errorless input can't acquire the correct grammar, how do children do it? One possibility might be that they receive both positive and negative evidence; Gold's analysis showed that under such conditions, the classes of languages that

include English are all learnable. The problem is that there is little substantiation for the idea that children actually receive much negative evidence (e.g., Brown & Hanlon, 1970; Newport, Gleitman, & Gleitman, 1977; Pinker, 1989), and if they do, not much indication that they notice or use it (McNeill, 1966; Braine, 1971). At most, some studies suggest that there are slight differences in the frequency of parents' corrections of well-formed vs. ill-formed utterances (e.g., Bohannon & Stanowicz, 1988; Chouinard & Clark, 2003), which – even if the child could use it to figure out what aspect of the utterance was ill-formed – is still probably not sufficient to reject all of the possible incorrect grammars (Gordon, 1990; Marcus, 1993).

In general, Gold's theorem demonstrates that without some constraint on the space of languages or the procedure used to learn them – or both – language acquisition from positive-only evidence is impossible. What sort of constraints might help? Gold shows that certain restrictions on the order of presentation can make even recursively enumerable languages learnable. Unfortunately, these restrictions (which require the presentation to be generated by a primitive recursive function) are implausible when applied to human language learning. Theorists have explored many other avenues, and I will give two in particular more detailed attention: allowing grammars to be probabilistic, and incorporating a less stringent standard of learnability.[2]

There are many ways to incorporate a less stringent learnability standard (e.g., J. Feldman, 1972; Wharton, 1974), but one of the most well-developed and useful is the Probably Approximately Correct (PAC) framework introduced by Valiant (1984). The PAC approach translates the learnability problem into the language of statistical learning theory, so that each language is associated with an indicator function that maps each of the strings to a real number between 0 and 1 (where a 1 indicates that the string belongs in the language). In order to be able to evaluate "how close" a language is to the correct one, a distance metric is imposed on the space of languages – so that instead of scoring 1 if the learner has guessed the correct language and 0 otherwise, the learner "gets credit" for being close. A class of languages is learnable within this framework only if it has finite VC dimension (Vapnik & Chervonenkis, 1971); this

[2]See (P. Niyogi, 2006) for an overview.

implies, remarkably, that *even finite languages* (as well as regular and context-free languages) are not learnable in the PAC setting. Although some languages are PAC learnable but not Gold learnable, it does not seem as if we can rely on PAC learning to solve the learnability problem.

What about allowing grammars to be probabilistic? It might seem that this would facilitate learning, and indeed it does in certain special cases, but in the abstract it actually makes the problem harder:[3] rather than having to simply converge on the correct set of grammatical rules (or the correct extension of sentences, as in Gold's formulation), the learner must now converge on the correct set of rules *and probabilities*. In fact, if nothing is known *a priori* about the nature of the probability distribution on rules – call it μ – then making the languages stochastic does not expand the class of learnable languages at all (Angluin, 1988; P. Niyogi, 2006). If, however, we can make certain assumptions about μ, then the entire class of recursively enumerable languages – which includes human languages – becomes learnable (Osherson, Stob, & Weinstein, 1986; Angluin, 1988). Are these assumptions plausible for human language?

It is difficult to be certain, but many have argued that they are probably not. The essential idea is that μ must be a member of a family of approximately uniformly computable distributions (Angluin, 1988). A family of distributions is approximately uniformly computable if the distribution on the strings so far can be approximated within some error by every individual μ_i in the family. This imposes a fairly stringent constraint: for instance, probability measures are obtained on context-free grammars by tying the probabilities to context-free rules. This imposes an exponential-decay distribution in which longer strings are exponentially less frequent than shorter strings. As a result, a learner who (correctly) assumes that the distribution is of this form will converge to the correct context-free grammar (Horning, 1969), but a learner assuming an arbitrary distribution may not.

[3]The same general point applies to the idea of jointly learning grammars (i.e., syntax) and meanings (i.e., semantics) - this simply makes the space larger. It may be possible that syntax and semantics might mutually constrain each other in such a way that it is easier to learn both, but this is not obviously true.

Horning's work is interesting not only because it demonstrates a positive learnability result for context-free grammars – albeit one that depends on the assumption that individual strings follow a exponential-decay probability distribution – but also because it incorporates notions from Bayesian probability theory, which is related to research in information theory based on the notion of Minimum Description Length (MDL). Both Bayesian and MDL approaches are based on the insight that incorporating a simplicity metric can provide a way to choose among all of the grammars (or, more generically, hypotheses) that are consistent with the data. Indeed, a remarkable proof by Solomonoff (1978) demonstrates that a learner that incorporates a certain simplicity metric will be able to predict any computable sequence with an error that approaches zero as the size of the dataset goes to infinity (see also Solomonoff, 1964; Rissanen & Ristad, 1992; Chater & Vitànyi, 2007). This is in some sense the perfect universal prediction algorithm. The drawback? It is not computable, meaning that it would take an infinite amount of time to calculate. Thus, although it is reassuring in an ideal sense, this says little about how children actually overcome the No Negative Evidence problem.

A different idea, called the Subset Principle, suggests that children manage to learn from positive-only evidence by ranking their hypothesis grammars in such a way that they can be disconfirmed by positive examples (Berwick, 1985). Thus, they would consider the narrowest languages first – i.e., grammars whose extensions are strict subsets of all of the grammars later in the ordering – and only if those grammars were disconfirmed would the next largest grammar be evaluated. According to mathematical learning theory, the Subset Principle is both necessary and sufficient for convergence to the correct language (at least in the limit). Problem solved? Unfortunately, no: it does not explain certain empirical phenomena of overgeneralization in language – people simply do not seem to be conservative learners in the way that a Subset Principle based learner would be. But how, then, do human learners solve the No Negative Evidence problem?

Problem #3: Feature Relevance

The problem of determining Feature Relevance, which is related but logically distinct from the No Negative Evidence problem, was first explicated by early philosophers. In one classic thought experience, Quine (1960) asked readers to imagine they were an anthropologist visiting a remote tribe, trying to learn the language. One day he sees a rabbit run by, and a one of the tribesman points to it, saying "gavagai." Can he infer that "gavagai" means *rabbit*? It certainly seems the natural conclusion, but there are logically an infinite number of possible meanings: it might mean *rabbit fur*, *undetached rabbit parts*, *rabbit running on grass*, *soft furriness in motion*, or even *the abstract concept of rabbits, as a species or general type, here represented by this example.* In fact, even if some meanings are ruled out by context, there will always remain an infinite number of possibilities consistent with the observations so far.

Goodman (1955) discussed a similar issue in the context of the famous "grue" problem. In this example, one problem of induction (determining the referent of a word) is replaced by another (determining the extension of a predicate). Having observed that all emeralds examined thus far are green, it is tempting to conclude that all future emeralds will also be green, particularly as one observes more and more green emeralds. Yet, as Goodman pointed out, it is equally true that every emerald that has been observed is grue: emerald before time t, and ruby after time t (assuming t has not yet occurred). Why, then, do we not conclude that emeralds are grue rather than green?

Both of these classic problems of induction have the a similar flavor: how to identify the correct hypothesis out of a potentially infinite set of possibilities. In some ways, the No Negative Evidence problem grapples with a similar issue as these, but it is important to note that it is logically distinct. The Feature Relevance problem is about how one decides which of an infinite number of dimensions or features are relevant to generalize along: Color? Color at time t? Color before and after time t? Objecthood? Furriness? Grassiness? The No Negative Evidence problem, by contrast, would apply even if there were only one feature: it is a question about how,

and to what extent, one should generalize beyond that input in the feature(s) already identified as relevant. This is an important distinction to keep in mind because the two problems are often conflated, leading to confusion about what has and hasn't been resolved.

Overview

All three of these learnability problems are essentially Poverty of the Stimulus problems at root: they simply differ in terms of the way in which the data is impoverished. The Feature Relevance problem arises from the fact that no finite dataset contains sufficient quantity of input to rule out which of an infinite number of possible features it should be generalized along. The No Negative Evidence problem arises because without negative evidence,, a learner cannot decide among all of the hypotheses that are consistent with the input. It differs from the Feature Relevance problem because the issue is no longer in deciding which of an infinite number of possible features are relevant; the issues is that, even *given* the feature (e.g., formal syntax in Gold's original formulation), it is impossible to logically eliminate hypotheses that overgeneralize beyond the input. Both of the Feature Relevance problem and the No Negative Evidence problem are PoS problems that depend critically on hypothesis spaces that are infinite in size. Another kind of PoS problem emerges from the lack of a certain kind of positive data, and would exist even if the hypothesis space were finite. Thus, although each of the three learnability problems are related, they are also distinct, and throughout this thesis we will see how these distinctions play out among particular examples in acquisition.

One commonality among all of these learnability problems – indeed, all learnability problems in general – is that they are simultaneously a *problem* and a *method*: a problem confronting a learner, and a method for the scientist to infer what sort of knowledge must be assumed in order to solve the problem. This knowledge takes the form of higher-order constraints (T) of some sort which, following Goodman (1955) I will sometimes also refer to as overhypotheses. The logic of these learnability problems requires only that some higher-order constraint *exists*, not that it be innate, although

it can be difficult to imagine a way in which it might be learned (especially *before* the lower-level knowledge it is meant to constrain).

Many types of higher-order constraints have been hypothesized in different domains and for different learning problems. All that the logical of the argument requires is that there be some constraint, not that it be domain-specific, nor even that it be knowledge *per se*: constraints due to perception, memory, or attention would also suffice. To choose a trivial example, no human categorizes objects according to their color along the ultra-violet spectrum; this is a higher-order constraint that limits which hypotheses about object categorization are considered, but it emerges because our perceptual systems do not represent ultraviolet colors, not because of anything cognitive or knowledge-based *per se*.

Of course, most of the time the natural response to these learnability problems is to hypothesize higher-order constraints that are both innate and (often) domain-specific. For instance, core systems of object representation (e.g., Carey & Spelke, 1996; Spelke & Kinzler, 2007) are theorized to explain why infants assume that objects obey spatio-temporal principles of cohesion, continuity, and contact (Spelke, 1990; Aguiar & Baillargeon, 1999). Core knowledge of psychology is used to explain why babies believe that agents are distinct from objects in that they can move without contact (Spelke, Phillips, & Woodward, 1995) and act in certain ways in response to goals (Woodward, 1999; Gergely & Csibra, 2003). Higher-level constraints on grammatical generalizations may be one aspect of, or play the role of, Universal Grammar (Chomsky, 1965). Constraints may also be learning algorithms, as in the Subset Principle (Berwick, 1985), or innate conceptual machinery, as in the case of Pinker's semantic bootstrapping hypothesis for verb learning (Pinker, 1989).

In the next section I will address three particular areas in language acquisition and cognitive development, corresponding to Chapters 3, 4, and 5 respectively, where we see these learnability problems play out in different ways.

Learnability in acquisition and development

Auxiliary fronting: Learning abstract syntactic principles

One learnability problem that has been much debated in cognitive science and linguistics concerns the phenomenon of auxiliary fronting in constructing English interrogative sentences (Laurence & Margolis, 2001; Lewis & Elman, 2001; Legate & Yang, 2002; Pullum & Scholz, 2002; Reali & Christiansen, 2005). This is an example of a PoS problem that results from the lack of a certain kind of positive data, namely, complex polar interrogatives. I consider this specific example with two larger goals in mind: to begin an exploration of the logical structure of poverty of stimulus arguments; and to address the general phenomenon of hierarchical phrase structure in syntax more specifically – the phenomenon that has been argued to underlie the learning of auxiliary fronting and many other specific rules.

At the core of modern linguistics is the insight that sentences, although they might appear to be simply linear sequences of words or sounds, are built up in a hierarchical fashion from nested phrase structures (Chomsky, 1965, 1980). The rules of syntax are defined over linguistic elements corresponding to phrases that can be represented hierarchically with respect to one another: for instance, a noun phrase might itself contain a prepositional phrase. By contrast, in a language without hierarchical phrase structure the rules of syntax might make reference only to the individual elements of the sentence as they appear in a linear sequence. Henceforth, when I say that "language has hierarchical phrase structure" I mean, more precisely, that the rules of syntax are defined over hierarchical phrase-structure representations rather than a linear sequence of words. Is the knowledge that language is organized in this way innate? In other words, is it a part of the initial state of the language acquisition system and a necessary feature of any possible hypothesis that the learner will consider?

Chomsky (1965, 1971, 1980) put forth several arguments for this position, most famously one based on the phenomenon of auxiliary fronting in English. English interrogatives such as "Is the man hungry?" correspond to declaratives with a fronted main clause auxiliary like "The man is hungry": the auxiliary is at the beginning of

the interrogative appears to map to the is in the middle of the declarative. One might consider two possible rules that could govern this correspondence between declarative and interrogative forms:

1.3. (a) Linear: Form the interrogative by moving the first occurrence of the auxiliary in the declarative to the beginning of the sentence.

(b) Hierarchical: Form the interrogative by moving the auxiliary from the main clause of the declarative to the beginning of the sentence.

The linear rule 1.3(a) can be implemented without reference to the hierarchical phrase structure of the sentence, but the hierarchical rule 1.3(b) cannot. We know that the actual grammar of English follows principles much closer to the hierarchical rule 1.3(b), but how is a child to learn that such a rule is correct as opposed to a linear rule such as 1.3(a)? Although the linear and hierarchical rules result in the same outcome when applied to simple declarative sentences like "The man is hungry", they yield different results when applied to more complex declaratives such as this:

1.4. The man who is hungry is ordering dinner.

The linear rule predicts the interrogative form in (1.5.a), while the hierarchical rule predicts the form in (1.5.b):

1.5. (a) * Is the man who hungry is ordering dinner?

(b) Is the man who is hungry ordering dinner?

Of course, 1.5(b) is grammatical in English while 1.5(a) is not. This difference could provide a basis for inferring the correct rule: if children learning language hear a sufficient sample of grammatical sentences like 1.5(b) and few or no ungrammatical sentences like 1.5(a), they might reasonably infer that the hierarchical rule rather than the linear rule correctly describes the grammar of English. Yet Chomsky argued that complex interrogative sentences such as 1.5(b) do not exist in sufficient quantity in child-directed speech, going so far as to assert that "it is quite possible for a person to go through life without having heard any of the relevant examples that would choose

between the two principles" (1971). In spite of this paucity of evidence, children three to five years old can form correct complex interrogative sentences like 1.5(b) but appear not to produce incorrect forms such as 1.5(a) (Crain & Nakayama, 1987).

Chomsky further argued that on *a priori* grounds, a general-purpose learning agent who knows nothing specifically about human natural languages would take the linear rule to be more plausible by virtue of its simplicity: it does not assume either the existence of hidden objects (e.g., syntactic phrases) or of a particular organization (e.g., hierarchical rather than linear). If the correct rule cannot be learned from data and is also dispreferred due to a general inductive bias favoring simplicity, the logical conclusion is that children come equipped with some powerful language-specific innate mechanisms that bias them to learn syntactic rules defined over hierarchical rather than linear structures.[4]

Word learning: The problem of reference

The issue of determining the referent of a word is one of the prototypical learnability problems confronting children as they develop. As in other domains, most theories of how to overcome this problem hypothesize some sort of innate overhypothesis or higher-order constraint. For instance, the whole object constraint is theorized to describe why children prefer to apply words to whole objects rather than parts (Heibeck & Markman, 1987; Markman, 1990). Other constraints include the mutual exclusivity principle, which suggests that children assume that objects have only one label, and the taxonomic assumption, under which children assume that labels pick out objects organized by kind rather than thematically (Markman, 1990).

These and other overhypotheses seem reasonable, and there is considerable evidence in favor of them, but it is also clear that they cannot be the whole story. After

[4]In his formulation, Chomsky was concerned with the question of structure dependence (whether the syntactic rules were defined over the structures of the language) rather than the question of what those structures were (i.e., hierarchical or linear). The analysis here is not an attempt to argue against the innateness of structure dependence *per se*, which inherently depends on the notion of movement rules. I do not evaluate linguistic representations that incorporate any such rules, and therefore am focused here more on the second question, (although the two are clearly quite closely intertwined, and the general framework might be interestingly applied to that question).

all, we *do* learn the names of parts of objects eventually (presumably one does not need to see a detached arm or head in order to acquire those terms!). Some of the words children learn violate the mutual exclusivity assumption: a child's pet may be both an *animal* and a *dog* (and, for that matter, a *pet*). While one could respond to these objections by modifying the constraints to be soft and allow exceptions, precisely specifying *which* exceptions to allow, and when, is nontrivial indeed.

There is deeper problem with the idea that the problem of reference is solved by assuming the existence of innate overhypotheses about word learning: namely, that at least some overhypotheses appear to be learned. For instance, consider the shape bias: by the age of 24 months old, English-learning children tend to assume that count nouns are organized by shape. When given a novel object paired with a novel label like "dax", they are likely to generalize it to items that are similar in shape but not texture or color (Landau, Smith, & Jones, 1988; Soja, Carey, & Spelke, 1991). There are many compelling reasons for believing that this bias is learned rather than innate. The distribution of count nouns in English is organized by shape, and children do not acquire the shape bias until their vocabulary reaches a certain size (Samuelson & Smith, 1999, 2000). Teaching them additional words before that point makes them not only acquire the bias earlier, but also results in faster learning of *other*, non-taught words (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). Furthermore, the shape bias does not appear to be the only overhypothesis about feature generalization that children apply to word learning: while shape alone is a good cue to category membership for artifacts, texture seems as good as shape for animal categories, and color rather than shape appears most appropriate for foods (e.g., Booth & Waxman, 2002; Macario, 1991).

Some have suggested that the shape bias is a matter of attentional learning, and emerges simply from noting the distributional overlap between count nouns and their referents (Smith, Jones, Yoshida, & Colunga, 2003; Colunga & Smith, 2004, 2005; Smith & Samuelson, 2005). Under this view, words are simply features that happen to correlate strongly with shape, and children learn this correlation based on simple association. Another explanation holds that word labels are a cue to kind or

category membership, and shape is useful because it is also a reliable cue (Soja et al., 1991; Xu, Carey, & Quint, 2004). Under this explanation, the shape bias is not linguistic *per se*: rather, it emerges from the ontological commitments children make concerning the nature of kinds (Xu, 2002; Xu et al., 2004; Bloom, 2000; Markson, Diesendruck, & Bloom, 2008). Thus, although this latter account certainly argues for more innate machinery, what is assumed innate is not the shape bias (or any particular feature bias) itself, but rather the assumptions the child brings to the problem of word learning that *result* in the shape bias.[5]

Another way to characterize this debate, therefore, might be that the two sides disagree about the nature of the *over*-overhypotheses that govern the acquisition of the shape bias – or, put another way, they differ on the nature of the innate knowledge that guides the acquisition of the higher-level constraints T.[6] One side states this explicitly, arguing that the contention under debate concerns what assumptions children make about how labels and shapes pick out kinds. But even the distributional account, though it claims that no strong or domain-specific assumptions are necessary, limits the features considered; they include shape and some others (size, color, texture), but not an infinite number, and the account is largely about how a learner determines which of *those* features is relevant. Everyone has implicitly accepted that something that is itself an overhypothesis (i.e., the shape bias) can be learned; they just disagree about what assumptions are necessary to explain how.

However, there is a hint of a paradox at the center of the notion of learned inductive constraints that we need to confront: how can an overhypothesis be learned before the things it is constraining and still act to constrain them? Does lower-level learning always occur before higher-level learning, as one might expect, or is it pos-

[5]While many of these researchers suggest that shape for young children is salient and important even in the absence of word learning, this is not due to some sort of (innate) general perceptual salience applied indiscriminately; rather, it is salient precisely when children have reason to believe it is a valid cue for kind categories (e.g., S. Gelman & Ebeling, 1998; Bloom & Markson, 1998; Diesendruck & Bloom, 2003). Indeed, although these researchers have not proposed a learning mechanism for how children realize that shape is a cue for category membership, they seem broadly sympathetic to the idea that one exists (e.g., Soja et al., 1991; Markson et al., 2008).

[6]Although, as we shall see in Chapter 4, these are not over-overhypotheses that are cached out directly in the model we present. I use the notion here of over-overhypothesis in a more general sense, as a constraint on a constraint that is itself learned.

sible to acquire an overhypothesis faster than the specific hypotheses it is meant to constrain? If not, how can we explain the shape bias? If so, what principles explain this acquisition? And what, if anything, does this imply about the inductive constraints that were previously assumed must be innate?

Baker's Paradox: The case of verb argument constructions

Another classic learnability problem in language acquisition concerns the generalization patterns of verb argument constructions (Baker, 1979; Bowerman, 1988; Pinker, 1989). Verbs vary syntactically as well as semantically: different verbs take arguments in distinct patterns, or constructions. For instance, a verb like *love* is associated with the transitive construction, which requires the verb to take an NP object (e.g., "He loves her"). Different verbs are associated with different constructions, and often cluster in recognizable patterns. Consider the following English sentence pairs:

1.6. (a) Dave gave a gift to Laura. / Dave gave Laura a gift.

(b) Tracy sent an e-mail to Rob. / Tracy sent Rob an e-mail.

(c) Steve told a joke to Lauren. / Steve told Lauren a joke.

(d) Julie read a book to John. / Julie read John a book.

You might expect, based on these, that an acceptable generalization would be to say that anything that can occur in the first construction (the prepositional dative) is also found in the second (double-object dative). Unfortunately, some verbs occur in one construction only:

1.7. (a) Toby reported the loss to Sara. / * Toby reported Sara the loss.

(b) Diane said "okay" to George. / * Diane said George "okay."

This is a classic example of the No Negative Evidence problem: though children are never told that the starred sentences are incorrect, they eventually learn to avoid them. This particular pair of constructions, called the dative alternation, is just one example; three other common alternations, extensively discussed by Pinker (1989) and elaborated in Levin (1993), among others, are:

1.8. Passive alternation

(a) Ben kicked Stephanie. / Stephanie was kicked by Ben.

(b) Ben resembled Stephanie. / *Stephanie was resembled by Ben.

1.9. Causative alternation

(a) The box opened. / Melissa opened the box.

(b) The dog barked. / * Melissa barked the dog.

1.10. Locative alternation

(a) Kelly loaded hay into the truck. / Kelly loaded the truck with hay.

(b) Kelly pushed hay into the truck. / * Kelly pushed the truck with hay.

Baker (1979), one of the first to point out this problem, claimed that children never produce the incorrect constructions: "the speech of children contains virtually no examples in which they overgeneralize the double-NP construction to verbs that do not allow it in the adult language." (p 543) More recent evidence indicates that Baker's claim is untrue – there is a stage at which children *do* overgeneralize[7] – but this simply makes their behavior more difficult to explain. If children make certain predictable errors due to overgeneralization, they are probably not relying on the Subset Principle; so how are they ultimately solving this learnability problem? What higher-order constraints T must we assume in order to explain their behavior?

Bringing it all together

Each of these three learnability problems is centrally concerned with the issues that arise when faced with an underconstrained problem of induction. The Feature Relevance problem focuses on the question of how a learner decides which of a potentially infinite number of features to generalize upon. The No Negative Evidence problem is about how to correctly limit generalizations along the feature(s) already identified as

[7]See Pinker (1989) and Chapter 5 for an overview.

relevant, if one lacks negative evidence about which ones are incorrect. And the most general of the three, the Poverty of the Stimulus problem, is about how to generalize when one lacks the evidence (of whatever kind) believed necessary to distinguish between multiple hypotheses that are otherwise all consistent with the data.

Existing learnability analyses have proven valuable for not only pointing out these problems in the first place, but also for providing a starting point in the struggle to understand how human learners solve them; but, as with any single approach, they have certain limitations. In this thesis I exploit recent advances in machine learning and computer science to revisit these arguments from a novel perspective. Although each specific learnability problem poses its own challenges, and all differ in particular details, several common themes – the core insights contributed by the Bayesian approach – emerge throughout.

Ideal analysis, but with real-world datasets

The three learnability problems I have discussed involve datasets with certain characteristics. The mathematical investigations of learnability in the limit apply only for datasets that are infinite in size. For the logical problem of induction presented by Goodman, the concern arises from the finite size of the dataset rather than any of the other characteristics it might have. And the Poverty of the Stimulus argument makes reference not to an entire dataset of language, but only to those certain subsets assumed to be relevant.

In none of these cases are large, real-world datasets incorporated into the analyses. In one sense this is a strength – it is genuinely useful to be able to identify what can happen in the limit of infinite data – but at the same time, it forms only part of the picture. Learnability analyses can be ideal either because the learner is ideal, or because the data is. This means that the space of different kinds of learnability analyses encompasses four options, graphically depicted in Figure 1-1. Most learnability analyses to date, as we've seen, have applied only to the top square: evaluating the effects of having ideal data as well as an ideal learner. The ultimate goal of cognitive science and linguistics is to understand the bottom square: how realistic learners use

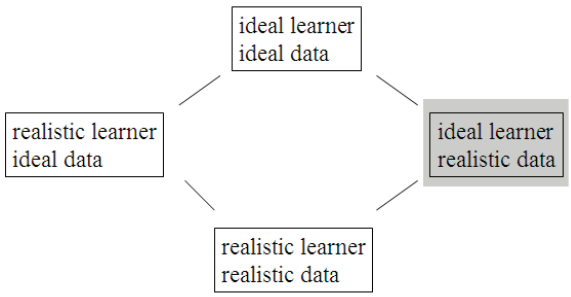


Figure 1-1: The landscape of learnability analyses. Learnability approaches differ based on the assumptions made about both the data and the learner. Most learnability analyses to date have applied only to the top square: ideal learners given ideal data. Ultimately we would like to be able to understand the bottom square: what realistic learners glean from realistic data. With this ultimate goal in mind, this thesis considers the shaded square on the right: ideal learners given realistic data.

realistic data over their lifetime. This thesis considers the shaded square on the right, exploring what is possible to learn in theory – i.e., given an ideal learner – based on the type of data that children actually see.

An advantage of this approach, besides moving the dialogue towards the ultimate goal, is that reframes the dialogue; instead of noting all of the many ways that learning is impossible in theory, it asks what *can* be learned in practice. Thanks to the analyses of mathematical learning theory, we know that it is impossible to learn non-finite languages in the limit without making certain strong assumptions about the presentation of the data, and that even viewing the problem probabilistically only helps if you know the general shape of the family of distributions to begin with. But these analyses apply to classes of languages defined according to the Chomsky hierarchy, not individual ones. Even if an ideal learner is not guaranteed to converge to the correct grammar on an arbitrary dataset, it might *actually* converge – or at least learn something linguistically interesting – on a dataset of real linguistic input. Natural language, particularly of the sort that children receive, is very different than an ideal corpus, containing a small number of sentence types that are idiosyncratically distributed in many ways. This is certainly a different sort of learning problem, and it is useful to explore precisely what assumptions need to be made to solve it.

Exploring full real-world datasets is important even in counterpoint to the anal-

yses that have either assumed that the dataset is minimal (but finite), or that have concentrated on only a particular subset of linguistic input. Goodman's riddle of induction points out that *some* constraints are necessary, leaving scientists with the question of "which ones?" Identifying those constraints – particularly if they exist on an even higher level, like over-overhypotheses – is much facilitated by being able to explore what an ideal learner *can* learn from real-world data. PoS arguments whose core contention concerns the lack of some piece of data are better evaluated in the context of a real-world dataset: this helps to avoid making implicit assumptions about which data is actually relevant to the learning problem.

Higher-order constraints can be learned

Although the idea that higher-order inductive constraints T can be learned is explored most fully in the context of the shape bias, it is a theme that runs through all of the analyses in this thesis. The importance of this idea is apparent in two major ways.

First, it forces a re-examination of the usual presumption in cognitive science that finding evidence for the existence of an inductive constraint means that the constraint *itself* must be innate. If inductive constraints can be learned – and sometimes learned faster than the specific hypotheses they are constraining – then it may be possible that many of the constraints and biases we assume to be built in are actually not. Under this view something must be innate, to be sure; but we may not be justified in concluding that any particular constraint is, even if it emerges early in development. More broadly, realizing how levels of knowledge can mutually reinforce and interact with one another as they are acquired can yield insight into what must be presumed in the first place.

Secondly, and perhaps more beneficially, this work demonstrates a paradigm for addressing the questions the first point opens up. Without a means to explore whether a hypothesized inductive constraint is truly innate, or is learned based on some even higher-level innate assumptions, the only benefit from this sort of re-examination would be, perhaps, to help restrain ourselves from jumping to conclusions. However, the Bayesian framework offers a means for performing that exploration by providing

a way to rigorously and systematically evaluate what sort of over-overhypotheses (or even over-over-overhypotheses) would be necessary to qualitatively explain human learning in the face of real-world data.

The simplicity/goodness-of-fit tradeoff

A central question in understanding how people learn and reason about the world relates to why we generalize beyond the input we receive at all – why not simply memorize everything? Needing to generalize creates a logical problem when there are an infinite number of possible features along which that generalization could be formed. Nevertheless, we must generalize because the ability to make inferences and predict data we haven't previously observed relies on our ability to extract structure from our observations of the world. If we do not generalize, we cannot learn, even though any act of generalization is, by definition, a simplification of the data in the world, resulting in occasional error. It is therefore important to simplify the data in such a way as to find the optimal balance between the gain in generalization and the cost due to error.

Achieving this balance is one of the fundamental goals of any learner, and indeed of any scientific theory or computational/statistical framework. Too much emphasis on simplicity means the learner is unable to learn from data, producing a high degree of error; too much emphasis on precisely memorizing the data means that the learner overfits, unable to capture the correct underlying generalizations. Bayesian models capture the tradeoff between simplicity and goodness-of-fit in an optimal way.[8] Because human learners – including children – care primarily about predicting future events, they too must adopt some version of this tradeoff.

As a result of performing this tradeoff, the amount and type of data can have a profound effect on the inferred theory. Especially when the representations involved are richly structured, what look like discrete qualitative shifts emerge simply because the tradeoff favors different theories as the data changes. In all three of the areas examined in this thesis, we will see how shifts in behavior that qualitatively parallel

[8]See Chapter 2 for a thorough discussion of this point.

human learning are a natural byproduct of Bayesian learning on realistic data.

Goals of this thesis

The central goal of this thesis is to explore three common learnability arguments by formalizing them in Bayesian terms; and, in so doing, to supply insight on several of the fundamental questions in cognitive science and language acquisition. What do we need to assume about a child's innate endowment – about her representational capacities, learning mechanisms, and cognitive biases – in order to explain the process and outcome of language acquisition? By focusing on three classic learnability arguments in the context of three major questions in language acquisition, I explore the issues of representation and learnability from multiple angles and demonstrate some of the power and flexibility of the this paradigm.

Bayesian learning offers a solution to the No Negative Evidence problem, resulting from the simplicity/goodness-of-fit balance. And while it does not solve the Feature Relevance or the Poverty of Stimulus problems in the abstract, it yields useful insights about both: that higher-level abstract constraints may be learned at least as rapidly as lower-level specific information; that evidence for certain types of abstract knowledge may result from characteristics of an entire dataset rather than specific items; and that what knowledge or constraints *are* innate may be specifiable at an abstract enough level that they could plausibly be domain-general even if the less abstract knowledge is not.

This research offers a new perspective about what we can and can't conclude from standard learnability arguments, and provides a framework that allows for an increasingly subtle, detailed, and rigorous analysis of claims about what must or must not be innate (as well as what that even means). The framework allows us to evaluate the bounds of "optimal" reasoning while also being able to vary with precision how this might depend on the data, the nature of the learner's mental representation, and both domain-specific as well as domain-general biases (whether due to the attentional, perceptual, memory-based, or learning mechanisms of the child). By no means do

I argue that such a framework should supplant other methods in cognitive science, but it is an essential tool in the toolbox as we move toward constructing a full and accurate picture of the human mind.

In the next chapter I provide a basic introduction to the Bayesian framework, including a discussion of some of the implications and issues that arise for anybody working within it. The subsequent three chapters each focus on the three particular areas in language acquisition and cognitive development introduced in here, which we can broadly classify by topic: abstract syntactic structure, verb argument structure, and word learning. Although similar themes emerge in each of these areas, the division by topic helps us explore the power and flexibility of the framework as I apply it to a range of problems, over several types of representation, and with multiple kinds of data. In the final chapter I discuss this work on a broader level, with a special focus on the themes underlying all of the individual projects, a consideration of the framework, and the direction of research as we move forward.

Chapter 2

An introduction to Bayesian modelling

Generalization is a central issue in cognitive science and language acquisition: how do we learn so much from such apparently limited evidence? This question arises in any situation in which the conclusions are underdetermined by the information available and it is therefore necessary to make inductive leaps – precisely the sorts of situations likely to be encountered by a human learner on a daily basis. Bayesian inference is a general-purpose computational framework for understanding and exploring how people might make these inductive leaps.

This chapter provides a basic introduction to how the Bayesian framework can be applied to important questions in cognitive science. The goal is to provide an intuitive and accessible guide to the *what* and the *why* of the Bayesian approach: what sorts of problems and data the framework is most relevant for, and how and why it may be useful for psychologists and cognitive scientists. I spend some time considering issues that are particularly relevant for any cognitive scientist in evaluating a computational model, including its biological plausibility, application to developmental issues, and the question of what these sorts of models can reveal about the human mind. This will be useful in subsequent chapters as we analyze particular models and ask questions about why they act as they do, and what conclusions we can draw about important issues in cognitive science on the basis of their behavior.

The basics

The fundamental question a Bayesian approach addresses is how to update beliefs and make new inferences in light of new data. Central to this framework is the assumption that beliefs can be represented as probabilities – that one's degree of belief in some proposition, hypothesis, or theory X can be expressed as a real number ranging from 0 to 1, where 0 means "X is completely false" and 1 means "X is completely true." Making such an assumption allows us to use the mathematics of probability theory to conduct inference: how to select from among a set of theories which one(s) best explain the observed data. The Bayesian framework does so by formalizing an important tradeoff between the complexity of the theory on one hand and how well it explains the observed data on the other; it prefers theories that offer a balance between the two.

Observed data is assumed to be generated by some underlying process – a mechanism explaining why the data occurs in the patterns it does. (Spoken sentences may be generated from some sort of mental grammar; words are generated from a mental lexicon; and both are partially affected by social and pragmatic factors as well). The job of the learner is to evaluate different hypotheses about the underlying nature of that process, and to make predictions based on the most likely hypotheses.

This is perhaps best illustrated in the context of a schematic example, depicted graphically in Figure 2-1. Data (the dots) are generated by processes that occupy different subsets of space: each process can generate data within its region, but not data outside of it. Each hypothesis constitutes a different theory about which subset(s) of space the processes occupy. One such hypothesis is illustrated in Figure 1: according to this hypothesis, there are two underlying processes, each denoted by one of the blue rectangles. If a learner believes that this hypothesis was correct, she might predict that she would be far more likely to observe a new datapoint in position a rather than position b, even if she had previously seen neither. In such a way, learning which hypotheses are most likely can aid a learner in inference and generalization.

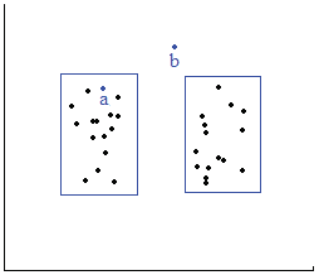


Figure 2-1: Example data and hypothesis. Graphical representation of data and one possible hypothesis about the how that data was generated. This hypothesis suggests that there are two generative processes, each depicted as a blue rectangle. Under this hypothesis, a new datapoint in position a is far more likely than one in b.

Hypotheses are compared using Bayes' Rule (Equation 2.1, below), which states that the probability of some hypothesis given the data (the posterior) is proportional to the probability of the data given the hypothesis (the likelihood) and the probability that the hypothesis is true, regardless of what the data is (the prior).

$$p(H|D) \propto p(D|H)p(H). \tag{2.1}$$

Intuitively, the posterior probability captures a natural balance between simplicity (measured by the prior) and goodness-of-fit (measured by the likelihood). Achieving this balance is one of the fundamental goals of any computational framework, and indeed of any scientific theory: too much emphasis on simplicity means the theory or model is unable to learn from data, and too much emphasis on goodness-of-fit means that it overfits, unable to capture the true generalizations within the data. Figure 2-2 illustrates this graphically: channeling Goldilocks, we might conclude that hypothesis A looks too simple, hypothesis C seems too complex, and hypothesis B is "just right." (Or, following Einstein, "everything should be made as simple as possible, but not simpler.") Bayesian models capture this intuition using techniques sometimes known as the Bayesian Occam's Razor (MacKay, 2004).

How well a hypothesis fits the data is captured by the likelihood, given by $P(H|D)$: the probability of the observed data given the hypothesis. Although the likelihood can sometimes be difficult to calculate in practice, it is straightforward to understand

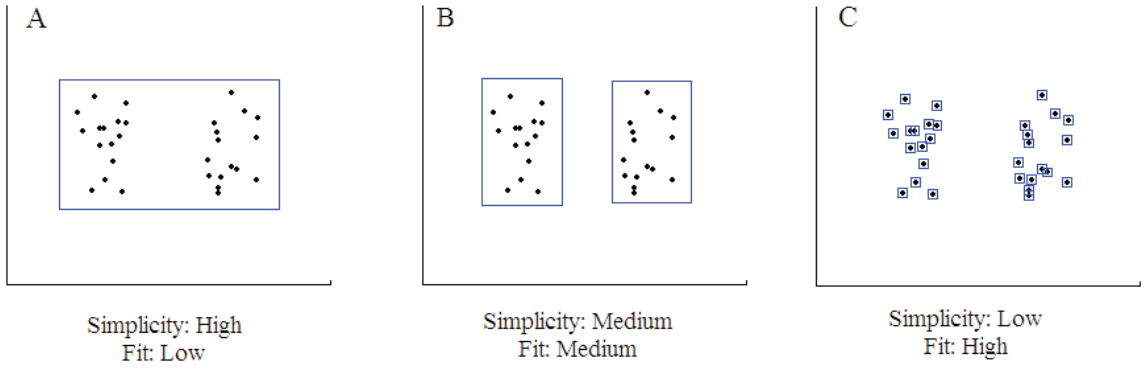


Figure 2-2: The Goldilocks Effect. Hypothesis A is too simple, C is too complex, and B is "just right." Hypothesis A is quite simple, but fits the observed data poorly: C fits closely but is highly complicated. The best description of the data should optimize a tradeoff between complexity and fit, as in B.

intuitively. For instance, hypothesis C in Figure 2-2 clearly has a high 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. Hypothesis C therefore fits the data extremely well. By contrast, hypothesis A has relatively low likelihood: it does not explain why the datapoints are found where they are. After all, according to A, the thirty datapoints would be just as likely if they were each randomly located in other places within the blue rectangle. The ratio of the observed datapoints to the area for predicted data is low for A, since the data could easily have been elsewhere, but high for C, since it couldn't. Likelihood is, essentially, this ratio[1]; thus, hypotheses that make specific predictions – those with more explanatory power – are favored there.

The prior probability of a hypothesis, or $P(H)$, is calculated in such a way that simpler hypotheses have higher probability. The definition of simplicity and the corresponding calculation of $P(H)$ are not the result of some externally-imposed *ad hoc* mechanism. Rather, they emerge naturally from the generative assumptions

[1]This is, incidentally, why it can be sometimes be difficult to calculate. In the example here, area is a straightforward calculation, but if the space of hypotheses is itself quite complicated, then calculating the "area" of that space involves summing up all of the data that can be generated by the hypothesis. This can be done by enumeration, but this is very time-consuming, and only practical for simple problems. In general, mathematical techniques are used instead; one of the biggest areas of research in machine learning is to develop more effective techniques and faster, more efficient algorithms for using them. Gilks, Richardson, and Spiegelhalter (1996), MacKay (2004), and A. Gelman, Carlin, Stern, and Rubin (2004) are good references for those interested in this issue.

underlying the Bayesian framework, in which hypotheses are themselves generated from a space of candidate hypotheses. For instance, the hypotheses in Figure 2-2 correspond to different sets of rectangular regions within the metric space. Simpler hypotheses require fewer "choice points" during that 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 something is, the more likely it is to mess it up – or, in other words, "The more they overtech the plumbing, the easier it is to stop up the drain."[2] I discovered this principle in junior high school when I tried to make a double-layer cake for a bake sale, never before having tried baking anything. Several explosions of flour, an unanticipated smoke alarm, and a burnt, unrisen cake later, I concluded that perhaps it would have been a better idea to start my cooking career with grilled cheese sandwiches or scrambled eggs instead. More complicated recipes have more ingredients and more directions, which means they have more choice points – more opportunities for things to go wrong. In a similar way, the more choices a hypothesis requires, the more likely it is that those choices could have been made in a different way, resulting in an entirely different hypothesis.

The precise prior probability of a hypotheses is therefore not arbitrarily assigned, but rather falls out in a principled way from how the hypotheses are generated. My "generative model" for cake recipes – in a sad reflection of my cooking knowledge at the time – was close to assuming that recipes consist of ingredients randomly thrown together; the generative model for the hypotheses in Figure 2-2 is one that can result in any possible combination of rectangular regions within the metric space. A different generative model results in a different – but no less principled – assignment of prior probabilities. For instance, if we assumed the regions could be circles rather than rectangles, then each region would require three choice points rather than four (the

[2]Scotty, *Star Trek IV*.

x and y coordinates of the center of the circle, plus its radius). Despite the different generative model, the logic favoring simple hypotheses is the same: multiple regions will still be *a priori* less likely than a few.[3] The generative model therefore matters for determining precisely what the relative probability of each hypothesis is, but most reasonable models give qualitatively similar relative probabilities to qualitatively similar hypotheses.

The set of all possible hypotheses is called the *hypothesis space*, which is depicted graphically in Figure 2-3. It can also be thought of as the range of hypotheses the model can entertain – the space of possibilities inherent in the setup. In the cooking example, the hypothesis space would be the space of all recipes I could have made given the ingredients in the house at the time – the number of random combinations of those ingredients and ways of cooking them. (Note that hypothesis spaces can be infinite in size). The hypothesis space corresponding to Figure 2-2 consists of all possible combinations of rectangles. Hypothesis spaces can also be additive or disjunctive: for instance, if we weren't sure about the shape of the regions generating the data, we could consider the hypothesis space made up of all possible rectangles *and* all possible circles. (If that were the case, prior probability would be assigned under a generative model that had an additional choice point for each region: is it a rectangle or a circle? Each individual hypothesis would therefore have a lower prior probability, although the probability of each relative to the other would be the same).

Learning and inference within a Bayesian model involves comparing the different

[3]It is always true that *within the set of choices defined by the generative model*, the hypothesis that requires fewer choices has higher probability. However, it is always possible, by changing the primitives of the model, to design some Bizarro World model that looks to us as though it assigns higher probability to more complicated-looking hypotheses. For instance, so far we have considered "primitives" – the basic elements we have to make choices about – to be simple regions like rectangles and circles. We could imagine instead that each primitive was defined within the model as, say, a set of thirty tiny regions (of which hypothesis C might be one instance); other hypotheses might be made by adding to or subtracting from those. In that case, hypothesis C might very well have higher prior probability than A. In this sense, then, the simplicity of a hypothesis is only meaningful relative to the set of primitives out of which the hypotheses are generated – and prior probability is, therefore, only a principled calculation given to those. The decision of what the primitives are is therefore an important choice that must be made by any modeler, and is an issue I consider in more depth later. For now, I will just note that this is a decision that is not restricted to Bayesian modelers, but must in fact be faced by any theorist, computational or otherwise: indeed, the Bayesian framework – by forcing us to make these assumptions explicit – can be a powerful tool for evaluating the primitives we choose.

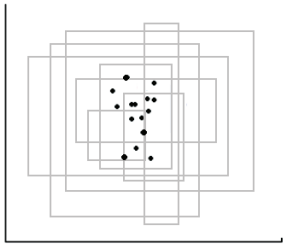


Figure 2-3: Hypothesis space for some sample data. Each rectangle consists of a possible hypothesis; some are supported closely by the data and some are not. The actual hypothesis space contains all possible rectangles, but I show just a few for illustrative purposes.

hypotheses in the hypothesis space, ultimately preferring those with the highest posterior probability. In the example above, the model learns which of the possible sets of regions constitutes the best theory about the underlying data-generating process. The hypothesis space itself is given: the theory must consist of a set of regions that exist within the positive quadrant of coordinate space, and the regions are specified by x, y, l, and w values within some range. This type of information must be built into any model, but these specifications need not be very strong or very limiting: for instance, one might simply specify that the range of possible values for x, y, l, and w lies between 0 and some extremely large number like 10^9, or be drawn from a probability distribution with a very long tail. The issue of what is "built in" will recur throughout this chapter, and I will return to it more broadly later in the chapter.

Hierarchical learning

One way to effectively increase the size of the hypothesis space is to allow learning at multiple levels of abstraction – learning higher-order information about specific hypotheses as well as the hypotheses themselves. We see this type of learning all the time in human cognition. Learning to cook involves realizing that making breads and pastries generally requires mixing ingredients that often include eggs and flour; that meat is often either fried or cooked in a pan, but not usually inserted into the toaster; and that spices and seasonings should be used by the pinch, not by the jar. These

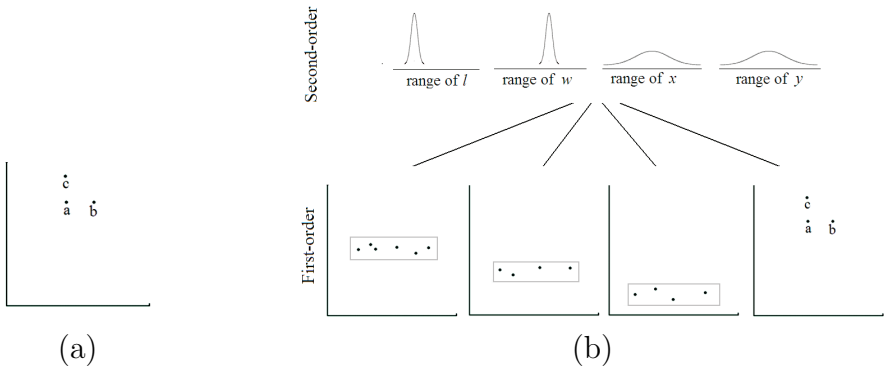


Figure 2-4: Learning higher-order information. (a) Given point a, one cannot identify whether b or c is more likely. (b) Given additional data, a model that could learn higher-order information about hypotheses might realize that regions tend to be long, thin rectangles oriented along the y axis. If so, points a and b are probably within the same region, but a and c are not.

realizations are all inferences about the nature of recipes rather than specific recipes themselves, and are especially useful when confronted with a novel yet underspecified recipe or a new set of ingredients. Even if you have never made roast goose before, you know that you shouldn't try to cook it by grinding it up and mixing it with eggs and flour – there is no need to ruin Christmas dinner to find this out.

A similar type of inference would occur if, presented with the data in Figure 2-4, a learner realized not only which rectangular regions were most likely, but also that the regions tend to long, thin, and oriented along the y-axis. Just as making higher-order inferences about cooking helps when confronted with a new situation, making this type of higher-order inference helps the learner when confronted with novel data. Such a learner would be able to infer that point b is probably in the same region as a but point c is not, even though b and c are equidistant from a.

A certain kind of Bayesian model, known as a hierarchical Bayesian model (HBM), can capture this sort of inference. It learns not only by choosing among specific hypotheses, but by also making higher-order generalizations *about* those hypotheses. In a non-hierarchical Bayesian model, the modeler sets the parameters that govern the hypotheses, as we've seen. In a hierarchical model, the modeler specifies the hyper-parameters – parameters on the parameters – and the model learns the parameters

themselves. So rather being given that the range of each of the x, y, l, and w values lies between 0 and 10^9, the hierarchical model learns the typical range of each – e.g., that l tends to be short while w tends to be long – and the modeler specifies the range of the ranges.

A surprising effect of learning in hierarchical models is that, quite often, the higher-order abstractions are acquired before the specific lower-level details: the model might realize that l tends to be short and w tends to be long before it has specified the size and location of each rectangular region with precision. This effect, which we might call the "blessing of abstraction"[4], is somewhat counterintuitive. Why are higher-order generalizations like this sometimes easier for a Bayesian learner to acquire?

One reason is that the higher-level hypothesis space is often smaller than the lower-level one. As a result, the model has to choose between fewer options at the higher level, which may require less evidence. For instance, the higher-level knowledge may consist of only three options: l and w are approximately equal, l is smaller than w, or w is smaller than l. Even if a learner doesn't know whether l is 10 units or 11 units long and w is 20 or 22, it might be fairly obvious that l is smaller than w.

More generally, the higher-level inference concerns the lower-level hypothesis space (and hence the dataset as a whole), whereas the lower-level inference is only relevant for specific datapoints. The red datapoint in Figure 2-4 is informative only about the precise size and location of a single region, the rectangle on the left. However, it – and every other single datapoint – is informative about all of the higher-level hypotheses. There is, in effect, more evidence available to the higher levels than the lower ones, and they are therefore learned more quickly.

This has interesting implications for the study of learnability and the question of innateness. The basic motivation for positing innate constraints on cognitive development is that without these constraints, children would be unable to infer the specific knowledge that they seem to acquire from the limited data available to them. What is critical to the argument is that some constraints are present prior to learning some of the specific data, not that those constraints must be innate. Approaches to

[4]I owe this coinage to Noah Goodman.

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, hierarchical Bayesian models provide such a learning mechanism. If an abstract generalization can be acquired very early and can function as a constraint on later acquisition of specific data, 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.

Representational capacity

Because a Bayesian model can be defined for any well-specified generative framework, the representational capacity of the Bayesian approach is limited only by the representations that can be specified by the generative process. In the example we've been considering, the representation (rectangles in a metric space) is fairly simple, and probabilistic inference is done over that representation. Other common representations include probability distributions in a metric space, which may be appropriate for phonemes as Gaussian clusters in phonetic space (e.g., N. Feldman & Griffiths, 2007); directed graphical causal models, which may be appropriate for causal reasoning (e.g., Pearl, 2000; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Gopnik et al., 2004); abstract graphical structures including taxonomies, which may be appropriate for some aspects of conceptual structure (e.g., Kemp, Perfors, & Tenenbaum, 2004; Roy, Kemp, Mansinghka, & Tenenbaum, 2006; Schmidt, Kemp, & Tenenbaum, 2006; Xu & Tenenbaum, 2007); objects as vectors of features, which may be appropriate for categorization and object understanding (e.g., Kemp, Tenenbaum, Griffiths, Yamada, & Ueda, 2006; Kemp, Perfors, & Tenenbaum, 2007; Shafto, Kemp, Mans-

inghka, Gordon, & Tenenbaum, 2006; Navarro & Griffiths, 2007); bags of words, which may be appropriate for semantic representation (e.g., Griffiths, Steyvers, & Tenenbaum, 2007); grammars, which may be appropriate for syntax (e.g., Dowman, 2000; Perfors, Tenenbaum, & Regier, 2006); argument structure frames, which may be appropriate for verb knowledge (e.g., Alishahi & Stevenson, 2008); Markov models, which may be appropriate for action planning and goal inference or part-of-speech tagging (e.g., Verma & Rao, 2006; Baker, Tenenbaum, & Saxe, 2007; Goldwater & Griffiths, 2007); and even logical rules, which may be appropriate for some aspects of conceptual knowledge (e.g., Goodman, Griffiths, Feldman, & Tenenbaum, 2007). This is not an exhaustive list, but it is sufficient to illustrate the representational flexibility inherent in the Bayesian approach.

The ability to capture a wide variety of both structured and unstructured representations provides an escape from an intellectual dichotomy imposed in part by the computational models available to theorists, especially those interested in questions of learnability in language. Although the question of whether human learners have (innate) domain-specific knowledge is logically separable from the question of whether and to what extent that knowledge requires structured representations, in practice these issues have often been conflated. Within cognitive science, computational models of how language might be learned have usually assumed that domain-general learning operates on representations without explicit structure (e.g., Elman et al., 1996; Rumelhart & McClelland, 1986; Rogers et al., 2004; Reali & Christiansen, 2005). The main proponents of innate language-specific factors, on the other hand, have typically assumed that the representations involved are structured (e.g., Chomsky, 1965; Pinker, 1984). Few cognitive scientists have seriously explored the possibility that explicitly structured mental representations might be learned via domain-general mechanisms. This lack exists in part because, historically, the computational formalisms most widely perceived to be capable of capturing domain-general human learning – neural networks and Markov models – incorporate representations without explicit structure. While both of these frameworks have made substantial contributions to cognitive science, it is also important to be able to evaluate the possibility

that structured representations exist explicitly but can still be learned. Bayesian models provide a computational framework in which to ask this question.

Bayesian learning in action

Although each of these specific models have distinct properties depending on the nature of the domain and dataset they are appropriate for, Bayesian learning in general has a number of notable characteristics. One such feature, which emerges naturally as a result of the tradeoff between simplicity and goodness-of-fit, is that the size of the dataset matters: different hypotheses may be favored for different amounts of data, even if the data is always generated by the same underlying process. This is a byproduct of sensible inference, and a hallmark of human cognition – I am more apt to think that all pastry recipes share a common structure (eggs, flour, baking in an oven) if I have seen many different recipes rather than only one. The Bayesian framework offers an explanation of why this kind of inference is reasonable, and provides a precise account for how inference changes as the amount of data increases – in short, how this sort of learning works, and why.

Let us return to Figure 2-2, but instead of comparing three different hypotheses on the same dataset, we can ask which hypotheses are preferred on datasets of different size. Figure 2-5 shows three datasets, all generated from the same underlying process, but varying in the number of datapoints. The best hypothesis fits the five datapoints in dataset 1 quite poorly, but because there are so few points this does not impose a substantial penalty relative to the high prior probability of the hypothesis. Thus, smaller datasets are often marked by a greater degree of overgeneralization because the dynamics of inference favor the simpler, more overgeneral hypotheses. As the data accumulates, the penalty imposed for poor fit is greater, since it applies to each datapoint that is not predicted accurately by the hypothesis. Thus, the same hypothesis that was dispreferred on the data in Figure 2-2 is actually most appropriate for the largest dataset, which contains many points clustered into thirty tiny regions.

A potentially misleading implication of Figure 2-5 is that as the number of dat-

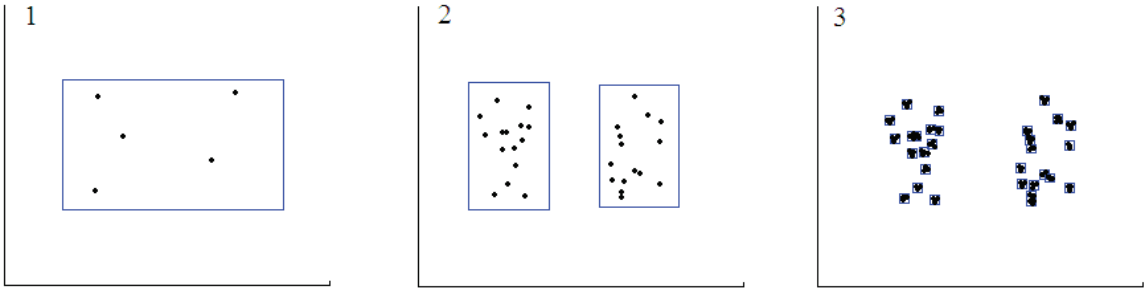


Figure 2-5: Role of dataset size. Three datasets with increasing numbers of datapoints and their corresponding best hypotheses. For dataset 1, there are so few datapoints that the simplicity of the hypothesis is the primary consideration; by dataset 3, the preferred hypothesis is one that fits the clustered datapoints quite tightly.

apoints increases, the most complex hypothesis will eventually always be preferred. This is not true. Rather, the hypothesis that will be preferred is the one that compresses the data most optimally: in general, this will be the hypothesis that is closest to the true generative process[5] (MacKay, 2004). In other words, if the data is truly generated by a process corresponding to thirty different rectangular regions, then the points will increasingly clump into clusters in those regions, and hypothesis C will be preferred. But if the data is truly generated by a process inhabiting two larger regions, then additional datapoints would look more like Figure 2-6 instead; in which case, hypothesis B would still have a higher probability. In general, as data accumulates a Bayesian learner will move towards increasingly complex hypotheses, but stops as the current hypothesis approximates the true generative process.

This initial preference for a simple hypothesis, followed by the eventual adoption of a more complex hypothesis, provides a natural solution to the No Negative Evidence

[5]Technically, this result has been proven for information-theoretic models in which probabilities of data or hypotheses are replaced by the lengths (in bits) of messages that communicate them to a receiver. The result is known as the "MDL Principle" (Rissanen, 1978), and is related to Kolmogorov complexity (Solomonoff, 1964; Kolmogorov, 1965). The Bayesian version applies given certain assumptions about the randomness of the data relative to the hypotheses and the hypotheses relative to the prior (Vitànyi & Li, 2000). Both versions apply only to the hypotheses in the hypothesis space: if no hypothesis corresponding to the true data generating process exists in the space, then it will never be considered, much less ultimately preferred. Thus, the hypothesis that is preferred by the model in the limit of infinite data is the "best" hypothesis only in the sense that it is closest to the true data generating process out of all of the hypotheses considered. This is a somewhat simplified account of the mathematical results, since the relationship between Bayesian inference and MDL-based approaches is complex; see Vitànyi and Li (2000); Jaynes (2003); MacKay (2004); Grünwald, Myung, and Pitt (2005) for more details).

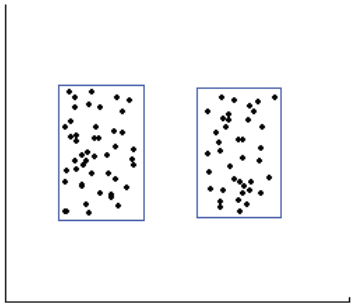


Figure 2-6: Interaction of dataset size and hypothesis choice. As data accumulates, the best hypothesis approaches the true generative model. Here, the hypothesis consisting of two subsets of space is preferred over one with more, since the data is distributed in such a way as to make the latter unlikely.

problem: deciding among hypotheses given positive-only examples. As the size of the dataset approaches infinity, a Bayesian learner rejects overgeneral hypotheses in favor of more precise ones, as the subset principle would do. 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 – one that explains both how and when overgeneralization should occur.

What will a Bayesian learner conclude after seeing just *one* datapoint? It may be tempting to infer from the discussion so far that in that case, the simplest possible hypothesis will be preferred; but this is not necessarily the case. It all depends on the tradeoff in complexity that it takes to represent individual datapoints as opposed to entire hypotheses: if it is "cheaper" to represent one or two datapoints than an entire hypothesis – no matter how simple – then simply memorizing the data would have a higher likelihood *and* a higher prior probability than an extremely vague, overgeneral hypothesis. This is sensible in real-world terms: if I have only seen one recipe in my entire life, it makes more sense to simply memorize it than to try to extract general "recipe principles." Consider also our graphical example. One datapoint would require two choices to specify – its x and y coordinates. By contrast, even the simplest hypothesis would require at least four (x, y, l, and w). Moreover, the single datapoint also has the highest possible likelihood, since it predicts the data

(itself) exactly. Only as the number of datapoints increases does the penalty in the prior become high enough to preclude simply memorizing each datapoint individually: this is when overgeneral, highly simple hypotheses begin to be preferred. Precisely how much data can be memorized depends on the representational complexity of the hypotheses and data, but the basic qualitative pattern occurs widely.

This pattern looks a great deal like the U-shaped curve observed in a variety of domains over the course of development (e.g., Marcus et al., 1992; Siegler, 2004). It falls naturally out of the dynamics of learning, as the balance between simplicity and goodness-of-fit changes in response to increases in the amount of data.

Some general issues

Several issues are typically raised when evaluating Bayesian modelling as a serious computational tool for cognitive science. Bayesian reasoning characterizes "optimal" inference: what does this mean? How biologically plausible are these models, and how much does this matter? And finally, where does it all come from – the hypothesis space, the parameters, the representations? The answers to each of these questions affect what conclusions about actual human cognition we can draw on the basis of Bayesian models; I will therefore consider each one in turn.

Optimality: What does it mean?

Bayesian probability theory[6] is not simply a set of *ad hoc* rules useful for manipulating and evaluating statistical information: it is also the set of unique, consistent rules for

[6]Bayesian methods are often contrasted to so-called "frequentist" approaches, which are the basis for many of the standard statistical tests used in the social sciences, such as t-tests. Although frequentist methods are often only appropriate for certain simple, idealized conditions that often fail to describe reality, provide no way to take prior information account, and indeed constitute a special case of Bayesian probability theory, Bayesian methods have historically been relatively neglected, and often attacked, in part because they are viewed as unnecessarily subjective (R. Fisher, 1933; Jeffreys, 1939; Savage, 1954; Lindley, 1956; Dantzig, 1957). This perception is untrue – Bayesian methods are simply more explicit about the prior information they take into account. Regardless, the issue of subjectivity seems particularly irrelevant for those interested in modelling human cognition, where accurately capturing "subjective belief" is part of the point. (There has also been a great deal of debate about the proper interpretation of probabilities within each framework; see the references above, as well as Jaynes (2003), for an overview of this issue).

conducting plausible inference Jaynes (2003). In essence, it is a extension of deductive logic to the case where propositions have degrees of truth or falsity – that is, it is identical to deductive logic if we know all the propositions 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 might include things like consistency and qualitative correspondence with common sense – if you see some data supporting a new proposition A, you should conclude that A is more plausible rather than less; the more you think A is true, the less you should think it is false; if a conclusion can be reasoned 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). Put another way, Bayesian probability theory is "optimal inference" in the sense that a non-Bayesian reasoner attempting to predict the future will always be out-predicted by a Bayesian reasoner in the long run (de Finetti, 1937).

Even if the Bayesian framework captures optimal inductive inference, does that mean it is an appropriate tool for modelling human cognition? People's everyday reasoning can be said to be many things, but few would aver that it is optimal, subject as it is to emotions, heuristics, and biases of many different sorts (e.g., Tversky & Kahneman, 1974). However, even if humans are non-optimal thinkers in many ways – and there is no reason to think they are in *every* way – it is impossible to know this without being able to precisely specify what optimal thinking would amount to. Understanding how humans *do* think is often made easier if one can identify the ways in which people depart from the ideal: this is approximately the methodology by which Kahneman and Tversky derived many of their famous heuristics and biases, and the flexibility of the Bayesian approach makes it relatively easy to incorporate

constraints based on memory, attention, or perception directly into one's model.

Bayesian modelling, in fact, is an implementation of scientific inquiry that operates on Marr's third (computational) level, which seeks to understand cognition based on what its goal is, why that goal would be appropriate, and the constraints on achieving that goal, rather than precisely how it is implemented algorithmically (Marr, 1982). Understanding at this level is important because the nature of the reasoning may often depend more on the learner's goals and constraints than it does on the particular implementation. It can also enhance understanding at the other levels: for instance, analyzing connectionist networks as an implementation of a computational-level theory can elucidate what sort of computations they perform, and often explain why they produce the results they do (Hertz, Krogh, & Palmer, 1991; MacKay, 2004).

Being able to precisely specify and understand optimal reasoning is also useful for performing ideal learnability analysis, which especially important in the area of cognitive development. What must be "built into" the newborn mind in order to explain how infants eventually grow to be adult reasoners, with adult knowledge? One way to address this question is to establish the bounds of the possible: if some knowledge couldn't possibly be learned by an optimal learner presented with the type of data children receive, it is safe to conclude that actual children couldn't learn it, either[7]. It would then be necessary to identify precisely what must be innate in order for that knowledge to be learnable in principle. Due to its representational flexibility and its ability to accurately calculate optimal inference even on large datasets, Bayesian modelling is an ideal tool for this sort of problem.

Biological plausibility

Because cognitive scientists are ultimately interested in understanding human cognition, and human cognition is ultimately implemented in the brain, it is important that our computational-level explanations be realizable on the neurological level, at least potentially. This is one reason for the popularity of the connectionist approach, which was developed as a neurally inspired model of the cognitive process (Rumelhart

[7]Assuming, of course, that the data in question accurately reflects children's natural input.

& McClelland, 1986). Connectionist networks, like the brain, contain many highly interconnected, active processing units (like neurons) that communicate with each other by sending activation or inhibition through their connections. As in the brain, learning appears to involve modifying connections, and knowledge is represented in a distributed fashion over the connections. As a result, representations degrade gracefully with neural damage, and reasoning is probabilistic and "fuzzy."

In contrast, Bayesian models may appear wholly implausible from the neurological perspective. One of the major virtues of the Bayesian approach – the transparency of its computations and the explicitness of its representation – is, in this light, potentially a major flaw: the brain is many wonderful things, but it is neither transparent nor explicit. How could representations like grammars or logics be instantiated in our neural hardware? How could our cortex encode hypotheses and compare them based on a tradeoff between their simplicity and goodness-of-fit? Perhaps most problematically, how could the brain – with a processing speed that is orders of magnitude slower than that of modern computers – actually implement optimal inference, which requires such a thorough search of such enormous hypothesis spaces that even the fastest computers can take days or weeks, if they can succeed at all?

These are good questions, but there is growing evidence for the relevance of Bayesian approaches on the neural level (e.g., Doya, Ishii, Pouget, & Rao, 2007). Probability distributions can in fact be represented by neurons, and they can be combined according to a close approximation of Bayes' Rule; posterior probability distributions may be encoded in populations of neurons in such a way that Bayesian inference is achieved simply by summing up firing rates (Pouget, Dayan, & Zemel, 2003; Ma, Beck, Latham, & Pouget, 2006). Spiking neurons can be modeled as Bayesian integrators accumulating evidence over time (Deneve, 2004; Zemel, Huys, Natarajan, & Dayan, 2005). Recurrent neural circuits are capable of performing both hierarchical and sequential Bayesian inference (Deneve, 2004; Rao, 2004, 2007). Even specific brain areas have been studied: for instance, there is evidence that the recurrent loops in the visual cortex integrate top-down priors and bottom-up data in such a way as to implement hierarchical Bayesian inference (Lee & Mumford, 2003).

This work, though still in its infancy, suggests that concerns about biological plausibility may not, in the end, prove to be particularly problematic. It may seem to us, used to working with serial computers, that searching these enormous hypothesis spaces quickly enough to perform anything approximating Bayesian inference is impossible; but the brain is a parallel computing machine made up of billions of highly interconnected neurons. The sorts of calculations that take a long time on a serial computer, like a sequential search of a hypothesis space, might be very easily performed in parallel. Whatever the future holds, indications so far serve as a chastening reminder of the danger of advancing from the "argument from incredulity" to any conclusions about biological plausibility.

It is also important to note that, for all of their apparent biological plausibility, neural networks are unrealistic in important ways, as many connectionist modelers acknowledge. Units in neural networks are assumed to have both excitatory and inhibitory connections, which is not neurally plausible. This is a problem because the primary learning mechanism, backpropagation, relies on the existence of such connections (Rumelhart & McClelland, 1986; Hertz et al., 1991). There is also no analogue of neurotransmitters and other types of chemical transmission, which play an important role in brain processes (Gazzaniga, Ivry, & Mangun, 2002). These issues are being overcome as the state of the art advances (see Rao, Olshausen, and Lewicki (2002) for some examples), but for the models most commonly used in cognitive science – perceptrons, multilayered recurrent networks and Boltzmann machines – they remain a concern.

The different computational techniques are therefore each biologically plausible in some ways and perhaps less so in others. Because we still know so little about the neurological mechanisms within the brain, it is difficult to characterize how plausible either approach is or how much the ways they fall short impact their utility. In addition, to a large extent, biological plausibility is irrelevant on the computational level of analysis. Even if it turned out that there was no possible way the brain could implement anything (even heuristically) approximating Bayesian inference – which seems unlikely in light of current research – Bayesian models would still be

useful for comprehending the goals and constraints faced by the cognitive system, and for comparing actual human performance to optimal reasoning. To the extent that connectionist models apply on the computational level, the same is true for them.

Where does it all come from?

For many, a more important critique is that it seems that, in some sense, Bayesian models do not seem to be *learning* at all. The entire hypothesis space, as well as the evaluation mechanism for comparing hypotheses, has been given by the modeler; all the model does is choose among hypotheses that already exist. Isn't learning, particularly the sort of learning that children perform over the first years of their life, something more than this? Our intuitive notion of learning certainly encompasses a spirit of discovery that does not appear at first glance to be captured by a model that simply does hypothesis testing within an already fully-specified hypothesis space.

The same intuition lies at the core of Fodor's famous puzzle of concept acquisition (Fodor, 1975, 1981). His essential point is that one cannot learn anything via hypothesis testing because one must possess it in order to test it in the first place. Therefore, except for those concepts that can be created by composing them from others, all concepts – including concepts like CARBURETOR and GRANDMOTHER – are, and must be, innate.

Fodor's argument makes sense only under a very restricted, essentially un-intuitive definition of what it means to learn. As an analogy, consider a standard English typewriter with an infinite amount of paper. There is a space of documents that it is capable of producing, which includes things like *The Tempest* and does not include, say, a Vermeer painting or a poem written in Russian. This typewriter can easily be formalized in Bayesian terms where each document is a hypothesis and the infinite set of documents producible by the typewriter is its hypothesis space. (The hypothesis space does not consist of the documents that the typewriter has actually produced because a hypothesis space is not an exhaustive, explicitly enumerated set of hypotheses[8]; rather, it is the latent space of possibilities that is implicitly defined

[8]Indeed, exhaustive hypothesis enumeration is intractable for all but the simplest models; most

by the generative process).

Is there a difference between documents that have been created by the typewriter and documents that exist only in the latent space? Of course there is: documents that have been created can be manipulated in all sorts of ways (reading, burning, discussing, editing) that documents latent in the space cannot. In the same way, there may be a profound difference between hypotheses that have been considered by the learner and hypotheses that are simply latent in the space: the former can be manipulated by the cognitive system – evaluated, used in inference, compared to other hypotheses – but the latter cannot. The process of transformation from an implicit, latent hypothesis into one that can be manipulated by the cognitive system is a process that looks, from most perspectives, rather like learning. Only once something is learned can we entertain it, talk about it, and manipulate it in other ways; to say that it exists in the latent space simply means we have the capacity to learn it. When I learn how to play chess, that means that I now identify chess pieces, entertain strategies for winning, and talk about the rules with other players, whereas before I didn't (though I had the capacity to learn it).

Fodor's argument, and this critique of Bayesian modelling, presumes a notion of learning that is very different from this, and indeed that is conceptually incoherent in many ways. If "learning" does not include acquiring some knowledge through hypothesis testing, then what *would* it consist of? The knowledge would have to spring into the hypothesis space somehow, but how? As we think about it further, we realize that this critique assumes a definition of learning that only applies if new knowledge enters the hypothesis space without any regularity or predictability whatsoever.

This can be proven by contradiction. Imagine that I could explain how new knowledge might be added to a hypothesis space; such an explanation would have to make reference to some rules or some kind of process for adding things. That process and those rules, however, would implicitly define a meta-space of their own. And because this meta-space is pre-specified (implicitly, by that process or set of rules)

perform inference via guided search, and only some of the hypotheses within the space are actually evaluated.

in the exact same way the original hypothesis space was pre-specified (implicitly, by the original generative process), the hypotheses within it are "innate" in precisely the same way that the original hypotheses were. In general, the only way for something to be learned in the Fodorian sense – the sense that underlies this critique – is for them to be able to spring, willy-nilly, into a hypothesis space in such a way that is essentially random (i.e., unexplainable via some process or rule). If this is truly what learning is, it seems to preclude the possibility of studying it scientifically; but luckily, this is not what most of us mean by learning.

One consequence of this is that *every* model, even the brain, must come equipped with a latent hypothesis space that consists of everything that it can possibly represent and compute; all learning must happen within this space. This is not a novel or controversial point – all cognitive scientists accept that *something* must be built in – but it is often forgotten as soon as a Bayesian model appears. Somehow, the fact that hypotheses are explicit and hypothesis spaces are clearly defined makes them appear more "innate" than if they were simply latent in the model. But even connectionist networks – which are often believed to presume very little in the way of innate knowledge – implicitly define hypotheses and hypothesis spaces via their architecture, functional form, learning rule, etc. In fact, connectionist networks can be viewed as implementations of Bayesian inference (e.g., Funahashi, 1998; McClelland, 1998; MacKay, 2004), corresponding to a computational-level model whose hypothesis space is a set of continuous functions (e.g., Funahashi, 1989; Stinchcombe & White, 1989). This is a large space, to be sure, but Bayesian models can easily have hypothesis spaces that are equivalently large.

Does this mean that there is no difference between Bayesian models and connectionist networks? In one way, the answer is yes: because neural networks are universal approximators (Hornik, Stinchcombe, & White, 1989), it is always possible to construct one that is an implementation of a Bayesian model. In practice, however, the answer is usually no: the two methods have very different strengths and weaknesses, and therefore their value as modelling tools varies depending on the questions being

asked.[9] Bayesian models optimally trade off between simplicity and goodness-of-fit; connectionist models perform a similar tradeoff, but generally non-optimally and in a more *ad hoc* manner, avoiding overfitting by limiting the length of training and choosing appropriate weights, learning rules, and network architecture.[10] In the Bayesian framework, what is built in is the generative process, which implicitly defines the assignment of prior probabilities, the representation involved, and the size of the hypothesis space; in the connectionist framework, these things are built in through choices about the architecture, weights, learning rule, training procedure, etc.

It is therefore incorrect to say one framework assumes more innate knowledge than another: *specific models* within each may assume more or less, but it can be quite difficult to compare them precisely, in part because connectionist networks incorporate it implicitly. Which model assumes more innate knowledge is often not even the interesting question. A more appropriate one might be: *what* innate knowledge does it assume? Instead of asking whether one representation is a stronger assumption than another, it is often more productive to ask which predicts human behavior better. The answer will probably depend on the problem and the domain, but the great advantage of computational modelling is that it allows us to systematically explore this dependence precisely. And an advantage of Bayesian modelling in particular is that we can be explicit about what assumptions we are making, thus facilitating that exploration.

One thing that is often overlooked in evaluating what is built into computational models is *preprocessing*: any model – Bayesian approaches as well as neural networks – requires the data to be entered in a usable form. If there is too much preprocessing, or too much of the wrong sort, it can be difficult to tell if the interesting learning has been done by the model or the modeler. This is a general problem confronted by any computational modeler, and is one of the reasons for the truism that "every model is wrong." Unless we can give the model completely raw, un-preprocessed data

[9]Many connectionist modelers see the two approaches as complementary (e.g., Rogers & McClelland, 2004), and I agree with this view.

[10]There is an interesting subfield called Bayesian neural networks studying how to construct models that make these choices for themselves, pruning connections in a Bayes-optimal way (e.g., MacKay, 1995; Neal, 1994, 1996).

– and the model is a complete copy of the brain itself – we will be simplifying and introducing error somehow. Of course, if we do that, we then won't have learned anything: the point of modelling *is* to simplify, and to hope that we have done so in such a way that it adds to our understanding rather than detracts. The ultimate aim is to develop models that take data that is preprocessed less and less; we are not there yet, but Bayesian models are a powerful tool toward that end.

Conclusion

Bayesian modelling is a useful approach in cognitive science, especially when used in conjunction with other types of computational modelling as well as experimental work. Its representational flexibility makes the Bayesian approach applicable to a wide variety of learning problems, and its transparency makes it easy to be clear about what assumptions are being made and what is being learned. The framework is valuable for defining an optimal standard as well as for exploring and illustrating the tradeoff between simplicity and goodness-of-fit. As a result, it has the potential to explain many aspects of human cognition.

In the next chapters I present specific Bayesian models that address problems in three particular areas of language acquisition. We will see how the principles of Bayesian inference, presented here in a more abstract form, come into play in each of these specific contexts.

Chapter 3

The acquisition of abstract linguistic structure

Chapter 1 introduced a classic learnability argument that 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 I use a Bayesian framework for grammar induction to argue for a different possibility. I show that, given typical child-directed speech and certain innate domain-general capacities, an unbiased ideal learner could recognize the hierarchical phrase structure of language without having this knowledge innately specified as part of the language faculty. I discuss the implications of this analysis for accounts of human language acquisition, and focus in particular about implications for Poverty of the Stimulus arguments more generally.

Hierarchical phrase structure in language

Scientific inquiry in language acquisition was influenced by Chomsky's observation that language learners make grammatical generalizations that appear to go beyond

The work in this chapter was carried out in collaboration with Terry Regier and Joshua Tenenbaum.

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

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

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

that has sparked many discussions of innateness, from Chomsky's original discussions to present-day debates: the phenomena of auxiliary fronting in constructing English interrogative sentences (Laurence & Margolis, 2001; Lewis & Elman, 2001; Legate & Yang, 2002; Pullum & Scholz, 2002; Reali & Christiansen, 2005). Nevertheless, this analysis should not be seen as an attempt to explain the learnability of auxiliary fronting (or any specific linguistic rule) *per se*. Rather the goal is to address the general phenomenon of hierarchical phrase structure in syntax – the phenomenon that has been argued to underlie the learning of auxiliary fronting and many other specific rules. I take as data an entire corpus of child-directed speech and evaluate hypotheses about candidate grammars that could account for the corpus as a whole. As a byproduct of this, the analysis allows us to explore the learnability of auxiliary fronting and other related specific aspects of syntax.

The debate

As we saw in Chapter 1, Chomsky put forth several Poverty of the Stimulus arguments for the claim that part of the initial state of the language acquisition system – and a necessary feature of any possible hypothesis considered by the language learner – is the knowledge that the rules of syntax are defined over hierarchical phrase-structure representations rather than a linear sequence of words. I presented the following abstract schema for PoS arguments in general:

1.1. (i) Children show a specific pattern of behavior B.

(ii) A particular generalization G must be grasped to produce behavior B.

(iii) It is impossible to reasonably induce G simply on the basis of the data D that children receive.

(iv) $\therefore$ Some abstract knowledge T, limiting which specific generalizations G are possible, is necessary.

This form of the PoS argument makes explicit the distinction between multiple levels of knowledge – a distinction that I consider in one form or another throughout

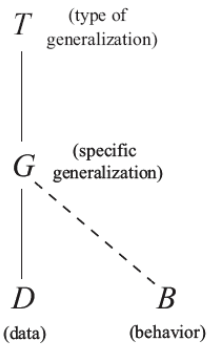


Figure 3-1: Graphical depiction of the standard Poverty of Stimulus argument. Abstract higher-level knowledge T is necessary to constrain the specific generalizations G that are learned from the data D, and that govern behavior B.

this thesis, and also illustrate schematically in Figure 3-1. In the case of auxiliary fronting, the specific generalization G refers to the hierarchical rule 1.3(b) in Chapter 1 that governs the formation of interrogative sentences (also shown, for ease of reference, below). The learning challenge is to explain how children come to produce only the correct forms for complex interrogatives (B), apparently following a hierarchical rule like 1.3(b), 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 like the linear rule 1.3(a).

1.3. (a) Linear: Form the interrogative by moving the first occurrence of the auxiliary in the declarative to the beginning of the sentence.

(b) Hierarchical: Form the interrogative by moving the auxiliary from the main clause of the declarative to the beginning of the sentence.

Note that 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. 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).

As I have discussed, 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:

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

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

(iii) $\therefore$ T must be innate.

Given this schema, the argument here can be construed in two different ways. On one view, I am arguing against premise 1.2(ii); I 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, I evaluate knowledge at both levels (T and G) together by drawing on the methods of hierarchical Bayesian models and Bayesian model selection (A. Gelman et al., 2004). Interestingly, these results suggest that less data is required to learn T than to learn the specific grammar G.

On another view, I am not arguing with the form of the PoS argument, but merely clarifying what content the knowledge T must have. I argue that phenomena such as children's mastery of auxiliary fronting are not sufficient to require that the innate knowledge constraining generalization in language acquisition be language-specific. Rather it could be based on more general-purpose systems of representation and inductive biases that favor the construction of simpler representations over more complex ones.

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

3.1. (a) Children do not show the pattern of behavior B.

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

(c) 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 response 3.1(a) 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 and Nakayama (1987) demonstrated that children do not spontaneously form incorrect complex interrogatives such as "Is the man who hungry is ordering dinner?" (1.5(b)), they make other mistakes that are not so easily interpretable. For instance, one might utter a sentence like "Is the man who is hungry is ordering dinner?", which is not immediately compatible with the correct hierarchical grammar but might be consistent with a linear rule. Additionally, recent research by Ambridge, Rowland, and Pine (2005) suggests that 6 to 7 year-old children presented with auxiliaries other than is do indeed occasionally form incorrect sentences along the lines of 1.5(a), such as "Can the boy who run fast can jump high?", as well as other kinds of errors.

A different response, 3.1(c), accepts that children have inferred the correct hierarchical rule for auxiliary fronting, but maintains that the input data is sufficient to support this inference. If children observe sufficiently many complex interrogative sentences like 1.5(b) while observing no incorrect sentences like 1.5(a), then perhaps they could learn directly that the hierarchical rule 1.3(b) is correct, or at least better supported than simple linear alternatives. The force of this response depends on how many grammatical complex interrogatives like 1.5(b) 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 and 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

[1]Adam (Brown corpus, 1973) and Nina (Suppes corpus, 1973); for both, see MacWhinney (2000).

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

Indirect evidence also plays a role in the second type of reply, 3.1(b), which is probably the most currently popular line of response to the PoS argument. The claim is that children could still show the correct pattern of linguistic behavior – acceptance or production of sentences like 1.5(b) but not 1.5(a) – even without having learned any grammatical rules like 1.3(a) or 1.3(b) at all. Perhaps the data, while poor with respect to complex interrogative forms, are rich in distributional and statistical regularities that would distinguish the linear rule 1.3(a) from the hierarchical rule 1.3(b). If children pick up on these regularities, that could be sufficient to explain why they avoid incorrect complex interrogative sentences like 1.5(a), without any need to posit the kinds of grammatical rules that others have claimed to be essential (Redington, Chater, & Finch, 1998; Lewis & Elman, 2001; Reali & Christiansen, 2004, 2005).

For instance, Lewis and 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 and Christiansen (2005) showed that the statistics of actual child-directed speech support such predictions. 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 1.5(a) and 1.5(b).

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 and Christiansen (2005) and Lewis and Elman (2001) measure success by whether they predict the next word in a sequence or by comparing the prediction error for grammatical and ungrammatical sentences. These networks lack not only a grammar-like representation; they lack any kind of explicitly articulated representation of the knowledge they have learned. It is thus difficult to say what exactly they have learned about linguistic structure.

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

Overview

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

These results emerge because an ideal learner must trade off simplicity and goodness-of-fit in evaluating hypotheses, just as we saw in Chapter 2. The notion that inductive learning should be constrained by a preference for simplicity is widely shared among scientists, philosophers of science, and linguists. Chomsky himself concluded that natural language is not finite-state based on informal simplicity considerations (1956, 1957), and suggested that human learners rely on an evaluation procedure that incorporates simplicity constraints (1965). Just as hypothesis B offers the "just right" balance between simplicity and goodness-of-fit in Figure 2-2, I argue that in a similar way, a hierarchical phrase-structure grammar yields a better tradeoff than linear grammars between simplicity of the grammar and fit to typical child-directed speech. But although these findings suggest that the specific feature of hierarchical structure can be learned without an innate language-specific bias, I do not argue or believe that all interesting aspects of language will have this characteristic.

One finding of this 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 this 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. Throughout this thesis I investigate what drives this perhaps counterintuitive result and discuss its implications for language acquisition and cognitive development more generally.

Method

I cast the problem of grammar induction within a hierarchical Bayesian framework[2] whose structure is shown in Figure 3-2. 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.

The 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:

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

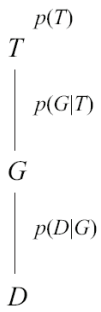


Figure 3-2: A hierarchical Bayesian model for assessing Poverty of Stimulus arguments. The model is organized around the same structure as Figure 3-1, but now each level of representation defines a probability distribution for the level below it. Bayesian inference can be used to make inferences at higher 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 biases for specific types of grammars can be encoded in the prior $p(T)$, although here I consider an unbiased prior, with $p(T)$ equal for all T.

$$p(G, T|D) \propto p(D|G)p(G|T)p(T). \tag{3.1}$$

We wish to explore learning when there is no innate bias towards grammars with hierarchical phrase structure. This is implemented in this 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).

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 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). This analysis is complementary to these previous Bayesian analyses. The main difference is that instead of addressing learnability issues in abstract and highly simplified settings, I focus on a specific question – the learnability of hierarchical structure in syntax – and evaluate it on realistic data: a finite corpus of child-directed speech. As with the input data that any child observes, this corpus contains only a fraction of the syntactic forms in the language, and probably a biased and noisy sample at that.

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

An ideal analysis of learnability

This analysis views learnability in terms of an ideal framework in which the learner is assumed to be able to effectively search over the joint space of G and T for grammars that maximize the Bayesian scoring criterion. (In other words, we set aside one learnability issue, centered on the computational tractability of performing the search, to focus on the issue of sufficiency of the data. We address the implications of this focus in the discussion). This proposal is therefore not a comprehensive or mechanistic account of how children actually acquire language; the full problem of language acquisition poses many challenges that I do not consider here. Rather, this analysis provides a formal framework for analyzing the learnability of some aspects of linguistic syntax, with the goal of clarifying and exploring claims about what language-specific prior knowledge must be assumed in order to make learning possible. I discuss the reasoning behind performing this sort of ideal learnability analysis in far more detail in Chapters 2 and 6, but the essential goal is to follow the spirit of how Chomsky and other linguists have considered learnability: as a question of what is learnable in principle. Is it in principle possible given the data a child observes to learn that language is governed by hierarchical phrase-structure rules, rather than linear rules, if one is not innately biased to consider only hierarchical grammars? If we can show that such learning is in principle possible, then it becomes meaningful to ask the

algorithmic-level question of how a system might successfully and in reasonable time search the space of possible grammars to discover the best-scoring grammar.

Of course, the value of this ideal learnability analysis depends on whether the specific grammars I 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). I therefore examine grammars generated in a variety of ways:

(1) The best hand-designed grammar of each grammar type.

(2) The best grammars found via local search, using the grammar from (1) as the starting point.

(3) The best grammars found in a completely automated fashion

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

The corpora

The corpus consists of the sentences spoken by adults in the Adam corpus (Brown, 1973) of the CHILDES database (MacWhinney, 2000). In order to focus on grammar learning rather than lexical acquisition, each word is replaced by its syntactic category.[3] Although learning a grammar and learning a lexicon are probably tightly linked, I 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, Wilson, & Lewis,

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

2005). In addition, ungrammatical sentences and the most grammatically complex sentence types are removed from the corpus.[4] The complicated sentence types are removed for reasons of computational tractability as well as the difficulty involved in designing grammars for them, but this is if anything a conservative move since these results suggest that the hierarchical grammars will be more preferred as the input grows more complex. The final corpus contains 21671 individual sentence tokens corresponding to 2336 unique sentence types, out of 25755 tokens in the original corpus.[5]

In order to explore how the preference for a grammar depends on the amount of data available to the learner, I 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.[6]

An additional variable of interest is what evidence is available to the child at different ages. As in Chapter 5, I approximate this by splitting the corpora into five equal sizes by age. The Adam corpus has 55 files, so I define the earliest (*Epoch 1*) corpus as the first 11 files. The *Epoch 2* corpus corresponds to the cumulative input

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

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

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

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, I wish to compare grammar types T that differ from each other structurally in the same way. Different Bayesian approaches to evaluating alternative grammar types are possible. In particular, we could score a grammar type T by integrating the posterior probability over all specific grammars G of that type ($\sum_G p(T, G|D)$) or by choosing the best G of that type ($max_G p(T, G|D)$). Integrating over all grammars is computationally intractable, and arguably also less relevant. Ultimately it is the specific grammar G that governs how the learner understands and produces language, so we should be interested in finding the best pair of T and G jointly. I 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, I 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. Since the 1950s, they have often been treated as a first approximation to the structure of natural language since the early period of generative linguistics (Chomsky, 1959). Probabilistic context-free grammars (PCFGs) are a probabilistic generaliza-

tion of CFGs commonly used in statistical natural language processing (Manning & Schütze, 1999; Jurafsky & Martin, 2000), and I incorporate powerful tools for statistical learning and inference with PCFGs in this work here. I recognize that there are also many aspects of syntax that cannot be captured naturally in CFGs. In particular, they do not express the sort of movement rules that underlie some standard generative accounts of aux-fronting: they do not represent the interrogative form of a sentence as a transformed version of a simpler declarative form. I work with CFGs because they are the simplest and most tractable formalism suitable for this purposes here – assessing the learnability of hierarchical phrase structure in syntax – but in future work it would be valuable to extend these analyses to richer syntactic formalisms.

I consider three different approaches for representing linear grammars. The first is based on regular grammars, also known as finite-state grammars. Regular grammars were originally proposed by Chomsky as a "minimal linguistic theory", a kind of "null hypothesis" for syntactic theory. They are models capable of producing a potentially infinite set of strings, but in a manner that is sensitive only to the linear order of words and not the hierarchical structure of syntactic phrases. A second approach, which I 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, I consider the one-state (1-ST) grammar, which maximizes simplicity by sacrificing goodness-of-fit. It permits any syntactic category to follow any other and is equivalent to a finite automaton with one state in which all transitions are possible (and is similar to a standard unigram model). Though these three approaches may not capture exactly what was originally envisioned as linear grammars, I 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. I design both context-free and regular grammars that span the range between these two extremes (much as the models in Figure 2-2 do); within each type, specific grammars differ systematically in how they capture the tradeoff between simplicity and goodness-of-fit. Table 3.1 contains sample productions from each of the specific grammars.[7]

I 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.

I consider three regular grammars spanning the range of the simplicity/goodness-

[7] All full grammars, corpora, and perplexity values (corresponding to all likelihood calculations) may be found at http://www.mit.edu/ perfors/posgrammars.html.

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 I consider. It has 85 non-terminals and 389 productions, some examples of which are shown in Table 3.1. 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

Table 3.1: 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 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

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

Grammars constructed by automated search

There is reason to believe that hand-designed grammars provide a good approximation of the best grammar of each type. Both context-free grammars are designed based on linguistic intuition, and the regular grammars are constructed from the context-free grammars in order to preserve as much linguistic structure as possible. Furthermore, grammars of all types have been chosen to reflect the range of tradeoffs between simplicity and goodness-of-fit. It would nevertheless be ideal to search the space of possible grammars and compare the resulting best grammars found for each type,

rather than simply comparing the best hand-designed grammars. Unfortunately, this type of search for context-free grammars presents a difficult computational problem, and current search algorithms cannot be relied upon to find the optimal grammar of any given type on large-scale corpora. Fortunately, this argument requires only a search over regular grammars: if the 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.

I 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 the 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 to a regular grammar containing those non-terminals and no others.[8]

As another comparison, I 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. The partial search was inspired by the work of Stolcke and 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 the appendix.

[8]It is not assumed that each syntactic category has one corresponding non-terminal, or vice versa; both may be ambiguous. Though the algorithm incorporates a prior that favors fewer hidden variables (non-terminals), it requires the modeler to specify the maximum number of non-terminals considered. I 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, I 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.

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. I formalize it using a generative model under which each grammar is selected from the space of all grammars of that type. More complex grammars are those that result from more (and more specific) choices. As we saw in Chapter 2, this method of scoring simplicity is quite general, not restricted to grammars or even language. Just as the more complex hypotheses in Figure 2-2 are those that require more free parameters to specify, more complex probabilistic grammars G require more choices during the generation process. If one were generating a grammar from scratch, one would have to make the series of choices depicted in Figure 3-3, 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 the vocabulary V, the prior probability is:

$$p(G|T) = p(n) \prod_{k=1}^{n} p(P_k) p(\theta_k) \prod_{i=1}^{P_k} p(N_i) \prod_{j=1}^{N_i} \frac{1}{V}. \tag{3.2}$$

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

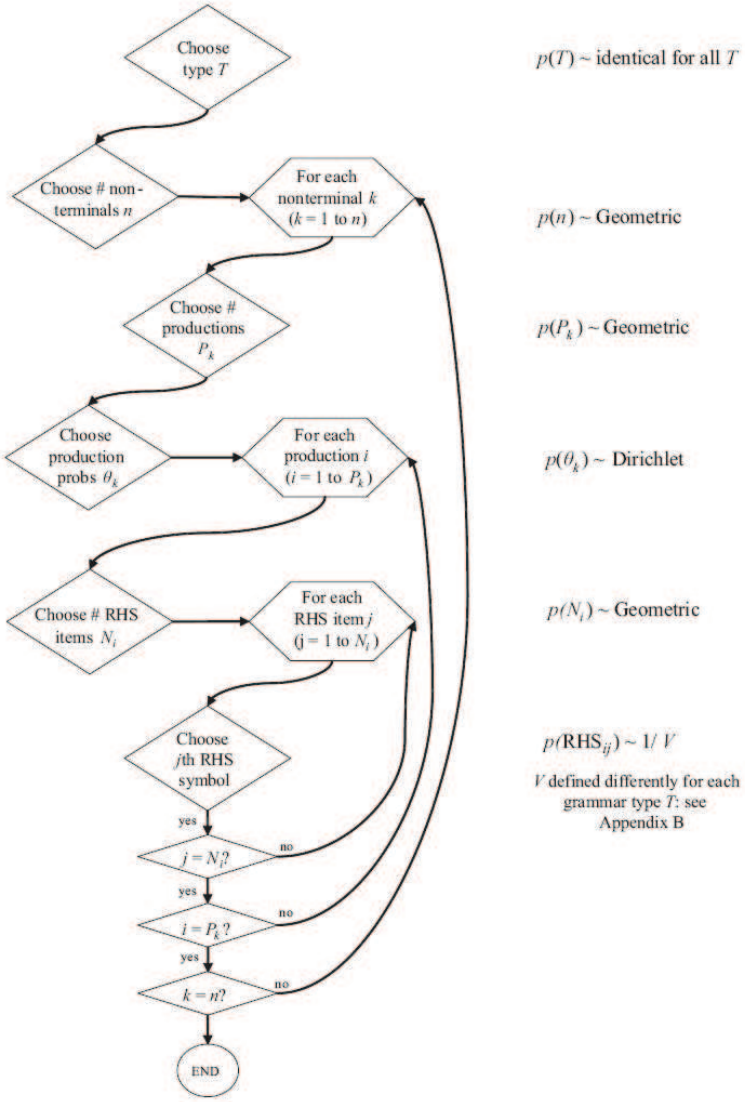


Figure 3-3: Flowchart depicting the series of choices required to generate a grammar. More subtle differences between grammar types are discussed in the appendix.

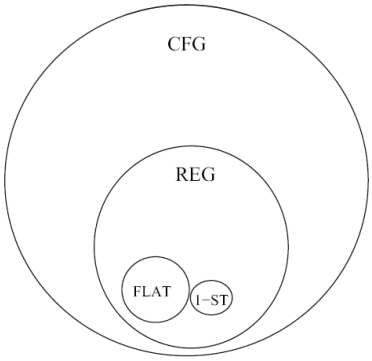


Figure 3-4: Venn diagram depicting the relation of the grammar types T to each other. The set of context-free grammars contains regular, flat, and one-state grammar types as special cases. Flat and one-state grammars are themselves special cases of regular grammars.

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

The subsets of grammars that can be generated by the several grammar types I 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. One-state and flat grammars are a subset of regular grammars, which are a subset of context-free grammars (see Figure 3-4). 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.

This preference for the simplest grammar type is related to the Bayesian Occam's razor (MacKay, 2004). 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 intuitive idea 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. This is the standard approach in probabilistic grammar induction (e.g., Stolcke & Omohundro, 1994; Manning & Schütze, 1999). However, this assumption is not well-matched to the induction problem we address here. Most centrally, these models generate a statistical distribution of sentences that differs systematically from the statistics of natural language in a way that might be expected to lead to a bias in grammar induction. In particular, they produce distributions that do not capture the well-attested power law organization of language (e.g., Zipf, 1932).[9] They also fail to capture the actual causal process underlying language production. Many factors – discourse requirements, semantics, priming and salience effects, subjective preference, the conversational context, etc. – all affect which sentences are spoken that are not captured by the production probabilities of the grammar; a model that presumes that sentences are emitted independently from one another based on those probabilities is therefore descriptively and explanatorily inadequate.

Motivated by the fact that conventional probabilistic grammars that generate every sentence independently fail to capture the power-law statistics of natural language, Goldwater, Griffiths, and Johnson (2006) introduced a modified approach to grammar induction known as adaptor grammars (see also Johnson, Griffiths, and Goldwater (2007)). I adopt a version of their approach here. In the adaptor grammar framework, the likelihood is calculated assuming a language model that is divided into two components. The first component, the generator, assigns a probability dis-

[9] See also Briscoe (2006) for a more recent overview.

tribution 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. (I sometimes refer to this component as simply the "grammar"). The second component, the adaptor, produces a finite observed corpus through a nonparametric stochastic process that mixes draws from the generating grammar with draws from a stored memory of previously produced sentence forms. 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.

An important special case of the adaptor framework, corresponding to particular settings of the adaptor parameters[10], evaluates a grammar based on the probabilities it assigns to the set of sentence types occurring in a corpus, independent of the frequencies with which these types occur (i.e., the sentence token frequencies). Goldwater et al. (2006) note that 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. Since we are concerned with grammar comparison rather than corpus generation, I focus in this work on the first component of the model. I thus take the data to consist of the set of sentence types (i.e., distinct sequences of syntactic categories) that appear in the corpus, and evaluate the likelihoods of candidate probabilistic grammars on that dataset.

It is important to note although the model is presented with type-based information only, I do not presume that human children ignore token frequencies in general.[11] The adaptor grammar framework explicitly captures people's well-attested sensitivity to token-based frequency information via the adaptor component of the model; in fact, to the extent that token-based frequencies follow a power-law distribution, the adaptor framework captures that sensitivity better than more standard approaches.

[10] In which $\beta =$ and $\alpha \rightarrow 0$.

[11] There are two different sorts of token-based frequency information, and the distinction between the two is somewhat important. Much of the evidence demonstrating a sensitivity to this sort of frequency information concerns the token frequency of individual lexical items. Sentence tokens, in our model, correspond to the unique strings of syntactic sequences (e.g., *n aux adj*) rather than unique lexical items such as dog or run. We do not presume that children ignore either kind of token-based frequency information – simply that different sorts of information are relevant to different tasks.

The model corresponds to assuming that language users can generate the syntactic forms of sentence tokens either by drawing on a memory store of familiar syntactic types, or by consulting a deeper level of grammatical knowledge about how to generate the infinite variety of acceptable syntactic forms in the language. Any sentence type generated by the former system would originally have been generated from the latter, but speakers need not consult their deep grammatical knowledge for every sentence token they utter or comprehend.

It is the deeper grammatical knowledge used to generate the set of sentence types – not the memory-based generation process that produces the observed frequency of sentence tokens – that I am interested in here. By focusing only on the first component of the language model (for which the data consists of sentence types) I presume that token frequencies – while important in many phenomena in language acquisition that I do not model here, including learning individual lexical items, categories, or constructions, or for explaining performance effects due to memory or attention – may be less relevant to the problem of identifying which particular sentence forms are grammatical or not. In the discussion I address this presumption in more detail.

The likelihood assigned to a grammar based on a dataset of sentence types can be interpreted as a measure of how well the grammar fits or predicts the data. Like simplicity, this notion of "fit" may be understood in intuitive terms that have nothing specifically to do with grammars or language. Consider again Figure 2-2: intuitively it seems as if hypothesis B is more likely to be the source of the data than hypothesis A, but why? If A were the correct model, it would be quite a coincidence that all of the data points fall only in the regions covered by B. Likelihood is dependent on the quantity of data observed: it would not be much of a coincidence to see just one or a few data points inside B's region if they were in fact generated by A, but seeing 1000 data points all clustered there – and none anywhere else – would be very surprising if A were correct.

This probabilistic preference for the most specific grammar consistent with the observed data is a version of the size principle in Bayesian models of concept learn-

ing and word learning (Tenenbaum & Griffiths, 2001; Regier & Gahl, 2004; Xu & Tenenbaum, 2007). As noted before, it can also be seen as a probabilistic version of the subset principle (Wexler & Culicover, 1980; Berwick, 1986). The effective set of sentences that the 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 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:

$$p(D|G) = \prod_{l=1}^{M} p(S_l|G). \tag{3.3}$$

The probability of any sentence type S_l given the grammar ($p(S_l|G)$) is the product of the probabilities of the productions used to derive S_l. Thus, calculating likelihood involves solving a joint parsing and parameter estimation problem: identifying the possible parse for each sentence in the corpus, as well as calculating the parameters for the production probabilities in the grammar. I 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). I evaluate Equation 3.3 in the same way, using the

maximum-likelihood parameter values but integrating over all possible parses of the corpus.[12] 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 3.3. This notion of simplicity in derivation captures an inductive bias favoring grammars that assign the observed sentences more economical derivations – a bias that is distinct and complementary to that illustrated in Figure 2-2, which favors grammars generating smaller languages that more tightly cover the observed sentences.

Results

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

Posterior probability on different grammar types

Hand-designed grammars

Table 3.2 shows the prior, likelihood, and posterior probability of each handpicked grammar on each corpus. When there is the least evidence in the input (corpus *Level 1*), the flat grammar is preferred. As the evidence accumulates, the one-state grammar scores higher. However, for the larger corpora (*Level 4* and higher), a hierarchical grammar always scores the highest, more highly than any linear grammar.

If linear grammars are *a priori* simpler than context-free grammars, why does the prior probability favor context-free grammars on more complex corpora? Recall that I 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

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

Table 3.2: Log prior, likelihood, and posterior probabilities of each hand-designed grammar for each level of evidence. Because numbers are negative, smaller absolute values correspond to higher probability. If two grammars have log probabilities that differ by n, their actual probabilities differ by e^n; thus, the best hierarchical grammar CFG-L is $e^{101}(\sim 10^{43})$ times more probable than the best linear grammar REG-M.

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

of compression offered by the context-free grammar is sufficient to overwhelm the initial simplicity preference towards the others. This is evident by comparing the size of each grammar for the smallest and largest corpora. On the *Level 1* corpus, the context-free grammars require more productions than do the linear grammars (17 productions for CFG-S; 20 for CFG-L; 17 for REG-N; 15 for REG-M; 14 for REG-B; 10 for 1-ST; 8 for FLAT). Thus, the hierarchical grammars have the lowest initial prior probability. However, their generalization ability is sufficiently great that additions to the corpus require relatively few additional productions: the context-free grammars that can parse the *Level 6* corpus have 69 and 120 productions, in comparison to 117 (REG-B), 169 (REG-M), 389 (REG-N), 25 (1-ST), and 2336 (FLAT).

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

As the amount of data accumulates, the posterior increasingly favors the hierarchical grammars: the linear grammars are either too complex or fit the data too poorly by comparison. This ideal learning analysis thus infers that the syntax of English, at least as represented by this corpus, is best explained using the hierarchical phrase structures of context-free grammars rather than the linear structures of simpler Markovian grammars. In essence, this analysis reproduces one of the founding

insights of generative grammar (Chomsky, 1956, 1957): hierarchical phrase-structure grammars are better than Markovian linear grammars as models of the range of syntactic forms found in natural language. Child learners could in principle make the same inference, if they can draw on the same rational inductive principles.

Berwick (1982) presented a different approach to formalizing the simplicity argument for hierarchical structure in syntax, using tools from automata theory and Kolmogorov complexity. At a high level, this analysis and Berwick's are similar, but there are two important differences. First, I evaluate learnability of a sizeable and realistic CFG for English on a natural corpus of child-directed speech, rather than simple languages (e.g., palindrome or mirror-symmetry languages on a binary alphabet) with idealized corpora (e.g., including all sentences generated by the grammar, or all sentences up to some maximum length or depth of recursion). Second, rather than considering only those grammars that fit the corpus precisely and evaluating them based only on their simplicity, I adopt a framework in which simplicity of the grammar trades off against how well the grammar fits the particular corpus.

We can see where these differences matter in comparing the scores of particular grammars at different levels of evidence. Although regular grammars never receive the highest score across grammars of all types, they all score higher on the smallest corpus (*Level 1*) than do any of the hierarchical grammars. On most levels of evidence, at least one regular grammar is preferred over CFG-L. CFG-L is in fact the grammar that is ultimately favored on the full corpus, but it overgeneralizes far more than the regular grammars, and for the smaller corpora the tradeoff between simplicity and fit weighs against it. Thus these results cannot be obviously predicted by simplicity-based learnability analyses developed for artificial grammars on idealized corpora. The tradeoff between simplicity and degree of fit, given the sparse and idiosyncratic nature of child-directed language input, is critical to determining what kind of grammar an ideal learner should acquire.

It is interesting that the smallest corpora are best accounted for by the flat and one-state grammars. The smallest corpus contains only eight sentence types, with an average 1.6 words per sentence; thus, it is not surprising that it is optimal to

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

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

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

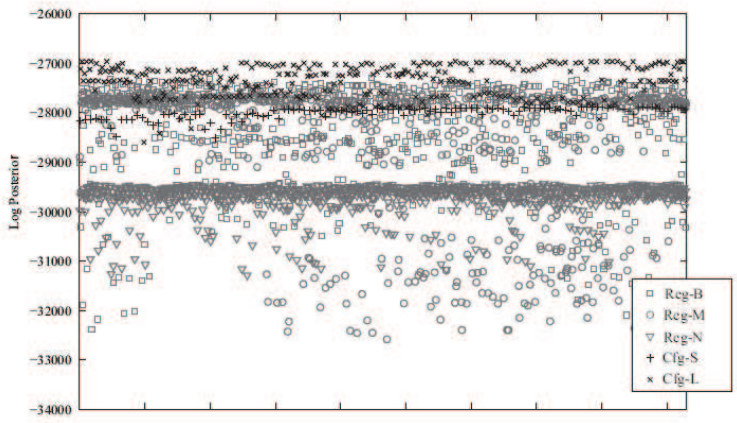


Figure 3-5: Posterior probabilities of each grammar considered during the search of regular and context-free possibilities. Each point represents one grammar; the x-axis is meaningless.

Local search from hand-designed grammars

To what extent are these results dependent on these particular hand-designed grammars? I address this question by analyzing the posterior scores of those grammars identified via local search. Figure 3-5 depicts the posterior probabilities of all of the grammars considered in the search. Many linear grammars found by the automatic search procedures have scores similar to the best hand-designed linear grammars, but none have posterior probabilities close to that of the best context-free grammars.

The posterior probabilities of the best grammars of each type found after the local search are shown in Table 3.3. The results are qualitatively similar to those obtained with hand-designed grammars: the posterior still favors a context-free grammar once the corpus is large enough, but for smaller corpora the best grammars are linear.

Identifying a regular grammar by automated search

In addition to identifying the best grammars resulting from a local search, I also examine the best regular grammar (REG-AUTO) found in a purely automated fashion using the unsupervised learning model developed by Goldwater and Griffiths (2007). The grammars with the highest posterior probability on each corpus are shown in Table 3.4. All of the REG-AUTO grammars have posterior probabilities similar

Table 3.3: Log prior, likelihood, and posterior probabilities of each grammar 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**

to those of the other regular grammars, but on the larger corpora none have higher probability than the best context-free grammars. Because the REG-AUTO grammars do not consistently have higher probability than the other regular grammars, we cannot conclude that they represent the "true best" from the space of all possible grammars of that type. However, the fact that the best regular grammars found by every method have a similar order-of-magnitude probability – and that none have been found that approach the best-performing context-free grammar – suggests that if better regular grammars do exist, they are not easy to discover.

Summing up these results, we find that a context-free grammar always has the highest posterior probability on the largest corpus, compared to a variety of plausible linear grammars. Though the ability of the hierarchical grammars to generate a higher variety of sentences from fewer productions typically results in a lower likelihood, this compression helps dramatically in the prior. A hierarchical grammar thus consistently maximizes the tradeoff between data fit and complexity.

Ungrammatical sentences

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

Table 3.4: 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**

them is not critical to the outcome.[13] To the best grammars of each type, I added the minimum number of additional productions required to parse the ungrammatical corpus. The hierarchical grammars still have the highest posterior probability (*Level 6* posterior: CFG-L: -29963; REG-B: -30458; REG-M: -30725; CFG-S: -31008; REG-N: -33466; 1-ST: -43098; FLAT: -92737). Thus, considering the ungrammatical sentences along with the grammatical sentences does not qualitatively alter these findings.

Sentence tokens vs sentence types

The likelihood was defined under a language model with separate generative processes: one for the allowable types of syntactic forms in a language, another for the frequency of specific sentence tokens. This definition effectively restricts the data considered by the model to only include sentence types, rather than each individual sentence token. Defining the likelihood in this way was a principled choice, paralleling standard linguistic practice, in which 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.

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

This result suggests that if the hierarchical structure of syntax is to be inferred from observed data based on Bayesian inference with probabilistic grammars, the

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

learner must have some sort of disposition to evaluate grammars primarily with respect to type-based rather than token-based data. Although one possibility is that this disposition emerges from a general insensitivity to token frequencies, we certainly do not suggest that this is the case: there is extensive psycholinguistic and developmental evidence demonstrating that people are sensitive to quantitative frequency variations in a wide variety of contexts. It is more plausible that grammar induction *per se* is based on more sophisticated grammar models (like 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 component of the representation that captures the deep principles of the grammar that generate acceptable forms. We will return to this issue in the discussion.

Age-based stratification

These results may have developmental implications, but these must be interpreted with caution. The 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. I therefore compare the posterior probabilities of the grammars on the Epoch corpora, which were constructed creating age-based divisions in the full corpus. Table 3.5 shows the probabilities of the best hand-designed linear

and hierarchical grammars on these corpora. Strikingly, a context-free grammar is preferred at every age. This is even true for grammars that correspond to just the first file (*Epoch 0*), which consists of one hour of conversation at age 2;3. It is also interesting that the prior probabilities of the CFG-S and CFG-L grammars beginning at *Epoch 3* do not change. Why is this? Recall that at each epoch and level, I 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? No. 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 this model. However, this analysis does show that the data supporting a hierarchical phrase structure for English are so ubiquitous that once a learner has some ability to assign syntactic categories to words and to parse sentences of sufficient complexity, it should be possible to infer that hierarchical grammars provide the best description of the language's syntax. It is interesting and theoretically important that the amount of data required to infer the existence of hierarchical phrase structure is much less than is required to infer all the rules of the correct hierarchical phrase-structure grammar. In terms of Figures 3-1 and 3-2, 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 I have not demonstrated this here, it is theoretically possible that during the course of acquisition, higher-level knowledge,

Table 3.5: Log prior, likelihood, and posterior probabilities of each grammar type on the *Epoch* corpora, which reflect an age split. The hierarchical grammars are favored at all stages, 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	-2443	**-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**

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. I will return to this point in the discussion.

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 3.6 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 it 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%),

Table 3.6: Proportion of sentences in the full corpus that are parsed by smaller grammars of each type. The *Level 1* grammar is the smallest grammar of that type that can parse the *Level 1* corpus. All *Level 6* grammars 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%

REG-B (91.0%), and REG-N (87.2%) trailing. Although the magnitude of these differences is not large, they once again follow the same pattern as that noted 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 ungrammatical 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 these grammars lack the expressivity needed to encode important syntactic constraints, such as agreement. However, it is interesting that the REG-M grammar, which generalizes less than either context-free grammar to the full corpus in Table 3.6, 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 more incorrect (ungrammatical) sentences than the others.

The hierarchical grammars also generalize more appropriately than the linear grammars in the specific case of auxiliary-fronting for interrogative sentences. As Table 3.7 shows, both context-free grammars can parse aux-fronted interrogatives containing subject NPs that have relative clauses with auxiliaries – Chomsky's critical forms – despite never having seen an example of these forms in the input. They can do so because the input does contain simple declaratives and interrogatives, which license interrogative productions that do not contain an auxiliary in the main clause. The input also contains relative clauses, which are parsed as part of the noun phrase using the production [NP → NP CP]. Both context-free grammars can therefore parse an interrogative with a subject NP containing a relative clause, despite never having seen that form in the input.

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

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

Table 3.7: Ability of each grammar to parse specific sentences. The complex declarative "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	–	Can eagles that are alive fly? (aux n comp aux adj vi)	–	–	–	–	Y	Y	Y
Int Complex	–	* Are eagles that alive can fly? (aux n comp adj aux vi)	–	–	–	–	Y	–	–

cannot parse it, and therefore cannot parse the complex interrogative sentence in which it occurs.

One might argue that because none of these grammars can parse the incorrect complex interrogative (e.g., "Are eagles that alive can fly?"), we have not engaged with the precise original argument. Although we do not seek to address the issue of structure dependence, which this question more directly bears on, it is nonetheless useful to compare the performance of these grammars to the same grammars containing the minimal number of additional productions necessary to parse such a sentence. 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 corpus, and the inside-outside algorithm automatically prunes them away. If it is forced not to, 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 to similar grammars that lack those productions.

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

Discussion

This model of language suggests that there may be sufficient evidence in the input for an ideal rational learner to conclude that language has hierarchical phrase structure without having an innate language-specific bias to do so. The best-performing grammars correctly form interrogatives by fronting the main clause auxiliary, even though the input contains none of the crucial data Chomsky identified. In this discussion, I 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. This modelling 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. 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 argument I consider here is concerned with whether a particular feature of linguistic syntax – hierarchical structure – must be innately specified as part of a language-specific learning module in children's minds. This 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. I 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 I have analyzed are sufficiently close to the optimal ones. Advances in computational linguistics and the development of more powerful models of unsuper-

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

Even if this 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 (Kearns & Valiant, 1989). 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. These results do not suggest that any of these possibilities is more likely than the others. The core argument concerns only the specific need for a bias to *a priori* prefer analyses of syntax that incorporate hierarchical phrase structure. I am 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, this model incorporates some assumptions about the representational abilities of the child. First of all, I assume that children have the (domain-general) capacity to represent various types of grammars, including both hierarchical and linear grammars. I am not claiming that the specific grammars I analyze are exactly the ones children represent; clearly all of the grammars I have worked with are oversimplified in many ways. But an assumption that children in some sense have the capacity to represent both linear and hierarchical patterns of sequential structures – linguistic or non-linguistic – is necessary to even ask the questions I consider here. If children were not born with the capacity to represent the thoughts they later grow to have, no learning in that

direction could possibly occur. This analysis also assumes that the learner represents the different grammar types as different grammar types, choosing between context-free, regular, flat, and one-state grammars. This stratification is not critical to the results, however. If anything, it is a conservative assumption, because it favors the non-hierarchical grammars more heavily than they would be favored if I treated all grammars as instances of a single general type.[14]

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). I am 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. Another reason is that the reality of linguistic structure is widely accepted in standard linguistics. There are many linguistic phenomena whose only satisfying explanations (to date, not in principle) have been framed in structured symbolic terms. Explicitly structured representations also provide the basis of most state-of-the-art approaches in computational linguistics (e.g., Charniak, 1993; Manning & Schütze, 1999; Collins, 1999; Eisner, 2002; Klein & Manning, 2004). Given how useful structured grammatical representations have been both for explaining linguistic phenomena and behavior and for building effective computer systems for natural language processing, it seems worthwhile to take seriously the possibility that they might be the substrate over which children represent and learn language.

A final set of assumptions concerns the way I represent the input to learning. I

[14]See the appendix for details.

have given the model a corpus consisting of sequences of syntactic categories, corresponding to types of sentences, rather than sequences of lexical items which would correspond to actual sentence tokens.[15] There are two assumptions here: first, that it is reasonable to use syntactic categories rather than individual lexical items, and second, that it is reasonable to use input consisting of sentence types rather than sentence tokens.

Both of these assumptions are not necessary in principle, but they greatly simplify the analysis. 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. I make no claims about how children might initially acquire these syntactic categories, and the analysis would not change if they themselves were shown to be innate (whatever that would mean). There is some evidence that aspects of this knowledge may be in place even in children below the age of two (Booth & Waxman, 2003), and that syntactic categories may be learnable from simple distributional information without reference to the underlying grammar (Schütze, 1995; Redington et al., 1998; Mintz, Newport, & Bever, 2002; Gerken et al., 2005; Griffiths, Steyvers, Blei, & Tenenbaum, 2005). Thus I think it is plausible to assume that children have access to something like the input I 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 I assumed here. It would be valuable to explore whether hierarchical grammars continue to score better than linear grammars if the input to learning consists of automatically labelled part-of-speech tags rather than hand-labelled syntactic categories.

[15] Tomasello (2000) and others suggest that children initially restrict their syntactic frames to be used with particular verbs (the so-called "verb island" hypothesis). This 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.

Evaluating simple probabilistic grammars on sentence types rather than sentence tokens is an approximation to using more sophisticated multi-stage probabilistic grammars such as adaptor grammars (e.g., Goldwater et al., 2006; Johnson et al., 2007), which better capture the observed power-law frequencies of natural language and the processes giving rise to those frequencies. Presenting the model with sentence types does not mean I am presuming that human learners ignore token frequencies; in fact, the adaptor grammar framework explicitly provides a way to capture token-based frequency information, and does so more accurately than conventional probabilistic grammars do.

Do children actually evaluate candidate grammars primarily in terms of observed sentence types rather than tokens? There is little empirical work that bears directly on this issue, but the possibility is at least not inconsistent with what we know of language acquisition. It derives from a sensible model of language, which separates the component of the language faculty used to generate legal syntactic forms from the component of the faculty that determines the distribution of sentences. This separation is reasonable because there are many factors that might affect which sentences are spoken that have nothing to do with the grammar itself: the conversational context, lexical and syntactic priming effects, and salience in memory, to name just a few. A sensible learner should want to distinguish between these factors when evaluating potential grammars, or at least try to remove some of the variation that is due to language usage effects rather than the intrinsic structure of the grammar.

If people do have a disposition to evaluate grammars on the basis of sentence types, does this 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. Determining whether such a disposition exists – and, if so, whether it is language-specific or not – is a question for future work.

It is important to note that the main claim, that the hierarchical phrase structure of language need not be specified as part of the language faculty, is supported by the finding that the model prefers the hierarchical grammars on the basis of type-based

input. As noted earlier, as long as there is some reasonable way to set it up such that the ideal learner learns the right thing, that shows it is learnable; this argument is even more powerful when the framework adopted, based on adaptor grammars, is if anything a more accurate statistical model of natural language.

While the preference for linear grammars given token-based input does not affect my positive conclusion about hierarchical learnability with context-free grammars, it does provide further motivation to explore more linguistically sophisticated hierarchically structured grammars that move beyond a simple context-free formalism. It is known that CFGs are not adequate to capture the full syntax of language. It is possible that if one worked with more sophisticated, linguistically powerful hierarchical grammars, they might provide better fits to sentence token data. This work also yields predictions that might be testable experimentally: for instance, one could evaluate whether people learn a different grammar in an artificial language learning experiment based on whether the evaluate the data with respect to type or token-based information.

In any case, all of the assumptions made in this analysis involve either abilities that are plausibly domain-general, or language-specific knowledge that is distinct from the knowledge being debated: I critically did not assume that learners must know in advance that language specifically has hierarchical phrase structure. Rather, I showed that this knowledge can be acquired by an ideal learner equipped with sophisticated domain-general statistical inference mechanisms and a domain-general ability to represent hierarchical structure in sequences – a type of structure found in many domains outside of natural language. The model contains no *a priori* bias to prefer hierarchical grammars in language – a bias that classic PoS arguments have argued was necessary. The learned preference for hierarchical grammars is data-driven, and different data could have resulted in a different outcome. Indeed, we find different outcomes when we restrict attention to only part of the data.

Relevance to human language acquisition

What conclusions, if any, may we draw from this work about the nature of grammatical acquisition in human learners? This 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, this 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. 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 (Jaynes, 2003; Howson & Urbach, 1993). A more deductive or falsificationist approach (Popper, 1959) to scientific inference might underlie Chomsky's view: an objective neutral scientist should maintain belief in the simplest rule – e.g., the linear rule for auxiliary-fronting – until counterevidence is observed, and because such counterevidence is never observed in the auxiliary-fronting case, that scientist would incorrectly stay with the linear rule. But under the view that scientific discovery is a kind of inference to the best explanation – which is naturally captured in a Bayesian framework such as ours – the hierarchical rule could be preferred even without direct counterevidence eliminating the simpler alternative. This is particularly true when we consider the discovery problem as learning the grammar of a language as a whole, where the rules for parsing a particular kind of sentence (such as complex auxiliary-fronted interrogatives) may emerge as a byproduct of learning how to parse many other kinds of sentences. The rational Bayesian learning framework adopted here certainly bears more resemblance to the practice of actual linguists – who after all are mostly convinced that language does indeed have hierarchical structure! – than

does a falsificationist neutral scientist.

Defining the prior probability unavoidably requires making particular assumptions. A simplicity metric defined over a very different representation would probably yield different results, but this does not pose a problem for this analysis. The classic PoS argument claims that it would be impossible for a reasonable learner to learn a hierarchical rule like aux-fronting. All that is required to respond to such a claim is to demonstrate that such a reasonable learner *could* learn this. Indeed, this 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 the conclusion that some reasonable learner could learn that language has hierarchical phrase structure.

Another issue for cognitive plausibility is the question of scalability: the largest corpus presented to the model contains only 2336 sentence types, many less than the average human learner is exposed to in a lifetime. Since these 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. I 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 analysis are defined over the former rather than the latter. Secondly, as I have shown, additional (grammatical) sentence types are more likely to be already parsable by a context-free grammar than by a regular grammar. This means that the appearance of those types will actually *improve* the relative likelihood of the context-free grammar (because they will no longer constitute an overgeneralization) while not changing the prior probability at all; by contrast, the regular grammar may more often need to add productions in order to account for an additional sentence type, resulting in a lower prior probability and thus a lower relative posterior score.

If the knowledge that language has hierarchical phrase structure is not in fact

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

Although Chomsky's original formulation of the PoS argument focused on the innateness of the hierarchical structure of language, recent characterizations of an innate language faculty have concentrated on recursion in particular (Hauser, Chomsky, & Fitch, 2002). An interesting aspect of these results is that although all of the best context-free grammars I 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).[16] It is difficult to know how to interpret these results, but one possibility is that perhaps syntax, while fundamentally

[16]See Perfors, Tenenbaum, Gibson, and Regier (submitted) for a more detailed exploration of this issue.

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, this 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.

Conclusion

I 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 may be possible for human children to do so as well. Two important open questions remain: how well an ideal learnability analysis corresponds to the actual learning behavior of children, and how well this computational model approximates this ideal. These specific conclusions are therefore preliminary and may need to be revised as we begin to learn more about these two fundamental issues. Regardless, I have offered a positive and plausible "in principle" response to the classic negative "in principle" poverty-of-stimulus arguments for innate language-specific knowledge.

This work also suggests a new approach to classic questions of innateness. By working with sophisticated statistical inference mechanisms that can operate over structured representations of knowledge such as generative grammars, we can rigorously explore a relatively uncharted region of the theoretical landscape: the possibility that genuinely structured knowledge is genuinely learned, as opposed to the classic positions that focus on innate structured knowledge or learned unstructured knowl-

edge, where apparent structure is merely implicit.

Some general lessons can be drawn. It does not make sense to ask whether a specific generalization is based on innate knowledge when that generalization is part of a much larger system of knowledge that is acquired as a whole. Abstract organizational principles can be induced based on evidence from one part of the system and effectively transferred to constrain learning of other parts of the system, as we saw for the auxiliary-fronting rule. These principles may also be learned prior to more concrete generalizations, or may be learnable from much less data than is required to identify most of the specific rules in a complex system of knowledge. In the following two chapters, I will explore two other topics within language acquisition – learning verb argument constructions, and acquiring word learning biases – in an effort to evaluate whether, and to what extent, these lessons apply more generally.

Chapter 4

Word learning principles

In Chapter 1 I discussed the problem of Feature Relevance, with an emphasis on its implications for the problems of word learning and the determination of reference. We saw that it may be reasonable to assume that the problem of Feature Relevance can be addressed, at least in part, by assuming the existence of some innate higher-order learning constraint (or overhypothesis), but that this solution cannot apply everywhere. In some domains, overhypotheses themselves appear to be learned, as is plausibly true for the case of the shape bias in word learning. However, we still understand little about how learning on multiple levels of abstraction can be done. Does lower-level learning always occur before higher-level learning, as one might expect, or is it possible to acquire an overhypothesis faster the the specific hypotheses it is meant to constrain? If not, how can we explain the emergence of the shape bias? If so, what principles explain this acquisition? And what, if anything, does this imply about the inductive constraints that were previously assumed must be innate?

The acquisition of feature biases

Although learning on multiple levels of abstraction is a theme throughout this thesis, in this chapter I focus on it in more detail in the context of the acquisition of word

The work in this chapter was carried out in collaboration with Joshua Tenenbaum, and the model is an extension of Kemp et al. (2007).

learning biases. This work addresses several major issues that arise in that literature.

Category learning

In previous work with Charles Kemp (Kemp et al., 2007), we demonstrated that a simple hierarchical Bayesian model can qualitatively capture human performance based on simple datasets modelled on Smith et al. (2002) and Jones and Smith (2002). However, the input we gave to that model was far "cleaner" than that typically faced by the child; it had relatively few features, and there was more within-category coherence than might be typically found in the real world. Was the successful acquisition of a shape bias the result of these factors? How "messy" would the input have to be before second-order overhypothesis learning would be impossible?

Our previous work also assumed that the learner already knows which categories individual items are in. The actual learning task faced by children is of course far more complicated, since category information is not given directly: children must realize that individual exemplars of basketballs and baseballs are both members of the category *ball*, as well as that balls (like other count nouns) tend to be similar in shape but not other features. How do children solve this learning problem? One answer is that perhaps they are *effectively* given category information via word labels: if they believe that labels are clear indicators of category membership, they could use this information to cluster items in the world into categories and thereby learn the shape bias. A different possibility is that children could learn the categories themselves, without requiring labels, if they assume that categories consist of clusters of items with coherent features.

As we saw in Chapter 1, a core debate about how the shape bias emerges focuses on the nature of the higher-level assumptions necessary to explain that emergence: is it due to attentional/associationist learning (Smith et al., 2003; Colunga & Smith, 2004), or does it result from certain ontological commitments children make concerning the nature of kinds (Xu, 2002; Xu et al., 2004)? Central to this discussion is the question of what assumptions children make about word labels. According to the "dumb attentional mechanism" account, no special assumptions about words need

to be made in order to explain its emergence (Smith et al., 2003; Colunga & Smith, 2004, 2005; Smith & Samuelson, 2005). The shape bias is a bias about words, but only in the sense that what is learned is the correlation between the structure of early noun vocabularies and organization by shape. This correlation is argued to be sufficient for an associative learner to acquire the second-order biases that children do – biases saying that count nouns are correlated with shape, but in other domains different features are correlated with labels. Simulations on connectionist networks qualitatively replicate the emergence of feature biases, given artificial input whose statistics reflect that of the early noun vocabulary of children speaking English and Japanese (Samuelson, 2002; Colunga & Smith, 2005).

Other researchers propose that the emergence of the shape bias cannot be fully understood via simple associative mechanisms. Rather, as Waxman (2004) suggests, "infants embark on the task of word learning not as *tabulae rasae*, but equipped with a broad, universally shared expectation that links words to commonalities among objects." According to this idea, labels cannot be viewed simply as another feature that varies coherently with features of objects in the world: instead, from birth, children have an innate bias to believe that word labels pick out categories in a special way (e.g., Soja et al., 1991; Xu, 1999, 2002). Even pre-linguistic infants appear sensitive to information conveyed by language, individuating objects (Xu, 2002) and forming object categories (Balaban & Waxman, 1996; Fulkerson & Waxman, 2007) on the basis of linguistic labels but not other correlated information like tones.

These researchers agree that the statistical correlation between early noun vocabularies and objects in the world plays some role in the shape bias in word learning, but argue that one of the more crucial parts is that children have an *a priori* bias to believe that words are privileged cues to categories. Under this view, the shape bias is not specific to naming, but rather emerges because shape is a reliable cue to category/kind membership, which itself is correlated strongly with naming (Bloom, 2000; Diesendruck & Bloom, 2003; Xu et al., 2004). Children show the shape bias when asked to reason about categories, whether they are named or not (Diesendruck & Bloom, 2003). In addition, word generalization patterns are argued to depend on

conceptual as well as perceptual knowledge: when novel target objects are described as artifacts 18- to 24-month-olds extend names on the basis of shape, but when they are described as animates, they use both shape and texture (Booth, Waxman, & Huang, 2005; Booth & Waxman, 2006).

In sum, then, there remains much confusion about the role of linguistic input in the emergence of higher-order feature biases. Do children presume that labels are strong cues to category membership, or is the emergence of the shape bias explainable without this assumption? The connectionist networks presented by Samuelson (2002) and Colunga and Smith (2005) do not address this question, because there is no way within those models to differentiate between words-as-categories, words-as-features, and no words at all.

Learning from a few examples

An interesting study by Smith et al. (2002), often taken to be strong evidence for the distributional account of the shape bias, provides strong evidence that it can be taught. 17-month-old infants lacking the shape bias were trained for eight weeks on two items from four novel categories with novel names (e.g., *zup*) that were strongly organized by shape. At the end of this training, the infants' first-order and second-order generalization was tested by presenting them with either an exemplar from one of the trained categories (first-order) or an exemplar from a novel category (second-order) and asking them to choose another object that they think would share the same label. In both cases, children preferred to extend the objects by shape, demonstrating that they had acquired the shape bias at an age when non-trained children do not have one. Most intriguingly, the children's outside-of-the-lab vocabularies increased, suggesting that the acquisition of the shape bias helped them to acquire English words faster.

This is a surprising finding because aspects of it seem difficult to explain under either the distributional or the more nativist words-as-categories view. Although the latter view may be consistent with a learning account that can explain how a few additional items can lead to the emergence of this feature bias, no such account has

been articulated. Smith et al. (2002) interpret their results to be arguing for the distributional account. However, although the finding supports the notion that the shape bias is learned, it may be difficult to explain how an associationist learner could acquire it based on so little data. Associationist learning is generally much slower, and less prone to rapid shifts based on few data points. The children in the study already spoke many words; why did learning just a few more have such a large effect?

Study 1: Category learning

In this section I present an extension of the model described in Kemp et al. (2007), with the goal of exploring how first-order learning (about categories) and second-order learning (of feature biases) might emerge together. The model extension has the ability to cluster individual items into categories as well as infer second-order knowledge, or overhypotheses, about how categories in general are organized. It can also capture different assumptions about the role of words. This permits us to evaluate under what circumstances, if any, the different assumptions would give rise to differences in performance, at least in this ideal learning setting.

Previous work

In earlier work with Charles Kemp (Kemp et al., 2007), we presented a hierarchical Bayesian model for the acquisition of feature biases. The basic model assumes that data consists of observations of categories, where each observation corresponds to the number of times the category is observed with each feature. For instance, one of the datapoints corresponding to the model schema in Figure 4-1 would indicate that the *ball* category occurs five times with the shape feature *sphere*.

Within the framework captured by this model, if we would like to predict what sort of features we are liable to observe in some category n (i.e., level 1 knowledge about features), the first step is to ask how the level 1 knowledge might be acquired. The answer to this question will make reference to a more abstract body of knowledge (level 2 knowledge). In this case, level 2 knowledge would be knowledge about the

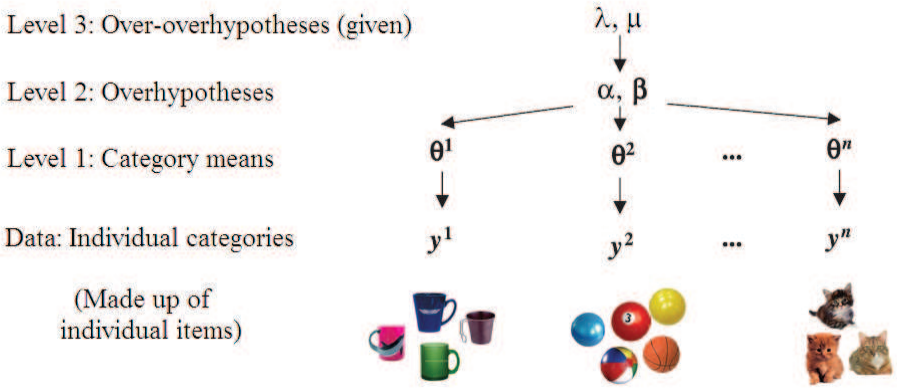


Figure 4-1: A hierarchical Bayesian model, Model L2. Each setting of $(\alpha, \boldsymbol{\beta})$ is an overhypothesis: $\boldsymbol{\beta}$ represents the feature distribution across all categories, and α represents the variability of each feature within each category. Data consists of category information corresponding to the features associated with each of the items in the category – for instance, the individual exemplars of cats, cups, and balls shown here. The hyper-parameters μ and λ are given; Level 1 and Level 2 knowledge are inferred by the model.

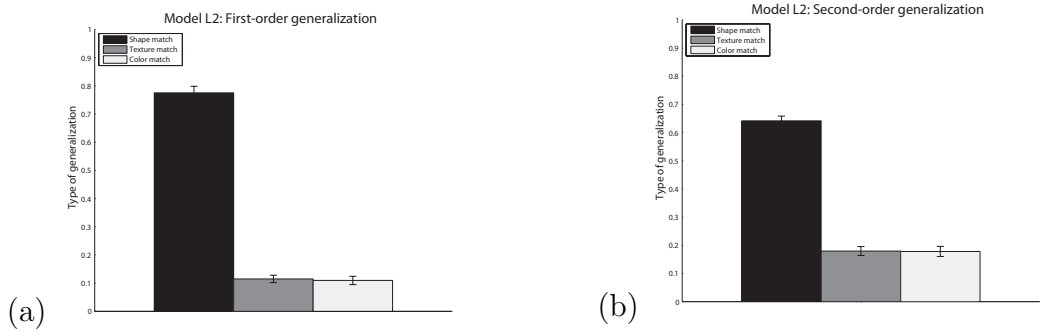
distribution of features across categories in the world, which can be represented by two parameters, α and $\boldsymbol{\beta}$. Roughly speaking, α captures the extent to which each individual category is organized by a given feature (or not), and $\boldsymbol{\beta}$ captures the average distribution of features across all categories in the world. For instance, low α would indicate that each item in a category tends to share a certain feature value, but does not say anything about *what* value that might be: if a category had low α for the shape feature, one would know that it was organized by shape, but not know precisely what shape it was.

As one might expect, level 2 knowledge depends on knowledge at a higher level, level 3, which is represented in this model by two (hyper-)parameters λ and μ. Because knowledge on higher levels grows increasingly abstract, it is difficult to translate level 3 prior knowledge in a clear and intuitive way – but, broadly speaking, λ and μ capture prior assumptions about α and $\boldsymbol{\beta}$, respectively. The λ parameter captures something about the range of values expected about the uniformity of features (would it be extremely unusual to see categories whose items all have exactly the same shape?), and the μ parameter the range of values of the expected distribution of features in the world.

Table 4.1: Data presented to the model, based on the Smith et al. (2002) study.

	Category 1		Category 2		Category 3		Category 4		
Feature	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Novel item
shape	1	1	2	2	3	3	4	4	5
texture	1	2	3	4	5	6	7	8	9
color	1	2	3	4	5	6	7	8	9
size	1	2	1	2	1	2	1	2	1

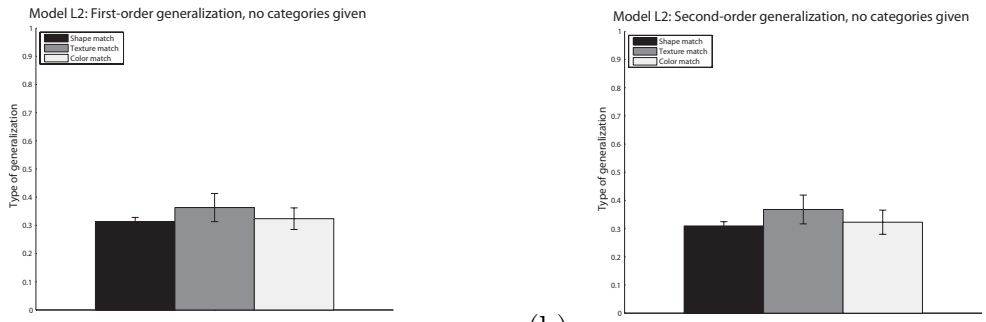
In Kemp et al. (2007), we presented this model with data modelled after the Smith et al. (2002) experiments. The data consisted of eight items divided into four categories and is described in detail in Table 4.1. The model, which I will call Model L2 because it performs inference on two levels of abstraction, forms correct first- and second-order generalizations on the basis of shape, as shown in Figure 4-2.



(a) (b)

Figure 4-2: Type of generalization predicted by Model L2 given input modelled after the Smith et. al. (2002) study. The model correctly learns that categories are organized by shape, rather than color or texture. (a) First-order generalization. (b) Second-order generalization.

However, this data assumes the individual items are already partitioned into categories. Is learning of the shape bias possible if the model does not know in advance what the categories are, if it is simply given eight individual items, but with no category information? Figure 4-3 shows the first- and second-order predicted generalization of Model L2 on this new input. It is apparent that if it is not given category information, Model L2 cannot learn the shape bias. This is, of course, exactly what we would expect, because the shape bias is a bias about what how features are organized within categories; without categories, the idea is meaningless.



(a) (b)

Figure 4-3: Type of generalization predicted by Model L2 given input modelled after the Smith et. al. (2002) study, but without category information. The model generalizes along each feature equally, since it lacks any categories along which to infer any bias.(a) First-order generalization. (b) Second-order generalization.

Extension A: Learning categories *and* overhypotheses

We cannot assume that children are necessarily shown how each individual item is clustered into categories. Yet they clearly do learn the shape bias (and other sorts of feature biases as well). How is this done? One possibility, as discussed earlier, would be that the labels children hear serve as strong cues to category membership, and that without labels they would be unable to form categories or make second-order generalizations about them. The other possibility we considered is that the categories could be learnable even without any associated word labels.[1]

I evaluate these possibilities by creating a category-learning extension of Model L2. This extension is denoted with the prefix "C" (e.g., Model C-L2). Whereas Model L2 assumed that the input data corresponded to the features associated with categories $\boldsymbol{y}$ (where $\boldsymbol{y^i}$ contains the counts of features corresponding to category i), Model C-L2 assumes that the input consists of the features corresponding to individual items j. It then tries to group the items j into the best categories $\boldsymbol{y^i}$. The model does not know how many categories there are; rather, it calculates the best partition of items into categories on the assumption that items with similar feature distributions are

[1]Of course, if a second-order bias were formed without any word information at all, it would not be a bias about word learning *per se*, but rather about category organization. Nevertheless, since this sort of bias has a clear relationship to the word learning biases we focus on in the literature, it is an interesting question to explore – especially since, if a bias could be learned based simply on the organization of coherent non-word features, it would not be difficult to exploit the same mechanism to handle additional word information.

more likely to be grouped together into a category. A prior favors smaller numbers of categories, but this prior can be overcome if the items are different enough from one another. The technical details of this model are described in the appendix.

Presenting Model C-L2 the same data as in Table 4.1, but without category information included, results in performance that is qualitatively quite different to Model L2 given the same data, and similar to that of Model L2 on input in which category information is included. Figure 4-4 shows model performance for first and second-level generalization, as well as a depiction of the categories learned. The model is able to group the individual items into categories based simply on feature coherence: the shape features are more coherent than texture or color, and size, while coherent itself, results in categories with overall more variation, so clustering by size is dispreferred.

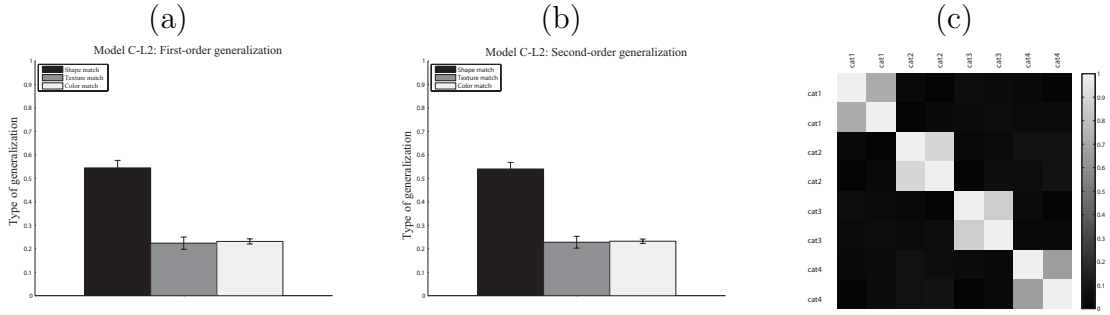


Figure 4-4: (a) Type of first-order generalization predicted by Model C-L2 given input corresponding to the data from Smith et. al. (2002), but not pre-classified into categories. (b) Second-order generalization. (c) The model learns the correct categorization of the objects, forming four categories.

These results appear to suggest that knowing the categories may not be necessary in order to acquire the shape bias, and in fact may not even make it much easier. However, such a conclusion would be rather hasty. This input is much cleaner (in terms of both the number and the coherence of features) than what is faced by children learning natural language. Thus, even if knowing the categories in advance might not be necessary when faced with this sort of input, it might still be necessary for a more realistic situation. What we really need is an understanding in the abstract of how knowing first-order information about category assignments helps an ideal learner make second-order generalizations about feature biases across categories. It is

unclear how much we can generalize results based on relatively small, clean datasets to the complexity of the situation faced by the child learner in real life.

Extension B: A systematic exploration of category learning

One approach to this problem would be to present the model with realistic, child-directed data, as we do in Chapters 3 and 5. But while it is possible to at least approximate the sort of input children receive when the data in question is about the syntactic forms they hear (thanks to the existence of corpora), it is a much different matter to know how to approximate the kind of data that would be relevant here. How many items, and of what categories, does a typical child encounter in his daily life? Even if we could roughly estimate this based on the first words in children's vocabularies, what features do we assume those items have? Perhaps we could estimate these by having adults list the features they associate with those items, but – in addition to problems typical to feature-listing methodologies, like overlooking extremely common or basic features (e.g., "breathes", "has cells") – doing so puts us in danger of assuming precisely what we wish to study. If, as some have argued (e.g., Goldstone, Steyvers, Spencer-Smith, & Kersten, 2000; Smith, 2005), the process of category learning affects our perception of features, then adult ratings might not accurately reflect children's input at all.

Another option would be to evaluate how factors like category coherence or number of items and categories in a dataset of dataset affect overhypothesis learning. Previously the model was presented with datasets in which the categories were completely coherent with respect to shape: knowing the shape of an object served as a 100% valid predictor of which category that object was in. It may be that if categories were less coherent, then knowing which items were in which categories would be more useful for learning overhypotheses about how they are organized.

To test this, I create a series of datasets that vary systematically by coherence. As before, there are eight objects from four categories, but the complexity of the dataset is increased somewhat by adding additional features, ten of which are arbitrarily called shape, and ten other. The other features always have 0% coherence, meaning

Table 4.2: Sample artificial datasets, varying in coherence. For illustrative purposes and to make the difference between different coherence levels most clear, these sample dataset contain two categories of five items each. The actual sample datasets first presented to the model contain eight items corresponding to four categories; later we systematically vary the number of items and categories between one and sixteen.

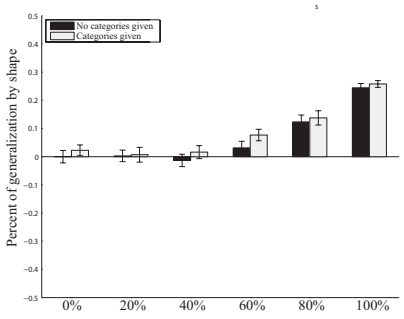
	Category 1					Category 2					
Feature	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Novel item
	100% Coherence										
shape_1	1	1	1	1	1	3	3	3	3	3	5
shape_2	2	2	2	2	2	4	4	4	4	4	5
shape_n	3	3	3	3	3	2	2	2	2	2	5
other_1	1	4	3	1	2	4	1	2	3	2	5
other_2	4	3	4	2	1	1	3	4	2	3	5
other_n	2	1	4	3	4	2	4	3	1	1	5
	60% Coherence										
shape_1	1	2	1	3	1	3	4	2	3	3	5
shape_2	1	4	2	2	2	4	1	4	4	3	5
shape_n	3	3	4	3	3	1	2	2	2	4	5
other_1	1	4	3	1	2	4	1	2	3	2	5
other_2	4	3	4	2	1	1	3	4	2	3	5
other_n	2	1	4	3	4	2	4	3	1	1	5

that they vary completely randomly with respect to category membership. The shape features are generated by the following procedure. First, we set them to have feature values that are 100% coherent with respect to the category, meaning they are perfect predictors of category membership. A coherence level of c is generated according to the definition that a feature value has a $100 - c\%$ chance of being random. For instance, if a feature (say, shape_1) is 60% coherent, then each item in some category m would, with 40% probability, have a shape_1 that does not match the defining shape_1 of category m. To illustrate what is meant by coherence, sample data is shown in Table 4.2.

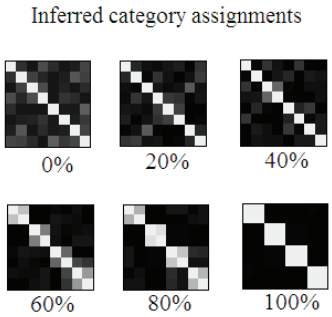
I create datasets that vary systematically in the coherence of their shape features. Dataset coherence ranges from 0% to 100% in increments of 20%. As before, second-order generalization is tested with an additional object with novel values along each feature: is the model more likely to assume that additional items in the same category will share the same shape features, or the same other features?

At each level of coherence, I compare the generalization between datasets in which the categories are given and datasets in which the model must infer the categories itself. Figure 4-5(a) shows the results. Somewhat surprisingly, there is no significant

difference in second-order generalization depending on whether the categories were given to the model or not. As Figure 4-5(b) reveals, if coherence is low enough it does not support category learning; but it would not be sufficient to support a second-order generalization even if the categories were known. In other words, once the features are coherent enough to support second-order generalizations, they are coherent enough to support categorization. Thus – at least for an optimal learner – being given category information appears to make little difference as to whether second-order overhypotheses are acquired.



(a)



(b)

Figure 4-5: (a) Second-order generalization by shape features on simulated datasets that vary in how coherent the shape features are (shown on the x-axis). Second-generalization, shown on the y-axis, is similar regardless of whether the model is given the category information or not. For low values of coherence there is no shape bias in both cases; but if coherence is high enough to infer the existence of a feature bias, it is high enough to infer category membership. (b) Category assignments inferred by Model C-L2. Consistent with the (a), recognizable categories only emerge when coherence reaches 60%.

Although this is interesting, it is still somewhat unclear how much this result depends on other aspects of the dataset, like the number of items or underlying categories. I therefore systematically vary all possible combinations of the number of items in the dataset (among 16, 8, 4, and 2 items) and the number of categories (among 8, 4, 2, and 1). Results are shown in Figure 4-6. There is once again no systematic difference in model performance depending on whether the categories are given or not. If there is sufficient coherence to support overhypothesis learning, there is sufficient coherence to support categorization. The other general patterns are all sensible: second-order generalization is better if category coherence is higher or there are more items in the dataset.

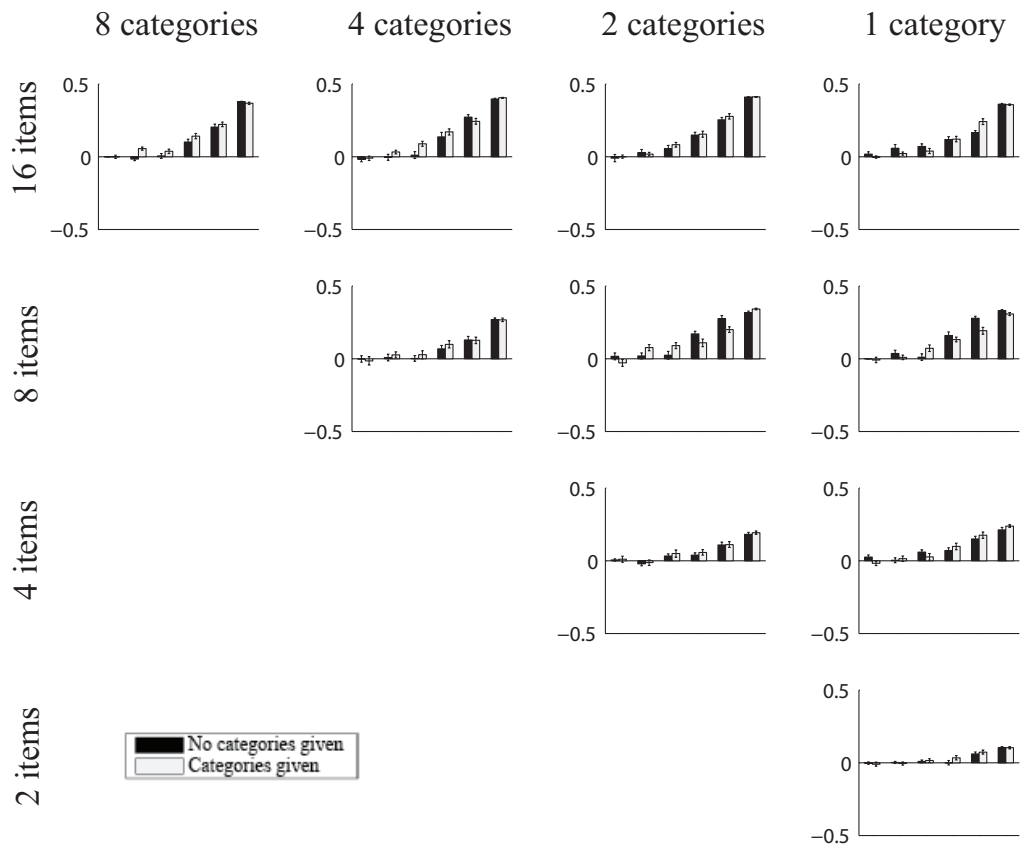


Figure 4-6: Second-order generalization by shape features on simulated datasets that vary in the number of items in the dataset (vertical) and the number of categories (horizonal). Within each subplot, the y axis indicates the degree of second-order generalization by shape features, and the x axis depicts coherence varying from 0% to 100%. As before, there is little systematic difference between second-order generalization by whether the categories were given to the model or not.

If there is no systematic effect of from making category information accessible, one implication might be that the assumptions children make about words should have no effect either. Whether words are just another feature or a strong cue to category membership, they carry more category information than was given to the model in the "no categories" case, but it performed equally well regardless.

Extension C: The role of words

The simulations so far include two conditions, one in which no category information at all is given and one in which category membership is known for all items. The latter condition is equivalent to assuming that labels are perfectly indicative of category membership and all items are labelled. To these two conditions, I add a condition in which labels are simply additional features (and, again, all items are labelled). The label information is there, but labels are not taken to be a strong indicator of category membership; it is thus more consistent with the distributional learning account of the shape bias.

Because it is unlikely that children hear labels for every item in their world, I also add two conditions where only 50% of the items are labelled.[2] I refer to these conditions as *semi-supervised*, and there is a semi-supervised version corresponding to each of the assumptions about words – whether words are taken as strong cues of category membership, or whether they are simply features in the model.

Results are shown in Figure 4-7. As expected, there is no significant difference in model performance regardless of what assumptions were made about words, whether all the items were labelled or only some were.

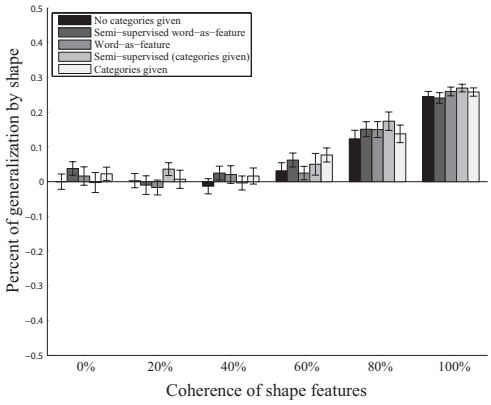


Figure 4-7: Second-order generalization by shape features on simulated datasets that vary in coherence. Second-order generalization is unaffected by the nature of the assumption made about words.

[2]I also evaluated conditions where 25% and 75% of the items were labelled, respectively; there is no qualitative difference between any of these conditions so here I only report on the 50% condition.

Is there any circumstance in which that the different theories about the role of words might result in different predictions? One possibility could be be if words do not correlate with the other features that pick out categories. Under the strongest "words are cues to categories" claim, one should form categories based on the labels even if the other features are consisted with a different category organization. But if words are features, then if the words are no longer correlated with the other features, one should learn to disregard words when forming categories. Under such an account, the shape bias would emerge, but would no longer be word-driven.

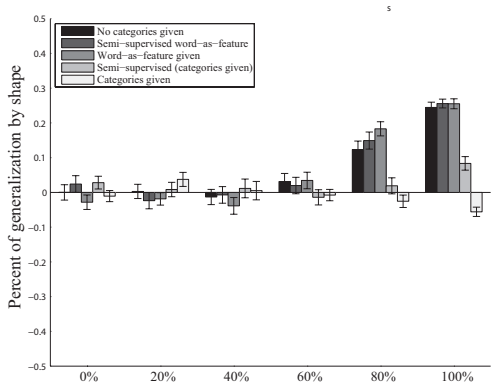


Figure 4-8: Second-order generalization by shape features on simulated datasets that vary in coherence. Here, the coherence of the shape features contradict the organization of the categories (given by word labels). If words are treated as features, rather than taken as strong cues of category membership, the model learns that categories are organized by shape, and that word labels should be ignored. But if words are taken as strong cues of category membership, the model can learn no such thing (since the categories picked out by the words now *aren't* organized by shape), and therefore does not acquire a shape bias.

Indeed, that is what we find. I present the model with the same dataset, except that the labels pick out categories opposite to those formed by the coherent correlation of the other (shape) features. Figure 4-8 shows that, as expected, if words are assumed to pick out categories then no shape bias emerges – the "categories" are no longer organized by shape. If words are assumed to be features, there is evidence of second-order generalization by shape, but this generalization is no longer word-based at all; the model has learned that the word features should be disregarded.

It is somewhat unclear how this result bears on the shape bias debate. A situation

in which naming patterns completely go against all other visible features of a category certainly seems quite unnatural, and it is possible to come up with explanations under both accounts that capture this behavior. The distributional learning account might simply point out that if words don't vary coherently with categories, then a shape bias in word learning should not be acquired. This might predict that a human presented with data like that shown to the model would learn to disregard labels in categorizing and forming second-order generalization. One future direction would be to perform such an experiment, although the situation is so unnatural it may still be difficult to determine how it applies to children's word learning.

Only an extreme version of the "words as categories" account would be unable to explain the model results, and few would argued that all words are assumed to pick out categories no matter what. Even if children begin with the expectation that words pick out categories in general, this expectation is clearly defeasible in some contexts. For instance, a study by Davidson and Gelman (1990), used to argue for a more nativist account of word learning, demonstrates that children will only generalize based on novel labels if the labels have some grounding in object appearance. This is a somewhat similar circumstance to the results reported here, but it is difficult to describe what is going on in terms that yield clear differences or predictions within this framework. Do children in the Davidson and Gelman (1990) study decide to disregard labels because they view words as features, or do they assume that while words pick out categories in general, there are exceptions? If so, how do children determine when this is? And how does this explanation differ predictively from the distributional account? It begins to look as if the difference between the two perspectives may be a difference that *makes* no difference, at least in terms of the implications for second-order generalization.

Study 2: Learning from a few examples

Another question inspired by the extensive literature on the acquisition of the shape bias is how to account for the rapid learning reported in the experiment by Smith et

al. (2002), where teaching 17-month-old children four new words that were cleanly organized by shape was sufficient to lead to the presence of a shape bias long before one typically emerges. Why did only a few examples provoke such a large change? This seems like the sort of question that a purely distributional, bottom-up associationist account is ill-equipped to handle, since associationist learning is generally slower and more gradual. Indeed, although connectionist models have been shown to capture many of the empirical phenomena in the literature about the shape bias (e.g., Samuelson, 2002; Colunga & Smith, 2005), I am not aware of any that explain this finding in particular.

One possibility is that category learning is useful because it enables children to create "shape caricatures" or abstractions, resulting in improved object recognition and rate of lexical acquisition (Smith, 2005). Evidence for this is that children between 17 and 25 months of age begin to be able to recognize stylized caricatures of objects (both known and new), and increases in this ability are related to higher vocabulary size. This suggests that language learning – in particular, learning the object names that identify categories – is what drives these changes in shape perception. According to this hypothesis, the Smith et al. (2002) experiment was so successful because the new items were so simple and so cleanly organized that they enabled children to form a more abstract notion of shape – and it is this that formed the basis of the second-order generalization about shape.

Extension A: Adding new items to simulated datasets

Our model provides a way to begin to evaluate whether this sort of abstract feature learning is necessary in order to explain the Smith et al. (2002) results. Unlike more associationist or bottom-up methods, hierarchical Bayesian learning can quickly form abstract generalizations on the basis of only a few datapoints. Can a model capable of this sort of learning – but lacking in an ability to form abstract features or "shape caricatures" – account for the type of rapid learning found in the Smith et al. (2002) experiment?

I investigate this possibility by adding two additional items to the datasets in

Extension B of Study 1. These datasets vary in coherence as well as the number of items (between two and sixteen) and categories (between one and eight). The two additional items, meant to play the same role as the new words in the Smith et al. (2002) experiment, are 100% coherent with respect to the shape features and 0% coherent with respect to the other features. Is adding them sufficient to create a shape bias? As a control condition, I create a series of datasets whose two additional items have the same average coherence as every other item in that dataset. This condition ensures that any observed shape bias does not emerge simply because of the addition of two objects, regardless of their nature.

Figure 4-9 shows the difference in the emergent shape bias in the "clean feature" condition as compared to the control condition, at all levels of coherence except 100%, where there is no difference between the conditions. In general, if the shape features are already highly coherent, there is relatively little effect of adding a few additional items, since the shape bias has already been learned. (This would be equivalent to running the Smith et al. (2002) training study on children who already had the shape bias). A shape bias emerges, however, on datasets whose other objects are otherwise not coherent enough to support making a generalization by shape. The simple addition of a few highly coherent objects to these datasets is enough for the model to abstract this as a general bias, just as children do.

Extension B: Mimicking children's vocabulary

Another possibly more relevant way to evaluate this question would be with a dataset that approximates the input children have received by the time they are taught the additional items. This is difficult to do accurately for all of the reasons discussed at the beginning of Study 1: what items have children seen? What are the features they pay attention to? How are the items organized within categories, and which categories do children know about? Though the problem of assigning features in particular is an extremely difficult one, it may be interesting nevertheless to construct a dataset that approximates children's reality as well as we know how. Work by Samuelson and Smith (1999) and Smith et al. (2002) is useful here, since they report on typical

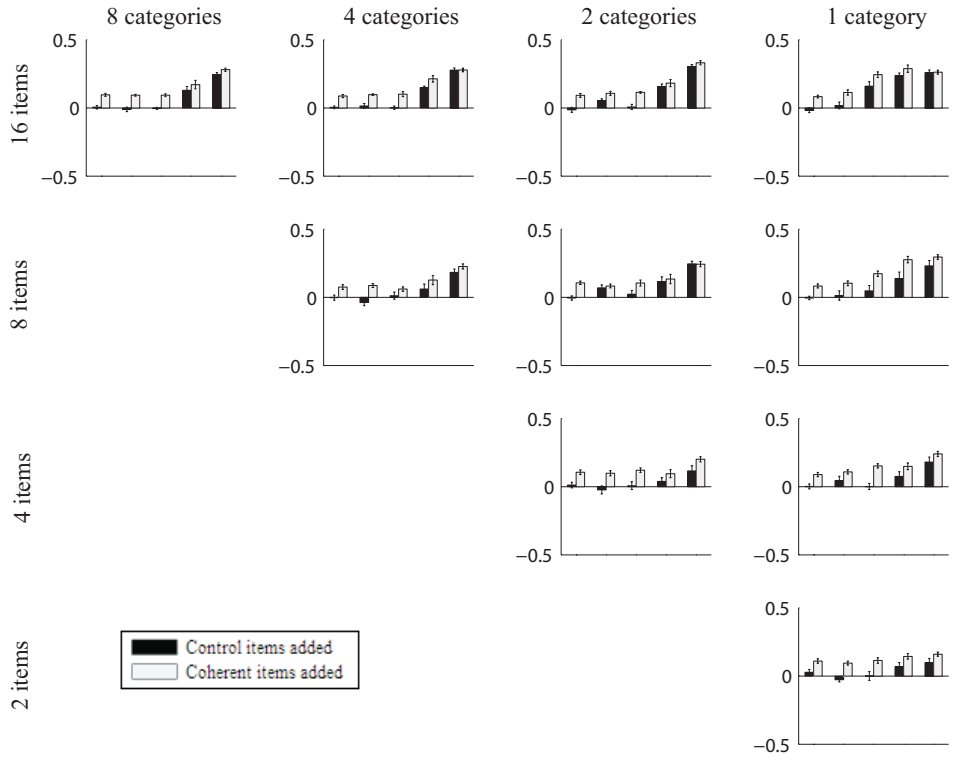


Figure 4-9: Role of adding two objects with highly coherent shape features, relative to a control condition. If the dataset is already highly coherent by shape, there is little additional benefit to adding the new objects, since it already supported a shape bias. But if the dataset is not already highly coherent, then these additional objects are sufficient for the model to infer a higher-level shape bias across the entire dataset.

vocabulary organization at different ages, as well as the average vocabulary size for the children in their study.

Based on this work, I therefore create a dataset meant to more closely capture children's actual input when they began the Smith et al. (2002) experiment. According to Smith et al. (2002), the average count noun vocabulary size for the 17-month-old children at the beginning of training was 16 items. Assuming that the organization of their vocabulary parallels that reported in Samuelson and Smith (1999), I estimate that 9 of the 16 items are organized by shape, 6 are ambiguous between being organized by shape or material, and 1 is organized by material. It is unclear how to estimate the number of exemplars children have for each of these categories, what de-

gree of feature coherence corresponds to "being organized by" shape, or what features children have. As a rough guess, I assume that the number of exemplars per category follows a Zipfian distribution. I also define three features in the dataset, corresponding to shape, color, and material, chosen because they are the three features reported on in the Samuelson and Smith (1999) study). Each feature has a sufficient number of values for each object to be entirely defined by that feature.[3] This is of course far simplified from the task facing children, but it may do as a first pass. I also assume that if an item is organized by shape or material, that means that it is 15% coherent in that feature and 0% coherent in the others; if it is ambiguous between any, then it is 10% coherent in both.[4]

Figure 4-10 compares model performance on this dataset with performance on the same dataset with four additional categories of two items each. As in the Smith et al. (2002) study, these categories are strongly organized by shape, which I implemented by setting their coherence on the shape feature to 100% and coherence on the other features to 0%. When given the dataset without the additional items, the model does not show a shape bias; but when those items are added, one emerges.

Discussion

The problem of Feature Relevance, introduced in Chapter 1, may be addressed at least in part by learning that can operate over multiple levels of abstraction. We saw this in a different domain in Chapter 3, where I showed that it might be possible to form abstract inferences about syntactic structure even before lower-level knowledge about the correct specific grammar is complete. The acquisition of the shape bias is a phenomenon for which there is also strong reason to believe that a higher-order inference is learned in time to constrain later generalizations. Previous work I did

[3]This is equivalent to assuming there are features such as, e.g., "cup-shaped" or "banana-shaped", rather than simply more generic features like "round."

[4]This coherence is quite low, but these are the values I had to choose in order that the model did not show a shape bias even without seeing the extra items. If there were more than one shape, one color, and one material feature, then individual items would be picked out by a fuzzy combination of many of the features of each type, and their coherence could be correspondingly greater without inducing a shape bias.

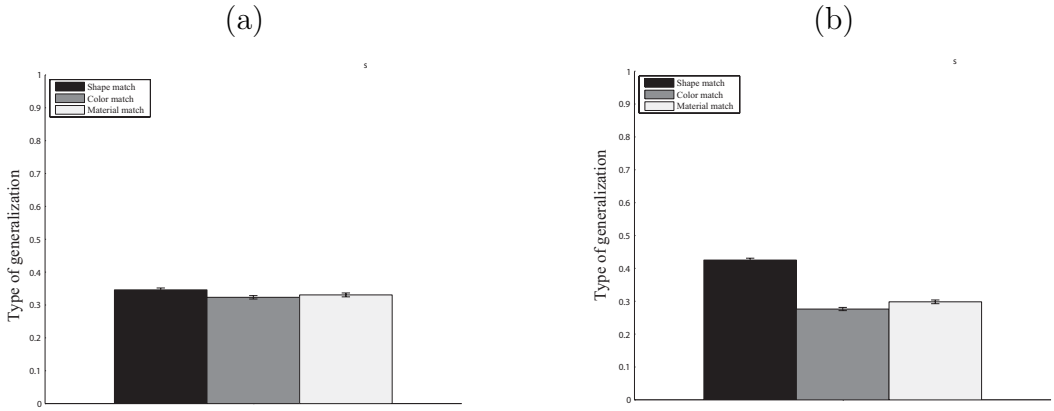


Figure 4-10: (a) Second-order generalization by shape on simulated dataset modelled after children's vocabulary at the beginning of the Smith et al. (2002) training study. The model, like children, does not show a shape bias. (b) Second-order generalization on the same dataset with four additional categories added, each strongly organized by shape. The model now shows a substantial shape bias.

with Charles Kemp (Kemp et al., 2007) presented a hierarchical Bayesian model for the acquisition of the shape bias, which was a valuable first step, but it assumed that children are simply given category information – clearly a simplification from the complicated, noisy world that human learners are faced with.

In this chapter I presented an extended version of the Kemp et al. (2007) model which is capable of clustering individual items into categories and inferring higher-order generalizations about those categories at the same time. This work demonstrates that learning overhypotheses like the shape bias is possible at the same time as category learning. The extended model, C-L2, provides a basis for exploring one of the oft-debated questions in the literature surrounding the acquisition of the shape bias – the assumptions that children make about word labels. Simulation results suggest that at least in this ideal learner setting, there is no significant difference between second-order generalization based on the presence or type of label information. This work also allows us to explore how and why adding only a few additional items can lead to a sudden emergence of a general higher-order bias.

In the following section I consider in more detail the implications and limitations of these findings. What do they tell us about the shape bias in language, and how do they bear on the learnability problems that motivated this thesis?

Words and categories

Perhaps the most surprising result in this chapter is the finding that there is no significant difference between second-order generalization based on whether category information is given to the model or not. This is true when the input is simple and abstract, as with the dataset based on the Smith et al. (2002) study, but it remains true for a variety of datasets that systematically vary in terms of coherence, number of items, and number of categories. Although the analysis could always be extended to cover datasets with more items or different numbers of features, the finding is robust to all of the variations I explored. From a purely computational perspective, this result is sensible: it reflects the insight that to the same extent that some input supports categorization, it supports forming second-order generalizations about the features along which categorization occurred. As a result, being given the category information makes little difference – if the categories are not strongly organized by a feature like shape, no second-order generalization about shape is learned even when the categories are known. Knowing the categories only leads to second-order generalization if there is enough organization to those categories to support such a generalization; but if that organization exists, an ideal learner will be able to infer the categories without being told what they are. Figure 4-11 demonstrates this graphically, with the goal of making it more intuitively clear how this is possible. As the category structure grows increasingly incoherent, the strength of first-order inferences (i.e., how each category is defined, e.g., "black dots") and second-order inferences (i.e., the generalization that categories tend to be long and thin) appear to vary in tandem.

The fact that this model is an ideal learner may play a critical role. It is possible, even likely, that the cognitive limitations of a non-ideal learner could make word labels far more important. Even if so, this analysis is still valuable. It points out that if word labels have a special status, that status is not due to the *informational* content they impart. Even the strong assumption that words pick out categories does not give the model appreciably more information on which to infer a shape bias. Informationally, a second-order feature bias is supported to the same degree that first-

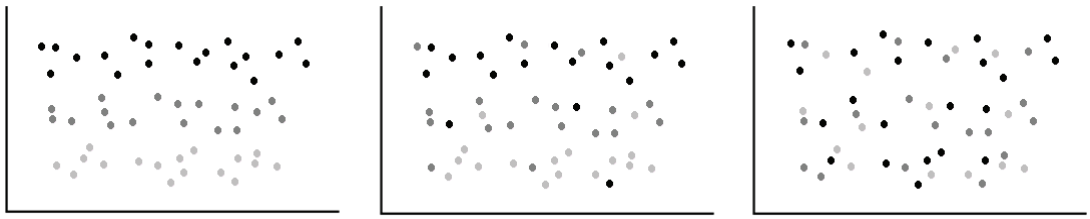


Figure 4-11: Graphical depiction of category structure as a function of coherence. The individual items, depicted via shaded dots, are 100% coherent with respect to the categories in the graph on the far left. It is easy to make first-order inferences (e.g., that the top category is shaped like a long thin rectangle, and is uniformly black) as well as second-order inferences (e.g., that all categories tend to be long thin rectangles and uniform in color). The middle graph shows categories with lower coherence, but that are still clear enough for both first- and second-order generalizations to be possible. In the final graph, coherence is low enough that both first- and second-order inferences are difficult to make. This graphically illustrates the degree to which both types of inference emerge in tandem.

order information about categorization is supported; knowing that category structure therefore does not provide significant information about second-order feature biases.

This is a departure from some arguments in the literature. The idea that children assume that words are privileged cues to category membership – while supported empirically by a variety of research (e.g., Xu, 1999, 2002; Fulkerson & Waxman, 2007; Bloom, 2000) – is sometimes justified theoretically as resulting from the additional information such an assumption would provide. Waxman (2004) suggests that children's initial expectations about words "support the formation of a stable repertoire of concepts" and help to establish reference. Xu (2002) suggests that a strong assumption that words map onto kinds "may be a mechanism by which infants first establish what kinds of things are in their environment and how these object kind concepts can play a role in object individuation." The results in this chapter suggest that, *contra* these ideas, the special status of words may not be due to the information they carry about category. Although assuming that words pick out categories might help even an ideal learner with first-order category learning, these results demonstrate that a strong assumption about words does not help with generalization beyond familiar items.

It is possible (and in some ways easiest) in interpret these results as consistent

with the distributional account of the shape bias, which suggests that words are simply another feature, and that learning is based on the coherent covariation of items that share a category. But it is also possible to interpret the results in ways that are consistent with the notion that word labels have some sort of special status as long as the learner is non-ideal in some way. Perhaps words serve as a memory aid, either because items that are labelled are easier to remember or their features are (e.g., Needham & Baillargeon, 2000). Maybe distinct labels highlight the differences between items, making it easier to identify which of many features matter – or, conversely, maybe different labels attenuate the differences between items (Robinson, Timbrook, & Sloutsky, 2006). Perhaps children pay special attention to information imparted linguistically or communicatively. Maybe perceptual biases arising from familiarity or other aspects of visual and auditory processing affect how children perceive and attend to word labels (Sloutsky & Robinson, 2008). Any of these possibilities are consistent with the findings in this chapter, which assume an optimal learner.

There is another option: perhaps word labels do not have an effect here only because the categories they pick out here are relatively impoverished. If the categories were associated with many, many other features, then solving the problem of Feature Relevance would become much more difficult. In that case, words may indeed contribute informationally, by picking out which features matter for categorization. More generally, in any circumstance where the relevant features are significantly outweighed by other features that coherently covary in another direction (or are simply random) then some external clue about which features are relevant may become necessary: words could serve such a purpose. This idea, which I present here in a rather qualitative and impressionistic sense only, may be susceptible to quantitative analysis in which the need for an external clue to Feature Relevance is a function of the number of features and their degree of covariation. This is an interesting avenue of future research.

The work presented in this chapter provides a new way to evaluate these possibilities. A thorough exploration would involve two main prongs of research: (1) experimentally testing to what extent the predictions of the model capture human

performance; and (2) adjusting the model to make it non-ideal in interesting ways. By yoking experiments with model predictions, we can determine precisely to what extent people use category information to form second-order generalizations. For instance, adults (or older children) could be presented with novel items corresponding to the input given to the model in Study 1. Does giving subjects category information or word labels improve their ability to form second-order generalizations? Do adults differ from optimal performance in any systematic ways? Experimental results along these lines would dovetail nicely with modelling work investigating different ways of making the model non-optimal. This general research program offers a valuable way to ask and test increasingly subtle questions about the role of words in category learning and the acquisition of the shape bias.

Feature learning and new examples

This work appears to address the finding of Smith et al. (2002), which showed that presenting 17-month-old children with just a few examples of items cleanly organized by shape was sufficient to lead to the presence of a shape bias (and increased vocabulary growth) long before one typically emerges. When the model is presented with a dataset that by itself does not support a shape bias plus two items strongly organized by shape, it shows increased generalization by shape relative to a control condition. Because the model is Bayesian, it is capable of making relatively strong inferences on the basis of relatively few datapoints. As discussed in Chapter 2, this is a byproduct of measuring likelihood with respect to the size of the hypothesis space. Since each datapoint (item) is presumed to be independent, the likelihood of seeing n items increases exponentially in the size of the space. Thus, seeing only two items that are strongly organized by shape is sufficient to conclude that there a shape bias of some sort.

A limitation of this finding, however, is that the model also forms a shape bias even when the other items in the dataset are not organized by shape at all. This does not appear to be a very sensible inference; wouldn't a rational learner conclude, instead, that the two new items simply follow a different pattern than the other data?

An extension of the model that can learn on the level of categories (as does Model C-L2) as well as kinds (as does the model in Kemp et al. (2007)) may be more appropriate for this sort of data. One would predict that if the other items in the dataset were different enough from the novel exemplars, such a model would, indeed, simply assume that the novel exemplars are in a separate "kind" and do not bear on the others. But if they were not too different, only a bit more well-organized by shape, there might be a version of the model that *does* learn the shape bias on the basis of only those. Further adventures in modelling will explore this possibility.

The results from Extension A of Study 2 may appear to suggest that a few new examples may have such a large effect simply because they are highly organized by shape, not because their features are in any way more stylized or abstract, as suggested by Smith (2005). However, although the initial results of the model are consistent with this suggestion, it is not yet well-supported. As before, a potential limitation is that our model is an ideal learner: it may be that a learner who was non-optimal in certain ways would not be able to form a rapid second-order generalization without inferring abstract features of some sort. We can explore this possibility by adjusting the model to have performance limitations of various sorts. Another issue is that the two new objects were well-organized on all of the shape features; a more realistic approximation might have items that are only well-organized on some of the shape features. Much work therefore remains to be done here, but there are several intriguing possibilities

We should also be extremely wary of generalizing from the model performance, especially on the dataset designed to capture children's actual category structure. I had to make many approximations and guesses in designing that dataset. It is unclear how many exemplars children have seen in the categories that correspond to their count noun vocabulary at this age. The actual features that children attend to are probably not simply shape, material, and color. Indeed, in order to properly test the hypothesis proposed by Smith (2005), we would have to incorporate multiple features of each kind, such that categories are only fuzzily organized around them.

Regardless of the results of that exploration, the suggestion of Smith (2005) yields

an intriguing challenge for modelers: how might a learner be capable of making inferences about abstract features at the same time as forming categories and acquiring higher-order learning biases? Is it possible to infer abstract features simply on the basis of how features co-occur with each other within categories, or is more information (say, features of features) necessary? Does forming an abstract feature or "shape caricature" help in generalization and inference, even if it may not be strictly necessary? Future work will address these questions and more.

Conclusion

In this chapter I presented a hierarchical Bayesian model capable of clustering individual items into categories and inferring higher-order generalizations about those categories at the same time. This work demonstrates that learning overhypotheses like the shape bias is possible at the same time as category learning. At least in this ideal learner setting, there is no significant difference between second-order generalization based on the presence or type of label information. We also began to explore how and why adding only a few additional items can lead to a sudden emergence of a general higher-order bias, as found in work by Smith et al. (2002).

On a broader level, I have used a Bayesian modelling framework to explore some classic issues of learnability in another domain. Although this work does not show how to solve the Feature Relevance problem, since there were always a finite number of features, it does provide one way to address it. Instead of assuming the innateness of any observed higher-order constraints about which features are relevant, it may be possible to learn such constraints. And although even higher-order assumptions (like over-overhypotheses) of some sort must be necessary in order to explain this learning, one take-home message of hierarchical Bayesian models is that the mere presence of a higher-order constraint cannot be reason for assuming that *that* constraint is innate. Our work suggests, in fact, that learning on multiple levels of abstraction can occur simultaneously, and that receiving additional information on a lower level may not speed up acquisition at a higher level.

In the next chapter I extend this model to a very different phenomenon: the acqui-

sition of verb constructions. Whereas here the focus was on exploring the dynamics of learning on multiple levels of abstraction, the next chapter centers more on the No Negative Evidence problem. As always, however, threads of all of the learnability problems and common themes about how to approach them will emerge. Finally, in the last chapter I will explore these threads and themes in richer detail.

Chapter 5

The acquisition of verb argument constructions

Chapter 1 introduced the learnability problems raised by the apparent lack of negative evidence in children's input, and we saw in Chapter 3 how a Bayesian learner might learn to avoid grammars that are too overgeneral, without eliminating overgeneralization entirely. In this chapter I address the learnability problems raised in Chapter 1, in particular the No Negative Evidence problem, in the context of a specific, well-studied example: the acquisition of verb argument constructions. I extend the simple hierarchical Bayesian model introduced in Chapter 4 to apply in this new domain, and show that its behavior qualitatively matches children's in several important respects. This framework suggests how the negative evidence problem may be solvable, and I finish by considering the assumptions and details built into the model that lead to this result.

Baker's Paradox and the puzzle of verb learning

Negative evidence – information about which constructions in a language are ungrammatical – appears to be largely missing in naturalistic speech to children (e.g., Brown

The work in this chapter was carried out in collaboration with Charles Kemp, Elizabeth Wonnacott, and Joshua Tenenbaum, and the model an is extension of Kemp et al. (2007).

Table 5.1: Constructions in the dative alternation.

Construction name	Abbreviation	Abstract form	Example
Prepositional dative	PD	NP_1 V NP_2 to NP_3	Debbie gave a pretzel to Dean.
Double object dative	DOD	NP_1 V NP_3 NP_2	Debbie gave Dean a pretzel.

& Hanlon, 1970), which poses a learning problem: without knowing how *not* to generalize beyond the language heard so far, a learner can only be certain that he isn't speaking ungrammatically if he is perfectly conservative, never saying anything he hasn't already heard. This poses obvious difficulties for scientists seeking to explain human language acquisition, because it is quite clear that in many ways, children are not conservative learners in this sense (e.g., Pinker, 1989). How, then, do they learn language, realizing not only what constructions are grammatical, but which ones are *un*grammatical as well?

A classic and well-studied example of this puzzle can be found in the study of the acquisition of verb argument constructions. In every language, different verbs take arguments in distinct constructions; for instance, the verb *give* is associated with a construction requiring the verb to have two object arguments in addition to the subject ("Debbie gave Dean the pretzel"). Not all verbs can occur in all constructions: "Debbie gave Dean" is ungrammatical, and not a sentence one would expect out of the mouth of a native adult English speaker. Rather, verbs appear to cluster together according to which constructions they can appear in. This phenomenon is known as subcategorization, and the clusters of each verb are called subcategories.

Consider the two constructions shown in Table 5.1. Given that many verbs in English occur in both, it seems natural to conclude that verbs that occur in one of the two can occur in the other. Unfortunately, this generalization – known as the dative alternation – does not apply for all verbs: for instance, *confess* is grammatical in the PD construction ("Jonathan confessed the truth to Doug") but not in the DOD (*"Jonathan confessed Doug the truth"). Despite never having been explicitly taught that *confess* is ungrammatical in the double object dative – and even though a near-synonym, *told*, is grammatical – fluent speakers of English appear to have no trouble avoiding the incorrect form. How is this explained?

In the following subsections I will evaluate several answers to this question, which requires a critical and detailed discussion of some of the major empirical phenomena involved. At the end of the section I will discuss these findings and elaborate on what they mean and what has yet to be explained.

Hypothesis #1: The data is sufficient

So far I have only asserted, rather than demonstrated, the truth of the claim that children do not receive negative evidence, or at least not enough. But is that actually true?

At the minimum, there is a broad consensus that there is little overt correction of ungrammatical constructions. In an examination of parent-child exchanges, Brown and Hanlon (1970) found that parents were no more likely to express explicit disapproval (or less likely to express explicit approval) when children spoke ungrammatically than when they didn't. In some sense, this is no surprise: if adults *did* frequently offer explicit correction, it would be extremely difficult to have any sort of conversation, given how often children speak ungrammatically.

Another, more realistic, possibility is that children might receive negative evidence in a more subtle or probabilistic form: perhaps adults give replies to ungrammatical utterances that, statistically at least, can be used differentiate them from grammatical ones. Indeed, there is evidence that parents more often repeat 2-year-olds' utterances when they are ungrammatical (Hirsh-Pasek, Treiman, & Schneiderman, 1984). Clarification questions, verbatim questions, and "move on" conversational signals may be statistically related to the grammaticality of children's utterances (Demetras, Post, & Snow, 1986). One study found that 34% of two-year-olds' syntactic errors are followed by implicit feedback of some sort, either via more frequent exact repetitions when the utterance is grammatical, or more frequent recasts and expansions when it is not (Bohannon & Stanowicz, 1988).

Is this type of *indirect* negative evidence sufficient? It is difficult to know how one would determine this definitively, but there are several reasons to think it might not be. The information is probabilistic, after all, and often the differences between

the responses to grammatical and ungrammatical utterances would require impressive statistical abilities to discern. Although there is evidence that children do have impressive statistical abilities in many ways (e.g., Saffran, Aslin, & Newport, 1996), it is not clear that they notice or track *these* differences in particular; and some have argued that even if they do, the differences are not large enough to solve the negative evidence problem (Gordon, 1990; Marcus, 1993).

Another problem is that implicit, subtle feedback does not tell the child what precisely is ungrammatical, only that *something* is (or might be). Was the word incorrect, or the pronunciation of the word? Was the verb used in an incorrect construction? Were the words in the wrong order? Were important words dropped? The child has no way of knowing unless the feedback is explicit. If the adult responses are reformulations – which highlight, by their structure, the location of the error – this problem may be mitigated somewhat; but, at best, this feedback is still only distributional. In a study by Chouinard and Clark (2003), between 20% and 67% of erroneous utterances were found to be reformulated, compared to between 2% and 38% of correct ones (see also Saxton, 1997). It seems clear, at least, that a Gold-style ideal learner would not be able to converge onto the correct language given statistical syntactic evidence of the sort considered here (Gordon, 1990), and as yet I am unaware of any formal, concrete proposals for what sort of learner could.

Hypothesis #2: Language learners do not overgeneralize

One logical possibility for how human learners solve the negative evidence problem is that, like the Subset Principle (Berwick, 1985), they are simply conservative, and do not generalize in any systematic way beyond the input they hear. Unfortunately, this explanation – or at least the strongest version of it – is quite clearly ruled out by the data: both children and adults will cheerfully, and frequently, use words and constructions that they couldn't possibly have heard before. Children have been documented producing novel causative forms like *Don't giggle me* and *You go it in* (Bowerman, 1982); novel passives like *It was bandaided* (E. Clark, 1982); and novel datives like *I go write you something* or *You ate me my cracker* (Gropen, Pinker, Hollander,

Goldberg, & Wilson, 1989). This is just the tip of the iceberg: Pinker (1989) cites and discusses evidence from dozens of studies showing that children spontaneously produce sentences that are clearly generalizations beyond what they have previously heard.

Hypothesis #3: Learners use semantic information

One of the most important suggestions for how learners overcome the "no negative evidence" problem is that there might actually be *positive* evidence about which verbs enter into an alternation and which do not, in the form of correlated morphological, phonological, or semantic features (e.g., Morgan & Demuth, 1996; Pinker, 1984, 1989). This "semantic bootstrapping" hypothesis is argued to overcome the learnability problem by providing a feature that the child could use to distinguish between verbs that do and do not occur in the alternation in question (Mazurkewich & White, 1984; Pinker, 1984). According to this hypothesis, the child notices the semantic and morpho-phonological features pertaining to different "narrow" verb classes, and learns to generalize the syntactic alternation pattern based on those classes (Pinker, 1989). For instance, many verbs that are not dativizable are of Latinate origin, and thus have distinct morphological and phonological features (Green, 1974; Mazurkewich & White, 1984) that predict adult speakers' ratings of the goodness of each construction (Gropen et al., 1989). There is also a strong correlation between verb meaning and syntactic structure (C. Fisher, Gleitman, & Gleitman, 1991). In addition, there is evidence that semantics play a role in constraining generalization: children are able to use semantic information to generalize novel verbs (Gropen et al., 1989; Gropen, Pinker, Hollander, & Goldberg, 1991), and assessments of the grammaticality of some constructions is related to the semantic class membership of the verbs in question (Ambridge et al., 2005).

Although the semantic distinctions between verbs that occur in different constructions are well-attested and appear to be productive in the adult grammar, there is reason to believe that the semantic bootstrapping hypothesis cannot explain everything about the acquisition of verb argument constructions. For one thing, semantic

and morpho-phonological classes do not suffice to capture the full verb-structure distribution (Bowerman, 1988; Braine & Brooks, 1995; Goldberg, 1995), and children may have more difficulty learning verbs that the semantic bootstrapping hypothesis suggests should be easy (Bowerman, 1990). Furthermore, it is demonstrably difficult to ascertain the meaning of many verbs based simply on environmental contingencies and co-occurrences (Landau & Gleitman, 1985; Gleitman, 1990; Gillette, Gleitman, Gleitman, & Lederer, 1991). Even if, as Pinker (1994) claims, this difficulty is only insurmountable for an associationist learner incapable of representing structured hypotheses about verb semantics or applying information from across multiple situations, it is clear that in practice, syntactic information is enormously useful for acquiring the meaning of certain verbs.

This hypothesis, called "syntactic bootstrapping", does not suggest – despite what its name may imply – that verb meaning is learned entirely (or even mostly) via syntactic information; instead, it proposes that words are acquired through a probabilistic process that depends on *multiple* cues, one of which is syntax. In particular, the acquisition of less concrete words like abstract nouns and mental state verbs proceeds by combining information from multiple sources, including extralinguistic world information and syntactic structure (Gleitman, 1990; Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005). Consistent with this, children appear capable of using syntactic information to make inferences about verb meaning (Naigles, 1990; C. Fisher, 1996). In some circumstances syntax may be even more reliable than contextual cues (Papafragou, Cassidy, & Gleitman, 2007).

Although the syntactic bootstrapping hypothesis is a powerful and well-supported theory, it addresses the problem of Feature Relevance, *not* the problem of No Negative Evidence. It is a theory about how a learner chooses among a logically infinite number of possible verb meanings; although it suggests that multiple cues can mutually constrain one another as that meaning is acquired, it does not bear on the question of syntactic overgeneralization raised by Baker's Paradox. By contrast, the semantic bootstrapping hypothesis is a theory about how syntax might be acquired – and, in particular, how certain syntactic overgeneralizations might be constrained by seman-

tic knowledge. Its ability to solve the No Negative Evidence problem is limited by the degree to which it relies on semantic knowledge that may be unlearnable without the aid of syntax (and implausible to assume is innate); but syntactic bootstrapping does not offer an alternate theory about how to resolve the No Negative Evidence problem. Thus, the original question remains: to the extent that correlated semantic and morpho-phonological features are not sufficient to constrain syntactic generalizations of this sort, how *do* children resolve the No Negative Evidence problem?

Hypothesis #4: Learners use indirect negative evidence

Another hypothesis, made originally by Braine (1971), suggests that learners might be able to use indirect negative evidence, inferring that if a certain form does not occur given enough input, then it is probably ungrammatical. One way this might occur is through *pre-emption*: if a verb is encountered in one construction, when another, more felicitous construction would provide the same information, the pragmatic conclusion would be that the unseen construction is actually ungrammatical (Goldberg, 1995). Children 4.5 years and older appear to be receptive to this sort of information (Brooks & Tomasello, 1999), but as of yet there is no evidence that younger children do.

Another form of indirect negative evidence is *entrenchment* – the idea that the more often a verb is observed in one construction only, the less likely it is to be generalized to new ones (Braine & Brooks, 1995). Both children and adults are more likely to rate overgeneralizations as grammatical if they occur in low-frequency rather than high-frequency verbs (Theakston, 2004; Ambridge et al., 2005). Indeed, even children as young as two and a half to three years of age are sensitive to frequency; the less frequent a verb is, the more likely children are to produce overgeneralizations (Brooks, Tomasello, Dodson, & Lewis, 1999; Matthews, Lieven, Theakston, & Tomasello, 2005).

People are sensitive to distributional information on both the verb-general and the verb-specific level. A large literature demonstrates that verb-specific syntactic biases about which structures are likely to follow individual verbs have strong effects on real-time processing (e.g., Trueswell, Tanenhaus, & Kello, 1993; Snedeker & Trueswell,

2004). Verb-general effects can also be seen; in particular, readers will sometimes interpret nouns occurring after a verb as a direct object even when that particular verb does not take direct objects (Mitchell, 1987; Juliano & Tanenhaus, 1993). This overgeneralization is more common if the verb is low in frequency.

How are these relationships between verbs and constructions acquired? While it is clear that semantic factors must play an important role in argument structure acquisition, one possibility is that balancing knowledge on multiple levels may critically depend upon distributional statistics. Support for this comes from a series of artificial language learning experiments by Wonnacott, Newport, and Tanenhaus (2008). In these experiments, adults were presented with languages containing novel verbs and two synonymous novel constructions. In all languages the syntactic input distribution provided the only cue to verb subcategory – semantic features did not correlate in any systematic way with subcategorization – but the co-occurrence of verbs and constructions differed across the languages. Overall, participants proved able to acquire both verb-specific constraints and to generalize. Critically, however,the tendency to generalize was affected by a third source of information: learners exposed to languages in which the majority of verbs occurred in both constructions were more likely to generalize verb to a construction in which it had not occurred in the input.

This evidence suggests that people incorporate information about distributional statistics on several levels of abstraction – from verb-specific information about particular lexical items, to verb-general inferences about verb classes or construction types, to knowledge on an even higher level about variability across verbs or constructions in language in general. But does this sort of distributional learning bear on the No Negative Evidence problem? And if so, what sort of learning mechanism can combine multiple levels of information in such a way to allow overgeneralization of low-frequency forms, but limit overgeneralization with increasing frequency in a principled and justifiable manner?

Bringing it all together

The weight of the evidence indicates that people solve Baker's Paradox, but probably not by making use of negative evidence directly, nor by employing strategy of strict conservativism. In the following sections I show that the hierarchical Bayesian model introduced in Chapter 4 resolves the No Negative Evidence problem on the basis of syntactic-only distributional information alone. This occurs because of general characteristics of Bayesian learning stemming from balancing simplicity and goodness-of-fit, as discussed in Chapters 1 and 2, which I address with respect to these specific problem here. In addition, because the model is capable of learning on multiple levels of abstraction, it can capture other empirical phenomena in the literature.

Study 1 reveals that the model captures adult performance in the artificial language experiments of Wonnacott et al. (2008), demonstrating how such performance results from making inferences about the sort of feature variability one would expect across verbs or constructions in language in general. Study 2 shows that the same model, given syntactic input for the verbs that occur in the dative alternation in a corpus of child-directed speech, learns the correct alternation problem in the absence of negative evidence. Moreover, it productively overgeneralizes verb constructions on the basis of frequency. These results are especially interesting in light of the fact that this model was originally developed to capture the emergence of feature biases in word learning (see Chapter 4), rather than anything particular about verb learning or verbal knowledge at all. Finally, in Study 3 I explore how learning is affected by semantic information being added to the input. Results show that the model is capable of using semantic information to make interesting syntactic generalizations much as people do, even without strong assumptions about the nature of the semantic representations or the nature of the semantic-syntactic linkage. Although I do not wish to imply that these assumptions are unnecessary or (even if they are) that children do not make them, the work suggests that a surprising amount may be learnable in principle without them, and provides a framework within which to rigorously explore the possibility in more detail.

Study 1: Modelling adult artificial language learning

Data: Artificial language input

The purpose of the artificial language learning experiment of Wonnacott et al. (2008) was to determine whether adults exposed to a novel language could acquire knowledge at both the verb-specific and verb-general levels.[1] Over the course of five days, subjects were taught a novel language including five nouns and eight verbs occurring in one of two possible constructions: *VAP* (verb agent patient) and *VPA-ka* (verb patient agent particle(*ka*). During training, participants were presented with a set of scene/sentence pairs, hearing sentences (*VAP* or *VPA-ka*) corresponding to video clips of scenes in which puppets acted out the sentence meaning. Because part of the purpose of the experiment was to explore performance given only syntactic information, both constructions had the same meaning.

Subjects were divided into two conditions based on the language they were exposed to. One group was presented with a language in which each of the eight verbs occurred in both constructions, but seven times as often in the *VPA-ka* construction as in the *VAP*. Wonnacott et al. (2008) dubbed this the Generalist language. The other group was exposed to the Lexicalist language, in which seven verbs occurred in the *VPA-ka* construction only and one verb occurred in the *VAP* only. In both languages, the absolute and relative frequencies of the *VAP* and *VPA-ka* constructions are the same, but the conditions differ widely in terms of the distribution of those constructions across individual verbs; this allows for the evaluation of whether learners can acquire and use verb-general as well as verb-specific statistical information. When presented with a novel verb in one construction, would participants in the Generalist condition be apt to infer that it can occur in both, even if one is more frequent? And would participants in the Lexicalist condition tend to think that it can occur only in one?

[1]For simplicity of presentation, I focus on Experiment 3 in their paper, although the model captures the results of all three experiments, all of which focus on the acquisition of knowledge about variability across individual verbs.

To test this, participants' grammatical knowledge was evaluated in two comprehension and one production test. I focus here on the results of the latter test, in which productive language use was evaluated in a procedure first established by Hudson Kam and Newport (2005). Subjects viewed a scene, heard the verb corresponding to the action in the scene, and were asked to complete the sentence aloud. This procedure allowed the testing of both familiar and novel verbs.[2] There were four novel verbs which were not heard during training at all but were presented four times during the test stage (just prior to the relevant production test): two of the verbs only in the *VAP* construction, and the other two only in *VPA-ka*. The novel verbs are particularly interesting because they reflect subjects' higher-level generalizations about variability across verbs, rather than just information inferred about those particular verbs, as shown in Figure 5-1a). People learning the Generalist language were likely to produce a novel verb in both constructions, though with a strong bias for the *VSO-ka* construction; they thus match the overall frequencies of each of the constructions in the language and ignore the fact that the novel verb itself had only occurred in one construction. By contrast, subjects learning the Lexicalist language, whose previous input consisted of verbs that only ever occurred in only one construction, tended to produce the novel verb only in the single construction in which it was observed.

Model: Level 2 and Level 3

The basic model, which was described in Kemp et al. (2007) and Chapter 4, applies to the problem of acquiring word learning biases[3] – but it can be reframed to apply instead to the acquisition of verb argument constructions by simply changing the input given. Where before the data consisted of observations of categories and the number of times they occurred with each feature, the data now consists of observations of verb types: the observations of each verb consist of the number of its tokens occur in each construction under consideration. Where before we wanted to predict what

[2]Wonnacott et al. (2008) refer to these previously-unobserved verbs as "minimal exposure" verbs, but for simplicity and continuity with Study 2 and 3, I will refer to them as novel verbs.

[3]The basic model we use corresponds to the one described in Kemp et al. (2007) rather than the category-learning extension developed at the end of Chapter 4.

sort of features we are liable to observe category n with (i.e., level 1 knowledge about features), we now would like to predict what sort of constructions we are liable to observe verb n in (i.e., level 1 knowledge about verb constructions). Where before, level 2 knowledge was knowledge about the distribution of features across categories in the world, it is now knowledge about the distribution of verb constructions. As before, this is represented by two parameters, α and $\boldsymbol{\beta}$. In this context, α captures the extent to which each individual verb occurs uniformly in one construction (or not), and $\boldsymbol{\beta}$ captures the average distribution of constructions across the entire language. Low α would indicate that each verb tends to occur uniformly in one construction (as in the Lexicalist language), but does not say anything about *which* construction that might be (here, verb-specific information would be necessary). The $\boldsymbol{\beta}$ would be the same for both Generalist and Lexicalist languages, since (by design) both had seven times as many instances of the *VOS-ka* as the *VOS* construction.

As before, this level 2 knowledge depends on knowledge at a higher level, level 3, which is represented by the hyper-parameters λ and μ. The λ parameter captures something about the range of values expected about the uniformity of constructions (would it be extremely unusual to see complete uniformity, as in the Lexicalist language? Is that the expected pattern?), and the μ parameter the range of values of the expected distribution of verb constructions across the language.

In this chapter I consider two models. Model L2, which we have seen before, assumes that the level 3 knowledge (λ and μ) is already known, and learns the α and $\boldsymbol{\beta}$ values that maximize posterior probability for the given data. Model L3, by contrast, learns λ and μ in addition to α and $\boldsymbol{\beta}$, and assumes that knowledge at Level 4 is given. Both models acquire level 1 knowledge about the expected constructions found in specific individual verb types. Both models also make quite simple representational assumptions, presuming only that individual verbs are represented as a vector of features (consisting of the number of observations of each construction) and that verb-general knowledge is represented by these higher-level parameters. The appendix describes the technical aspects of these models in more detail.

Results: Learning construction variability

I present Models L2 and L3 with data corresponding to the input given to adult subjects over the course of the five days of training in the Generalist and Lexicalist languages. As with the human subjects, I wanted to evaluate how previous exposure to the two languages would change how learners dealt with a small amount of verb-specific information. The model was therefore also presented with a single exemplar of a completely novel verb occurring either in the *VAP* or *VPA-ka* construction.[4] Results are shown in Figure 5-1. Both models (L2 and L3) qualitatively replicate the difference between conditions, demonstrating that the model makes inferences on both the verb-general and verb-specific level, much as humans do. Novel verbs in the Generalist condition are assumed to occur in both constructions, while the same novel verbs in the Lexicalist condition are assumed to occur in only one.

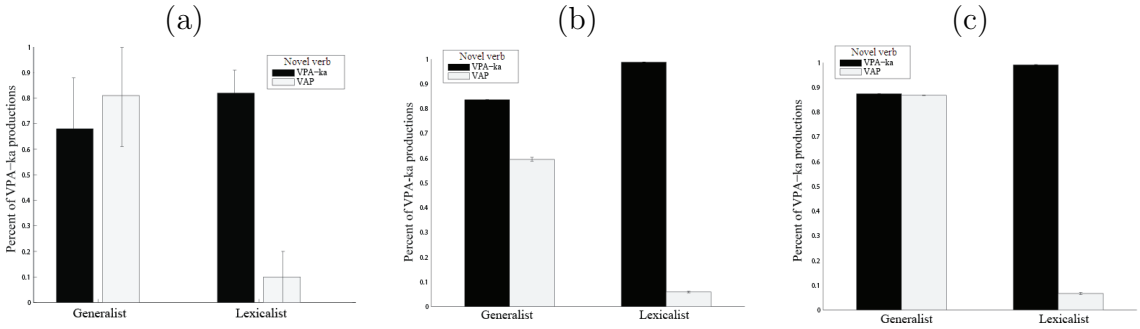


Figure 5-1: Comparison of model performance with human production for novel verbs in an artificial language. (a) Adult performance. Subjects in the Generalist condition were likely to produce a novel verb in both constructions, matching the overall frequency of each in the language, rather than the single construction it was heard in. Subjects in the Lexicalist condition, whose previous input consisted of verbs that occurred in only one construction at a time, tended to produce the novel verb only in the single construction it occurred in. (b) Model L2. (c) Model L3. Both models qualitatively replicate the difference in human performance in the each condition. Model L3, which can learn at a higher level of abstraction, matches human performance more closely.

Model L3 outperforms L2 in the Generalist condition, more accurately predicting human production for the novel verb occurring in the less-frequent construction in the

[4]Humans were presented with four novel verbs rather than one, but because the model does not have the memory limitations that humans do, it is more appropriate to evaluate its generalization given only one novel verb.

language (*VAP*). Even though that verb was heard in the *VAP* construction only, both humans and Model L3 predict that it will occur in the other (*VPA-ka*) construction nearly 87.5% of the time (which is the base rate of *VPA-ka* in the language as a whole). Although Model L2 qualitatively captures the difference between the Generalist and Lexicalist conditions, it is quantitatively less accurate, extending the *VAP* form to *VPA-ka* 60% rather than 87.5% of the time. The reason for this difference is that the hyperparameters (λ and μ) over α and $\boldsymbol{\beta}$, which are "built in" for Model L2, weakly restrict the range of α and $\boldsymbol{\beta}$ to avoid extreme values. The Generalist condition, in which each of the individual verbs' distribution of constructions precisely mirrors the distribution in the language as a whole, is best captured by values of α and $\boldsymbol{\beta}$ that happen to be dispreferred by the hyperparameters of Model L2. One might select different hyperparameters, but that would be arbitrary and *post hoc*; it could also cause the model to be unable to capture a different dataset. By contrast, because Model L3 can learn the hyperparameters λ and μ, it infers that the more extreme values are more appropriate for this dataset.

In the all of the subsequent sections, both Model L2 and L3 were analyzed, and results were qualitatively similar for both. For space and readability reasons, I report only the results from Model L3, which usually slightly outperforms Model L2.

Study 2: Modelling the dative alternation

Data: Corpus of child-directed speech

The previous study explored performance of the model given the data presented to adults in an artificial language learning task, but it is important to determine how relevant this is to the task facing the child. In this study I therefore present the model with real-world data taken from a corpus of child-directed speech. Because the dative alternation is a central, well-studied example relevant to Baker's Paradox, I choose to focus on verbs that occur in it. The data is collected from the sentences spoken by adults in the Adam corpus (Brown, 1973) of the CHILDES database (MacWhinney,

2000), and consists of the counts of each construction (PD and POD) for each of the dative verbs (as listed in Levin (1993)) that occur in the corpus.

An additional variable of interest is what sort of evidence may be available to the child at different ages. This can be loosely approximated by tallying the number of occurrences of each verb in each construction in subsets of the corpus split by age (see Table 2 in the appendix). The Adam corpus has 55 files, so the first segment, *Epoch 1*, contains the verbs in the first 11 files. The *Epoch 2* corpus corresponds to the cumulative input from the first 22 files, and so on up until the full corpus at *Epoch 5*. The dative verbs in the first file only, corresponding to approximately one hour of input, constitute *Epoch 0*. 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.

Model: Learning verb classes

The previous study demonstrated that both models can acquire verb-general variability information, but this may not suffice to address the complexity of the problem faced by a child. In natural language, verb-general statistics may be shared among only verbs in a certain class rather than over all the verbs in the language. For instance, some verbs occur in both constructions in the dative alternation, but others occur in only one. A learner that could only make inferences about verb-general statistics across the language as a whole would not be capable of realizing that there were these two types of verbs. Presented with a novel verb occurring only once in one construction, such a learner might be more likely to generalize it to both than one who realized that it might belong to a non-alternating class.

I therefore add the ability to discover verb classes (or "kinds"[5]) to both Models L2 and L3 and denote this extension with the prefix K (i.e., Model K-L2 and K-L3). As in Kemp et al. (2007), this results in a model that assumes that verbs may be grouped

[5]Because "class" is more sensible nomenclature for verbs, I will refer to them as classes throughout this chapter, but these are the same entities referred to as "kinds" in the Kemp et al. (2007) work and in Chapter 4.

into several classes, where each class is associated with its own hyperparameters. The model is not told how many classes there are, nor which verbs occur in which class; instead, it forms the categories based on the data, in combination with a prior under which all class assignments are possible, but fewer classes are favored. The goal of learning is to simultaneously infer how verbs are assigned to classes, along with the values of the hyperparameters that describe each class. It is described more fully in the appendix.

Results: Overgeneralization with frequency and quantity of data

Figure 5-2 shows class assignments predicted by Model K-L3. It captures the intuitively sensible pattern: those verbs that occur in one construction tend to be found in a separate class from verbs that occur in both. Sensibly, when there is less data (i.e., at the earlier *Epochs*), the model is less certain: the class assignments for subsets of the full corpus are generally less sharp than they are for the entire corpus. Frequency also plays a role; the model is more certain about class assignments of the high-frequency verbs like *give* and *call*, and much less confident about the class assignments of the low-frequency verbs like *sing*. In part because of this lack of certainty, we would expect the model to be more likely to overgeneralize the low-frequency verbs beyond the constructions in which they occur in the input.

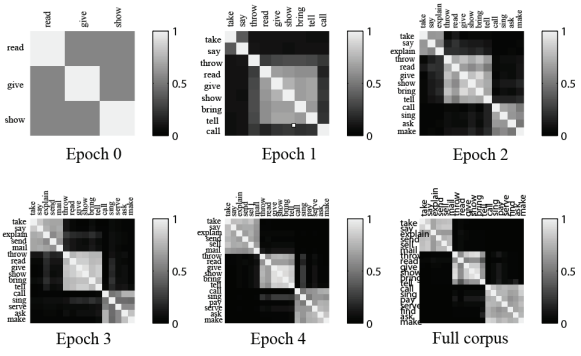


Figure 5-2: Class assignments given by Model K-L3. Lighter colors indicate increasing probability that the verbs in that row and column are assigned to the same class. The diagonal is always white because each verb is always in the same class as itself.

There are two ways of testing this prediction. First, we can examine model predictions for how to produce novel instances for each of the input verbs. These results are shown in Figure 5-3. It is evident the model overgeneralizes more often for the low-frequency verbs. The predicted construction distribution for high-frequency verbs like *give* or *call* is very similar to the observed distribution (shown in the associated pie chart). But low-frequency verbs like *explain* or *serve*, which only occur in one construction in the input, are nevertheless somewhat likely to be produced in the other construction. This is because there is still some possibility that they are actually members of the alternating class; as more and more verb tokens are heard and these verbs are still only heard in one construction, this becomes less and less likely.

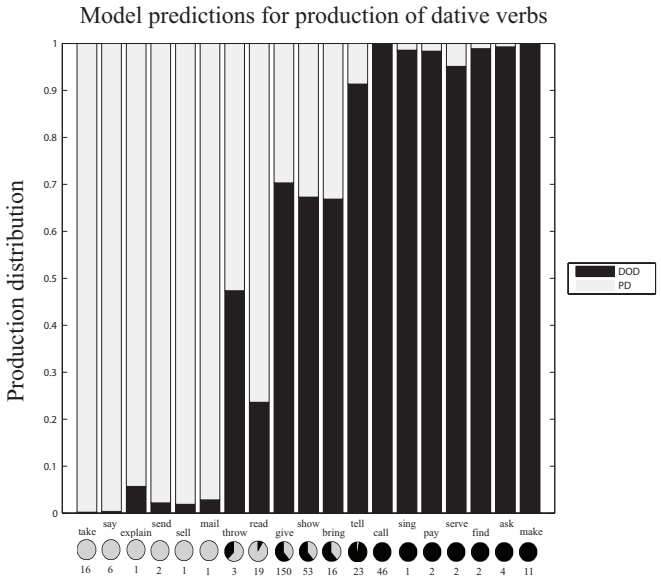


Figure 5-3: Production predictions of Model K-L3 for each verb in the full corpus. High-frequency verbs' constructions are produced at a distribution close to their empirical distribution, while low-frequency verbs are more likely to be overgeneralized (i.e., produced in a construction that they did not occur in the input). The production distribution is denoted with the stacked bars; the associated pie chart depicts each verb's observed distribution, and its empirical frequency is the number under the pie chart.

Second, we can examine how the model produces novel instances of verbs it has seen before, at each *Epoch*. Figure 5-4 shows the degree of overgeneralization for each of the verbs that occurred in just one construction at each *Epoch* in Model K-L3.[6]

[6]We exclude verbs that have already occurred in both constructions in the data, because by

This is calculated by finding the difference between the proportion of times the verb is observed vs. produced in the DOD construction.[7] If this difference is zero then it means the model produces the verb constructions precisely at the same frequency as they occurred in the corpus. The larger this difference is, the more the model has "smoothed," or overgeneralized away from, the observed data.

The results indicate that as the frequency of the verb increases, overgeneralization decreases: the difference between observed and predicted approaches zero. There is also an interaction with *Epoch*: verbs of equivalent frequencies are overgeneralized more in earlier *Epochs*. For instance, verbs that occur once in the full corpus are overgeneralized one-third as often as verbs that occur once at *Epoch 2*. The reason for this is that there is more data in the corpus at later *Epochs*, and the model is therefore more certain about the probable constructions it infers for even the low-frequency verbs.

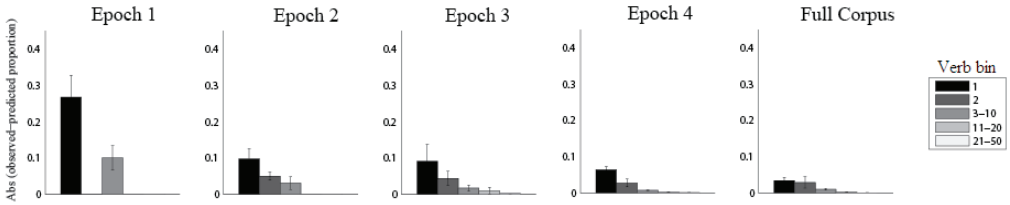


Figure 5-4: Degree of overgeneralization of non-alternating verbs by *Epoch* for Model K-L3. The y axis reflects the degree of overgeneralization, calculated by taking the absolute value of the difference between the proportion of the time the verb is observed vs. produced by the model in the DOD construction. Verbs of different frequencies are grouped in bins along the x axis: thus bin 1 contains all of the verbs that occurred only once in the corpus, bin 2 contains verbs occurring twice, and so on.

Both models appear to be learning in the absence of negative evidence: although they are never "told" explicitly that some do not occur in one construction or another – nor do they receive any correction of early overgeneralizations – the models eventually form classes of verbs along precisely these lines. This qualitatively captures two of the major phenomena found in the acquisition of verb argument constructions: more frequent verbs being overgeneralized more rarely, and a general decrease of over-

definition, they cannot overgeneralize beyond the constructions in which they have been observed.

[7] One could equivalently calculate this for the PD rather than DOD construction; since there are only two, this captures the same information.

generalization with age. The importance of frequency is made clearer by combining all of the *Epochs* together, which is shown in Figure 5-5.

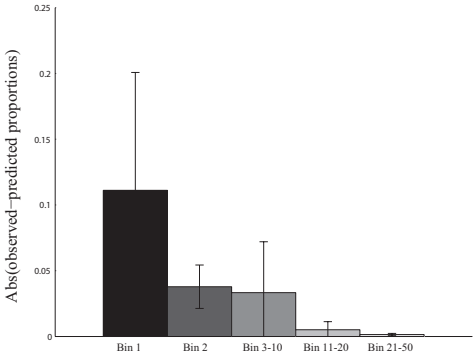


Figure 5-5: All *Epochs*: degree of overgeneralization of non-alternating verbs by for Model K-L3. This graph makes clear the effect of frequency on overgeneralization. As before, verbs of different frequencies are grouped in bins, such that bin 1 contains all of the verbs that occurred only once in the corpus, bin 2 contains verbs occurring twice, and so on.

To what degree is the change in overgeneralization with age and verb frequency due to the fact that the model can group verbs into classes? I address this question by comparing performance of the class-learning model K-L3 with Model L3 from Study 1. The results are shown in Figure 5-6. The ability to learn verb classes leads to less overgeneralization at all *Epochs*, but many of the qualitative effects are similar for both models. Both capture three of the major empirical phenomena: learning in the absence of overt negative evidence, and decreasing overgeneralization with increasing age as well as verb frequency. This is because these aspects of model performance result from general characteristics of Bayesian learning, rather than particular assumptions made by any specific model. I address this point in more detail in the discussion section.

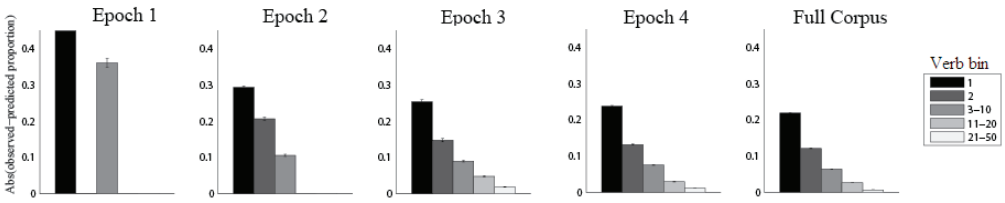


Figure 5-6: Model L3: Overgeneralization of non-alternating verbs by *Epoch*.

One interesting empirical finding is that neither model captures is the "U-shaped learning curve" reported for many aspects of language acquisition (e.g., Marcus et al., 1992). In U-shaped learning, periods marked by overgeneralization errors occur after an initial stage of correct production; yet in this model, overgeneralization is greatest at the beginning. U-shaped learning can sometimes emerge from Bayesian models, if the representation is complex enough that for the first few datapoints it is actually more parsimonious to simply memorize the exemplars directly (see Chapter 2 for further discussion of this point). This does not occur in this model, but that may not be a concern, since the evidence for U-shaped learning curves in the acquisition of verb constructions (as opposed to verbal morphology) is equivocal at best. The difficulty is that correct performance involves avoiding overgeneral constructions entirely, and unless a child's entire productive output is sampled, it is difficult to determine whether the child knows the verb and never overgeneralizes incorrectly, or simply failed to do so during the sampled segments. If, as has been suggested (Tomasello, 2000), the youngest children learn verbs on an item-by-item basis and do not form the abstract notion of constructions until they are older, then we must examine individual verbs rather than constructions as a whole. Of the eight verbs associated with overgeneralization errors found in the spontaneous speech of five children in Gropen et al. (1989), only three were used grammatically elsewhere in the corpus; of those three, two were used grammatically before the error (Adam, *fix* and *put*) and one was used after (Ross, *say*). Does this support a U-shaped acquisition pattern? The data are simply too sparse to tell. Even if it is true that very young children have abstract knowledge of verb constructions at a young age (Conwell & Demuth, 2007), sparse sampling still poses a problem: the rarer an error is, the more unlikely it is to be observed in a corpus at the time it first emerges (Tomasello & Stahl, 2004).

For instance, Gropen et al. (1989) examined the spontaneous speech of five children in the CHILDES database. The children used both constructions in the dative alternation (PD and DOD) idiosyncratically: 28 verbs occurred in both constructions, 22 potentially alternating verbs were used in the DOD construction only, and 24 were used in the PD construction only. No child acquired constructions in the same order

for all of their verbs. Although overgeneralization errors were rare[8], those that did occur were preceded by grammatical uses of both DOD and PD forms – which might suggest a U-shaped acquisition pattern, but there is reason to be cautious here.

In sum, then, it is unclear whether the lack of U-shaped learning mimics children's acquisition of verb argument constructions or not. However, the model does capture other more widely-observed and well-established empirical phenomena: learning in the absence of negative evidence, and decreasing overgeneralization as age and verb frequency increase.

Study 3: Exploring the role of semantics

Thus far I have focused on the acquisition of verb constructions based on only syntactic information, since that is central to the puzzle raised by Baker's Paradox. However, learning verb constructions involves the acquisition of semantic knowledge as well: not only realizing what syntactic forms are grammatical, but also recognizing what relationship those forms have to the semantic meaning of the verb. The precise role that semantic features play in acquiring verb constructions is still debated, and the answer is almost certainly more complex than I attempt to capture in this model.

The results from Study 1 and Study 2 suggest that at least some aspects of the learnability problem are in principle solvable based on syntactic input alone. Nevertheless, it is undeniably true that humans learn and use semantic as well as syntactic information. For instance, children taught novel motion verbs in PD syntax extend the verbs to DOD syntax somewhat differentially according to the semantics of the verb – more often when the verb depicts transfer of possession than motion to a location (Gropen et al., 1989). Does this model have the ability to use co-occurring semantic and syntactic information to mutually reinforce each other in this way?

[8]It is difficult to determine how, precisely, to measure "rarity" of a form. Gropen et al. (1989) found an incorrect DOD form in 1 out of every 3,924 sentences, yielding an error rate of 0.0003. As they note, however, many sentences occurred in contexts where potentially dativizable verbs were inappropriate. A different measure might include only those, or might calculate over specific verbs (either types or tokens); the latter measure yields an error rate that is considerably higher (between 5% and 14%), although still not large in absolute terms.

Data: Adding semantic features

I evaluate how semantics may be incorporated into the model by adding to the corpus of dative verbs a semantic feature that precisely parallels the semantics of each of the three classes. This semantic feature therefore has three possible values: one corresponding to the class of alternating verbs (which I will call semantic class A), one to those verbs occurring only in PD syntax (semantic class P), and one to verbs occurring only in DOD syntax (semantic class D). For instance, *give*, which occurs 106 times in DOD syntax and 44 times in PD syntax, has a semantic feature occurring 150 times in semantic class A; *call*, which occurs 46 times in DOD syntax, occurs 46 times in semantic class D; and *say*, which occurs 6 times in PD syntax, has a semantic feature occurring 6 times in semantic class P.

Semantics are assigned in this way for several reasons. Semantics are not obvious from corpora in the same way that syntax is, so it would be difficult to assign semantic features in a way that matches the actual input, especially since we do not know what features are obvious or available to children in those situations. Furthermore, this highly simplistic version of semantics allows for a clear exploration of precisely how people might generalize when semantic features are (and are not) correlated with syntax in the expected ways. In particular, the semantic bootstrapping hypothesis suggests that semantic features are correlated not with the syntactic constructions, but rather with the underlying "narrow" semantic classes – this is the reason they are believed to be a useful source of evidence for solving Baker's Paradox. Even though Study 2 suggests that these features may not be necessary to solve the No Negative Evidence problem, it is still interesting to ask how people might generalize when the correlation of semantic and syntactic features follows this pattern.

To test the generalizations made by the model, I evaluate it on this corpus of dative verbs (with semantic features included). In order to mimic the Gropen et al. (1989) experiment and test second-order generalization, I add one additional novel verb. The six conditions differ in the nature of that novel verb, as shown in Table 5.2. In three conditions the syntax of the novel verb is DOD, and in the other three it is

Table 5.2: Conditions based on semantic and syntactic features of novel verb.

	Syntactic form	
Semantic form	PD	DOD
D	Other Non-Alternating	Same Non-Alternating
A	Alternating	Alternating
P	Same Non-Alternating	Other Non-Alternating

PD. Each syntactic form is paired with each semantic form. When the syntactic form corresponds to the same semantic form it matches in the input data, I refer to that as the *Same Non-alternating Form* condition (i.e., PD syntax, semantic class P; and DOD syntax, semantic class D). If the semantic form occurs in the alternating class in the input, that is the *Alternating Form* condition (i.e., both PD and DOD syntax, semantic class A). And if the semantic form and syntactic form conflict based on the input corpus, I refer to that as the *Other Non-alternating Form* condition (PD syntax, semantic class D; DOD syntax, semantic class P). As a baseline, I compare these six conditions to the sort of generalization that occurs when there are no semantic features at all.

Model: Inference on multiple features

The models in Study 1 and Study 2 incorporate only a single feature, syntax, but can be trivially extended to include multiple features, just as the model in Chapter 4 incorporates multiple category features. As in that model, we assume that each feature is independently generated, which allows inference to proceed exactly as before except that each α and $\boldsymbol{\beta}$ is learned separately for each feature. The posterior probability for the full model is therefore the product of the probabilities along each feature.

Results: Using semantics for generalization

As shown in Figure 5-7(a), Model K-L3 generalizes according to semantics in just the way children do in the Gropen et al. (1989) experiment. Both children and the model are more likely to produce the construction they were not presented with

if its semantic feature is associated with the alternating class of verbs. Sensibly, if the semantic feature matches the syntax of the same non-alternating form, then the model rarely produces the unattested construction.

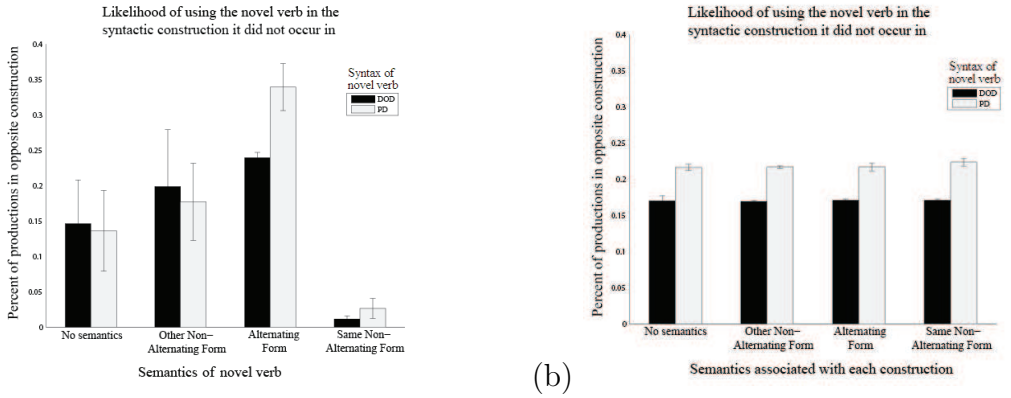


Figure 5-7: Percent of generalization of novel verb to the syntactic construction it was not previously observed in. (a) Model K-L3; (b) Model L3. Model K-L3 behaves qualitatively like human responses, generalizing most to the other construction when the semantic feature is consistent with the non-alternating class, and least when it is consistent with the same alternating class. Model L3 does not, indicating that the ability to group verbs into appropriate classes is necessary to capture this aspect of human behavior.

One interesting result is that, in general, the model is more likely to overgeneralize a verb occurring in PD syntax to DOD syntax than vice-versa – but only if the semantics is alternating. This is because among the verbs that alternate, the DOD construction is approximately twice as frequent. Therefore when a novel with alternating semantics is encountered in PD syntax, the model probabilistically puts the novel verb in the alternating class, and therefore assumes that it tends to have a similar distribution over constructions as other verbs in that class – i.e., that it should occur more often in the DOD construction. But because the model also takes verb-specific information into account (and because the verbs in the class are fairly variable) it does not assume that its distribution will be identical to that of the class. In other words, because a verb with alternating semantics will be assumed to belong to a class in which the DOD construction is more frequent, novel verbs occurring in the PD construction will be more often produced in the DOD construction than novel verbs in the DOD construction will be produced in the PD construction.

The fact that the success of the model depends on its ability to form separate verb classes is apparent when we examine the performance of Model L3, shown in Figure 5-7(b). This model, which cannot form separate classes, generalizes each novel verb identically, regardless of its semantics. In essence, the class information is the vehicle for capturing the relationship between semantics and construction usage.[9] Even without class information, the model is more likely to overgeneralize verbs occurring in PD syntax to DOD syntax (rather than vice-versa); this is because it is still capable of learning statistics on the level of the language as a whole, and in the language as a whole the DOD construction is somewhat more frequent than the PD construction.

One final interesting result is that in the *No Semantics* and *Other, Non-Alternating Form* conditions – i.e., when presented with a novel verb where either there is no semantic feature, or when the semantics is inconsistent with the syntactic form – Model K-L3 and Model L3 show opposite behavior. In Model K-L3 the novel verbs are overgeneralized slightly more from DOD syntax to PD syntax, whereas in Model L3 the opposite occurs. Why is this? Examining the class assignments in Figure 5-8 tells the story. Class groupings are as we would expect in the *Alternating* and *Same, Non-Alternating* conditions. However, because the semantic features in the *Other, Non-Alternating Form* class contradict the semantic features, it is difficult to know what to make of these verbs. The model therefore acts sensibly by assuming that they are likely to be in an entirely new category of their own – a category about which little is known. The model is slightly more likely to make this assumption if the novel verb occurs in DOD syntax since there is already a high quantity of data in the DOD-only class, which makes it fairly certain that the class is not associated with semantic feature P. It is somewhat more uncertain whether the class of verbs occurring in PD syntax is associated with semantic feature D, because there are fewer examples in that class.

[9]It is easy enough to imagine a model that *could* capture the relationship between semantics and syntax, perhaps by building it in via explicit hypotheses about semantics-syntax linking rules. But it is interesting that even in this model, which has no such hypotheses, the relationship can be learned as long as it has the ability to form separate verb classes.

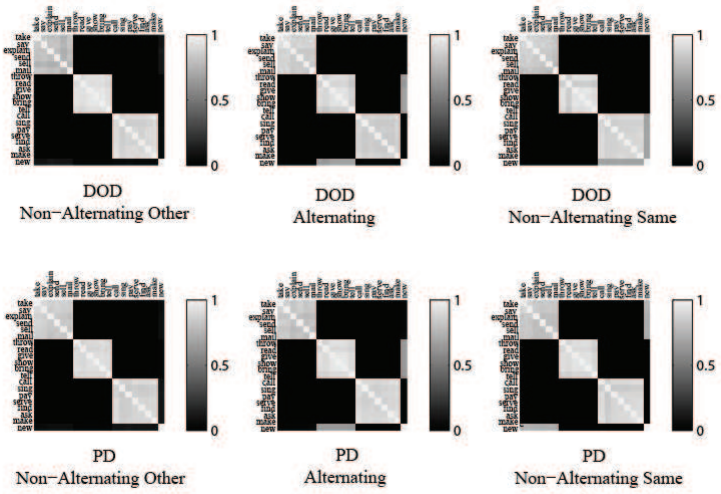


Figure 5-8: Verb class assignments for corpus with semantic as well as syntactic features, in six conditions of novel verb. In the *Alternating* and *Same, Non-Alternating* conditions, the novel verb is grouped into the appropriate class; in the *Other, Non-Alternating* condition, the novel verb is grouped in a class of its own.

One might object that it is unrealistic to assume the existence of one semantic feature that precisely follows the correct verb classes. It may be far more accurate to assume that there are many different features, each somewhat noisy, that are only statistically associated with the correct classes. Does this model qualitatively match human behavior even if the semantic features are less clean?

To evaluate this I present the model with the same corpus, except instead of one semantic feature with three possible values, each verb is associated with three semantic features, and each feature is associated with the correct verb class 60% of the time. The results in the same generalization task as before, shown in Figure 5-9, are qualitatively identical to the previous results. It is evident that it is not necessary for the semantic feature(s) to be perfectly clean in order to explain appropriate generalization on the basis of semantics as well as syntax.

Even this is a simplification of a situation that is undoubtedly far more complex: certainly, one of the things that makes verb learning difficult is that there are many features (semantic and otherwise) that are simply irrelevant, uncorrelated with the correct class assignments, perhaps even correlated with other regularities among the verbs. To what extent does this result depend on this simplification? I addressed

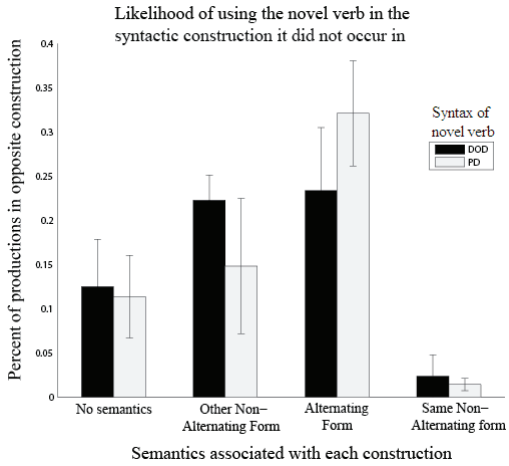


Figure 5-9: Verb class assignments for corpus with noisy semantic as well as syntactic features, in six conditions of novel verb. As before, in the *Alternating* and *Same, Non-Alternating* conditions, the novel verb is grouped into the appropriate class; in the *Other, Non-Alternating* condition, the novel verb is grouped in a class of its own.

the broad question of Feature Relevance (of which this is a subset) in more detail in Chapter 4, but we can explore it in this context by adding additional features to the verbs in this analysis.

I therefore present the model with the same data as before – one syntactic feature corresponding to the construction counts, and three semantic features noisily associated with the correct verb class – and add to it seven additional features. Three of those are noisily correlated with the syntactic construction counts, and four covary in such a way as to identify four completely different verb classes. (Note that they do not covary with each other; each of the features picks out a different class organization). Adding these additional features should act to partially outweigh the semantic features. I evaluate the generalizations as before, based on the semantic features, and the results are shown in Figure 5-10. As we might expect, the qualitative pattern is still the same, but the generalizations are far less marked.

Of course, there will be some extreme at which the additional features are noisy enough, or pick out other categories strongly enough, to completely wash out any effects of the semantic and syntactic features. For instance, I can add six additional features to this already noisy dataset, and specifically set every features to be consistent with a partitioning that puts all verbs in the same class. As a result, in this

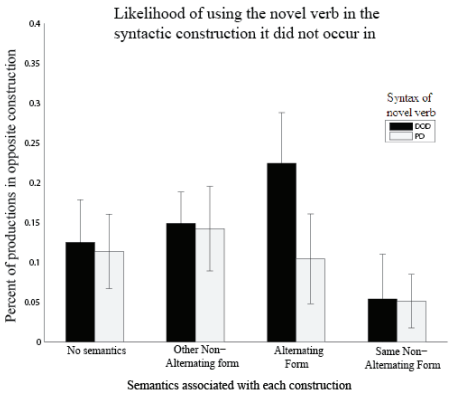


Figure 5-10: Verb class assignments for corpus with noisy semantic and syntactic features, plus seven additional noisier features, in six conditions of novel verb. As before, in the *Alternating* and *Same, Non-Alternating* conditions, the novel verb is grouped into the appropriate class; in the *Other, Non-Alternating* condition, the novel verb is grouped in a class of its own. These effects are much less marked than when the features are cleaner.

dataset most of the features either directly support an analysis in which all verbs are in one class, or support conflicting analyses that (at least somewhat) have effects that cancel each other out. As Figure 5-11 demonstrates, the resulting predictions this time do *not* qualitatively capture human performance.

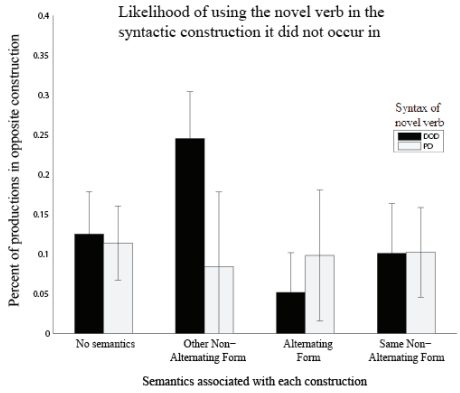


Figure 5-11: Verb class assignments for corpus with noisy semantic and syntactic features, plus 13 additional noisier features, in six conditions of novel verb. Unlike before, the model does not behave sensibly, largely because it tends to group all of the verbs into one class, causing overall imbalances in construction distributions across the dataset as a whole to play a larger role.

As I discussed in Chapter 4, generalization is a function of the coherence and number of features that pick out the correct class assignments. This brief exploration

demonstrates that this sort of learning is somewhat robust to noise and error, but not is not infinitely so. A great deal of further work is necessary to understand precisely how and to what extent additional features matter – to flesh out what the shape and nature of that "generalization function" is.

Discussion

In this chapter I have presented a domain-general hierarchical Bayesian model that addresses how abstract learning about feature variability can be combined with verb-general learning on the level of construction-based verb classes and verb-specific learning on the level of individual lexical items. It captures the qualitative patterns exhibited by adults in an artificial language learning task, as well as those exhibited by children over the course of acquisition. This model suggests that Baker's Paradox can be resolved by a certain kind of learner, even based on syntactic-only input. Furthermore, it does so in a (largely) domain-general way, without making strong language-specific representational assumptions about verb constructions.

In the following section I evaluate these conclusions in more detail. I take care to orient this research with respect to other computational models in order to highlight its contributions and limitations.

Abstract learning about feature variability

Study 1 demonstrates how it may be possible to learn abstract knowledge about variability across verb types, just as adult subjects do in the experiment performed by Wonnacott et al. (2008). Both humans and our model acquire different generalizations based on whether the input comes from a language in which all verbs occurred in both constructions or a language in which each verb occurred in only one construction. This sort of statistical learning is part of a larger process of balancing information on multiple levels of abstraction – a theme we observe throughout this thesis – and the work demonstrates how this balance may be achieved.

To my knowledge there are no computational models that can capture the acqui-

sition of higher-level knowledge on the level of verb or construction variability. Many models address the issue of learning more lexically-specific and detailed verb information (e.g., Dominey, 2003; Chang, 2004). However, although several models are capable of learning verb-general information about classes or construction types (e.g., Dowman, 2000; Onnis, Roberts, & Chater, 2002; Alishahi & Stevenson, 2008), even these do not acquire even higher-level information about their variability, and would therefore be unable to account for human behavior in the Wonnacott et al. (2008) task. The work by Dowman (2000) and Onnis et al. (2002) is focused mainly on the problem of negative evidence, and compares segments of toy grammars of varying degrees of complexity. The work by Alishahi and Stevenson (2008), which is in many ways an impressive model capable of capturing many aspects of semantic and syntactic acquisition, nevertheless cannot account for higher-order knowledge about feature variability. Their model and this one have numerous similarities: both are Bayesian, both can flexibly determine how many constructions or verb classes are appropriate given the data, and both capture similar qualitative patterns in acquisition. But because their model does not do inference on the highest levels (level 2 or level 3), it is capable only of learning about which semantic and syntactic features to expect for each construction, rather than making more abstract judgments about how variable they are expected to be.

Is the ability to learn on this level important in order to explain the acquisition of verb constructions in natural language? Because many of the qualitative aspects of acquisition can be captured by models that cannot learn on that level, it is possible that this sort of learning is not necessary. Nevertheless, several considerations should make us hesitant to draw this conclusion. First, the experiments of Wonnacott et al. (2008) suggest that humans (or at least adult humans) are *capable* of this kind of learning; it seems odd if people to have learning abilities that are too powerful or unnecessary to the task to which they are applied. It might be that this sort of learning is a domain-general capacity, useful when applied to other domains (such as, perhaps, the acquisition of word learning biases, as in Chapter 4), that may not be strictly necessary for the acquisition of verb constructions but that people are happy

to apply nevertheless.

Another possibility is that there may be conditions under which the ability to learn about variability is particularly useful. For instance, the model presented here is capable of learning the distinction between alternating and non-alternating verb classes on the basis of syntactic input alone. The model of Alishahi and Stevenson (2008) can form constructions, but only on the basis of differences in features rather than on patterns of feature variability. As a result, it would be unable to form the distinction between non-alternating and alternating verb classes without additional semantic features to assist. Each individual verb usage would occur only in PD or DOD syntax, and without semantic information differentiating alternating from non-alternating verbs, the model would only be capable of inferring a maximum of two constructions (DOD and PD). It would thus be able to learn that individual verbs are alternating or non-alternating (and thus resolve the negative evidence problem for those verbs). However, because it would have no reason to form a single alternating construction, it would not – as our model does – be able to infer that a completely novel verb appearing once in each construction is alternating, but that a verb appearing twice in one may not be. This is the sort of generalization made by adult subjects in the artificial language of Wonnacott et al. (2008), and it is an open question whether children or adults make it in more naturalistic circumstances.

A solution to the No Negative Evidence problem

One implication of this work is that it demonstrates how Baker's Paradox may be resolved on the basis of syntactic information alone. The Bayesian learner presented here, given the syntactic information from a corpus of dative verbs used in child-directed speech, can learn the appropriate categories of verbs. In so doing it solves the negative evidence problem, correctly realizing that verbs that have been observed often in one construction but never in another probably are not grammatical in both, but that verbs that have been observed rarely in one construction and never in another might be. In essence, the learner takes indirect negative evidence into account, and is a formal instantiation of the notions of entrenchment and pre-emption suggested by

other researchers (Braine, 1971; Braine & Brooks, 1995; Goldberg, 1995). Consistent with this hypothesis, the model – like people – is more apt to overgeneralize lower-frequency verbs or verbs earlier in the process of acquisition.

This performance is not an idiosyncratic property of specific choices made in setting up the model, but is rather the result of a general property of optimal inference, as we saw in Chapter 2. Because Bayesian inference trades off complexity (via prior probability) and goodness-of-fit (via likelihood), it tends to prefer hypotheses that are neither too simple nor too complex, but rather "just right" (as in Hypothesis B in Figure 2-2). As a result of this, as the number of datapoints increases, the likelihood increasingly favors the theory that most closely matches the observed data, and overgeneralization decreases. The likelihood in Bayesian learning can thus be seen as a principled quantitative measure of the weight of implicit negative evidence – one that explains both how and when overgeneralization should occur.

Because this pattern of inference is a general result of Bayesian inference, other computational approaches to the acquisition of verb argument constructions provide the same natural solution to Baker's Paradox (Dowman, 2000; Onnis et al., 2002; S. Niyogi, 2002; Alishahi & Stevenson, 2008). For instance, Dowman (2000) illustrates the same principle explicitly, by comparing toy grammars with and without subclasses of non-alternating verbs; as the amount of data increases, the more complex grammar is preferred and overgeneralization disappears. Their work involves a simplistic toy grammar segment and an idealized artificial corpus rather than the more naturalistic child-directed data considered here, but both models show the same ability to deal sensibly with the problem of negative evidence. More similarly to this work, Onnis et al. (2002) use a Bayesian model to demonstrate the learnability of an alternation based on statistics from corpora of child-directed speech. Their model succeeds in this for the same reason ours does. Our model makes different (in many ways simpler, more domain-general) representational assumptions, and is in other ways more flexible and powerful, with the ability to learn on multiple levels of abstraction, and the ability to determine flexibly how many classes of verbs there are. But in terms of the problem

of negative evidence, all of these models – ours included – solve it in the same way.[10]

Of course, there *are* models – computational as well as less formal ones – that do not provide the same natural solution to Baker's Paradox. These models can be broadly classified into two types: those that, like the Subset Principle, do not generalize at all (Berwick, 1985), and those that limit generalization on the basis of non-syntactic features (e.g., Pinker, 1989). While the non-generalization approach provides a solution to the problem in the sense that it explains how a learner might, in principle, arrive at the correct grammar, it does not capture an important empirical phenomenon: namely, that both children and adults *do* overgeneralize. Models that constrain generalization on the basis of correlated features, of which Pinker's (1989) proposal is perhaps the most elaborated and elegant example, can result in periods of overgeneralization. The difference is that the generalization is ultimately constrained by the presence of the other, correlated features rather than by the accumulated weight of indirect negative evidence. This sort of model is sensitive to indirect negative evidence only in a trivial sense, in that it behaves differently based on whether a certain sort of data is present or absent; a direct dependence on frequency is not a part of this approach in the same way that it is for the class of models that performs some sort of tradeoff between simplicity and goodness-of-fit.

One aspect of the problem that none of these models address – ours included – is the question of how the child knows which sort of evidence is important: an aspect of the problem of Feature Relevance.[11] Pinker raised this point about indirect negative evidence, saying that "it can't be true that the child literally rules out any sentence he or she hasn't heard, because there is always an infinity of sentences that he or she hasn't heard that are grammatical... So the question is, under exactly what circumstances does a child conclude that a nonwitnessed sentence is ungrammatical?

[10] In fact, even connectionist models (e.g., Allen & Seidenberg, 1999; Desai, 2002) implicitly incorporate a sort of tradeoff between simplicity and goodness-of-fit. Often the tradeoff is non-optimal, since the preference for simplicity emerges out of choices about network architecture, number of training epochs, and other modelling choices rather than the mathematics of probability theory, but as long as any tradeoff is being made, overgeneralization will decrease with increasing amounts of data.

[11] Aside from the models that specify that the child must innately know which features are relevant, that is.

This is virtually a restatement of the original learning problem." (Pinker (1989), p. 14). How does the child know that those particular syntactic forms are the interesting and relevant ones? How does she realize what semantic features are important? This knowledge has just been given to this model, and this work makes no particular claims about how it comes about. However, it is essential to realize that what I have done is not simply restated the learning problem, as Pinker suggests: rather, I have suggested an answer to one problem (the No Negative Evidence problem, or how to rule out logically possible alternatives without negative evidence), leaving another still unsolved (the Feature Relevance problem, or how to know which of a potentially infinite number of dimensions to generalize along). Though the latter problem remains a difficult and open question, it is not the *same* problem. The logic of Baker's Paradox would be the same even if there were only one possible dimension of generalization: the dilemma exists because one can never be certain that an unobserved datapoint (along that dimension) is truly ungrammatical, or simply unobserved. How the learner knows which dimensions to pay attention to is a different issue, an aspect of which I considered (albeit in a different domain) in Chapter 4.

Representational assumptions

Because this model was originally developed for a completely different domain, it makes few domain-specific assumptions about the representation of verbs and verb constructions. Both semantic and syntactic information is represented as a vector of features. An advantage of the simpler representation is that it highlights which phenomena emerge due to the nature of the data and the characteristics of Bayesian (optimal) inference, rather than because of the domain-specific representation. The only domain-specific knowledge in the model is in the assumption that the syntactic forms corresponding to verb constructions are already known by the learner to be the features that inference must operate over. Actually learning that, say, NP_1 V NP_3 NP_2 is a construction is itself a complex learning problem, and not one that this work bears on. But a significant part of the debate about verb construction learning is about the issue I focus on in Study 2 and Study 3: how to generalize properly over

the constructions seen in the input once it is clear what they are.

I do not claim that all aspects of verb learning (or even most of the interesting ones) can be accounted for by a model that makes the sparse representational assumptions this one does. In one sense this sort of representation is an advantage, because it clarifies and highlights precisely what aspects of learning derive from the Bayesian paradigm and which derive from the specific representation; and it allows us to explore the contribution of additional semantic features in the abstract, without worrying about their accessibility or precisely what they are. The tradeoff, of course, is that we are therefore abstracting over many details that might be critical for understanding the acquisition of particular verbs or asking the question of what particular semantic, conceptual, or syntactic knowledge must be built in order for the child to perceive which features are important and relevant.

Although Study 2 may appear to imply that some aspects of verb construction learning could be accomplished without semantic information, I am not suggesting that semantics are not an important part of learning verbs. Indeed, Study 3 was motivated by the fact that verb learning must include semantic as well as syntactic knowledge. We found that it is necessary to include some sort of semantic representation in order to capture the sort of generalization captured by the experiments of Gropen et al. (1989). Interestingly, although I did not build in innate semantic-syntactic rules of the sort theorized by both the semantic and syntactic bootstrapping hypotheses, as long as the model could form classes of verbs it was capable of capturing relationships between syntax and semantic class. I am hesitant to draw any strong conclusions from this result, but it does suggest the utility of further work in this direction, exploring to what extent innate linking rules are required to abstract semantic-syntactic relationships when the data and representations are more complicated and realistic.

Another interesting possibility suggested by the analysis here is that perhaps syntactic information was so effective because I gave the model clean features that already picked out precisely the constructions of interest. If both the semantic and syntactic features available to the child are far more noisy – and hidden amongst

many irrelevant features in the environment – then it may be that semantic and syntactic features only become accessible through a process of mutual bootstrapping. One could even imagine a model in which one or a few extremely coherent or salient features provided information sufficient to pick out features that would otherwise be too incoherent or difficult to observe – as may happen for particularly abstract semantics, like those corresponding to verbs of mental state (Gleitman et al., 2005; Papafragou et al., 2007). This might capture results reported by Ambridge, Pine, Rowland, and Young (2008), which found that for younger children (age 5-6) there was only a small effect of semantic class on generalization, whereas for older children and adults, there was a larger and significant one. Similarly, Brooks et al. (1999) found that entrenchment effects (in syntax) emerged before semantic class effects. Addressing these ideas, or others that explore the roles of syntax and semantics more fully, will probably require a richer semantic representation than the current model instantiates. Future work will explore this idea in more detail.

Conclusion

This chapter extends the hierarchical Bayesian model introduced in Chapter 4 to explore several issues relevant to the acquisition of verb argument constructions. Although the model was originally developed to address the acquisition of feature biases in word learning, it can also explain how abstract learning about feature variability can be combined with verb-general learning on the level of construction-based verb classes and verb-specific learning on the level of individual lexical items. This work demonstrates clearly how the No Negative Evidence problem can be solved by a Bayesian learner due to its ability to balance simplicity and goodness-of-fit when evaluating hypotheses for generalization. Additionally, I demonstrate how this model can acquire the dative alternation based on syntactic-only input from actual child-directed speech while capturing the qualitative patterns exhibited by adults in an artificial language learning task and by children over the course of acquisition. Furthermore, it does so in a (largely) domain-general way, without making strong language-specific

representational assumptions about verb constructions.

In the next and final chapter I consider how the work in this chapter, in combination with that from Chapters 3 and 4, can shed light on the learnability problems motivating this thesis. What broader questions do each of these specific models address? What themes underlie each component of this research, and what general lessons can we infer from this work? Finally, where should we go in the future?

Chapter 6

Discussion

The main goal of this thesis was to explore three common learnability arguments by formalizing them in Bayesian terms. By focusing on these classic learnability arguments in the context of three major questions in language acquisition, I was able to explore the issues of representation and learnability from multiple angles, and to shed new light on these old questions.

In this chapter I step back and position this work in a larger context. What are its implications for those studying language acquisition and cognitive development? What are its major limitations? In an effort to make this discussion maximally accessible (particularly because I revisit many points that I have already addressed to some degree elsewhere) this chapter will be written in a more colloquial tone.

Simplicity/goodness-of-fit tradeoff

Q: You suggest that a Bayesian learner can actually *solve* the No Negative Evidence problem. How is this possible?

I answer this question abstractly in some detail in Chapter 2 and touch on the reasoning again in Chapters 3 and 5, but in a nutshell, the solution emerges because Bayesian inference trades off the complexity of a hypothesis (reflected in its prior probability) and how well that hypothesis fits to the observed data (reflected in the

likelihood). As a result, it tends to prefer hypotheses that – like B in Figure 2-2 – strike a balance between being overly simple (hence missing important details) and overly complex (hence overfitting). Because the importance of fitting the data increases with the number of datapoints, there is an interesting pattern as the amount of data increases. When there are only a few datapoints, Bayesian learners tend to favor simpler hypotheses, because the importance of simplicity outweighs the lack of fit to those few datapoints. But as the number of datapoints increases, fitting them badly matters more and more (since there are more of them, and each one is penalized). As the size of the dataset approaches infinity, a Bayesian learner rejects larger or more overgeneral hypotheses in favor of more precise ones, as the Subset Principle would do. But with limited amounts of data, the Bayesian approach can make more subtle predictions.

You can see how this would solve the No Negative Evidence problem: as the learner observes more and more data, he becomes less and less likely to think that something that was unobserved is nevertheless permissible. Essentially, the likelihood captures the notion of a suspicious coincidence. If a gumball machine gives me ten blue gumballs in a row, I'm going to be fairly certain that it only gives blue gumballs; it's at least more likely than thinking that somehow, coincidentally, I just happened to never get green or red or yellow ones. The nice thing about Bayesian inference is that it gives a *quantitative* way of determining when the coincidence grows too great. Why am I not likely to think that the machine gives only blue gumballs after just receiving one, but will think it does after ten? Being able to calculate the tradeoff between simplicity and likelihood allows us to answer this question. In essence, the likelihood in Bayesian learning is a principled quantitative measure of the weight of implicit negative evidence – one that explains both how and when overgeneralization should occur.

A side effect of this reasoning is that frequency becomes very important. Because quantity of data is the deciding factor in determining how overgeneral the hypothesis should be, a Bayesian learner overgeneralizes more when there is less data. This is what we saw in Chapter 3, where the 1-ST grammar (which accepts any sentence

as grammatical) was preferred on smaller corpora and the context-free grammars were preferred on larger ones. It also explains the frequency effects found in Chapter 5, where the low-frequency items tended to be overgeneralized much more than the higher-frequency ones.

Q: This doesn't sound all that different from people who argued that indirect negative evidence is the way to solve the No Negative Evidence problem. Isn't this just a version of the same thing? And isn't this, therefore, vulnerable to the same counter-arguments?

It is indeed a very similar idea. One major difference is simply that conceptualizing the notion of indirect negative evidence in Bayesian terms allows us to justify why a learner might deal with indirect negative evidence in this way. There is a lot of research explaining why Bayesian learning constitutes optimal inference; in essence, a non-Bayesian reasoner attempting to predict the future will always be out-predicted by a Bayesian reasoner in the long run (e.g., de Finetti, 1937; Jaynes, 2003). It is also hardly a new idea to point out that making use of a simplicity criterion can enable a learner to limit overgeneralization in the "right" way (Solomonoff, 1964, 1978; Rissanen & Ristad, 1992; Chater & Vitànyi, 2007). A byproduct of reasoning with such a criterion in mind, while also being concerned with goodness of fit, is becoming sensitive to indirect negative evidence – information about what *isn't* there, as well as what is.

The other major difference is that the Bayesian framework provides a method that offers quantitative and precise predictions, rather than simply a qualitative idea. Without quantitative details, it can be difficult to test or falsify specific claims. How much negative evidence is sufficient? How does this depend on the complexity of the representation and the nature of the evidence that *is* observed? Formalizing the idea of indirect negative evidence in Bayesian terms allows us to rigorously address these questions.

The main counter-argument to the idea of indirect negative evidence is that point that people need to already know the relevant dimensions of generalization in order

to be sensitive to it. I discussed this in Chapter 5, since Pinker's objection is a version of this: "under exactly what circumstances does a child conclude that a nonwitnessed sentence is ungrammatical? This is virtually a restatement of the original learning problem." (Pinker (1989), p. 14). How does the child know that those particular syntactic forms are the relevant ones? This is, indeed, a hard problem – but it is not the No Negative Evidence problem *per se*. The logic of the No Negative Evidence problem remains the same whether there is one possible dimension of generalization, or an infinite number: the dilemma comes because one can never be certain that an unobserved datapoint (along that dimension) is truly ungrammatical, or simply unobserved. Indeed, in the formulation by Gold (1967) there is only one dimension of relevance: which formal grammar is appropriate.

So, I agree that the simplicity/goodness-of-fit tradeoff doesn't solve the Feature Relevance problem. I hope that one virtue of my analysis has been to make clear that the Feature Relevance problem is a separate problem from the No Negative Evidence problem, which a Bayesian learner *can* overcome.

Q: But how can this be reconciled with the logic of the No Negative Evidence argument? I haven't seen you show where it is wrong.

You're right that I haven't argued against the logic of the No Negative Evidence problem at any point. It is true that without negative evidence one cannot logically rule out hypotheses that are consistent with, but supersets of, the observed data. If the hypothesis space is infinite – as it is in Gold's 1967 analysis, and as it is in many interesting problems in language acquisition – we cannot simply wait until seeing all of the data before concluding that the correct hypothesis is the most conservative one. It is impossible to deductively rule out all of the infinite numbers of consistent superset hypotheses on the basis of positive-only finite data. It was impossible at the beginning of this thesis, and it remains impossible now.

However, a Bayesian learner does not deductively rule out consistent superset hypotheses; it simply assigns them less and less probability as the amount of data increases. This effectively converts the logical problem into a probabilistic problem;

and while a learner may never be 100% confident that it has converged to the correct hypothesis, it can quickly get to the point that the probability of other hypotheses is vanishingly small.

So this isn't a logical solution. It is, however, a solution in the sense that it offers an explanation for how an actual learner might deal with and overcome the problem *in practice*. This ties back into a point I made first in Chapter 1, when I talked about the importance of exploring how an ideal learner might handle real datasets. This sort of analysis, instead of focusing on all of the many ways that some abstract learning problem is impossible in theory, reframes to ask what *can* be learned in a more realistic situation, or what sort of learner can deal with realistic data in a reasonable way. Since we're ultimately interested in how human learners *actually* learn from data, this seems like an important question.

Q: You say that Bayesian learning solves the No Negative Evidence problem because it instantiates the tradeoff between simplicity and goodness-of-fit. But isn't "simplicity" a completely arbitrary metric?

Well, no. As I discuss in Chapter 2, the definition of simplicity emerges naturally from the generative assumptions underlying the Bayesian framework, in which hypotheses are themselves generated from a space of candidate hypotheses. If we conceptualize hypotheses as being generated through a series of choices – as in the dot examples in Chapter 2, or the grammars in Chapter 3 – then simpler hypotheses are those that require fewer "choice points." This is an intuitively sensible notion of simplicity. I think it's safe to say that we would all agree that the more complicated something is, the easier it is to mess it up: there are more choices you have to make, more places where you could zig instead of zag.

The precise prior probability of a hypotheses is therefore not arbitrarily assigned, but rather falls out in a principled way from how the hypotheses are generated. In Chapter 2 I mentioned the idea of a "generative model for recipes" whose "primitives" are recipe ingredients and measurements. The analysis in Chapter 3 assumes a generative model for grammars whose "primitives" are terminals, non-terminals, and

rules (which are built from the terminals and non-terminals). Though I didn't discuss the models in Chapter 4 and 5 explicitly using this terminology, their primitives are parameters and hyperparameters of distributions. In every case, the logic favoring simple hypotheses is the same: multiple choices are *a priori* less likely than a few.

I have one important caveat. While is always true that *within the set of choices defined by the generative model* the hypothesis that requires fewer choices has higher probability, one could always change the primitives of the model to design some Bizarro World model that looks to us as though it assigns higher probability to more complicated-looking hypotheses. For instance, if I defined a generative model for grammars whose primitives were specific 50-rule grammars that could be added to and subtracted from in complicated ways, it might look like one of those grammars was *a priori* "simpler" than a grammar with one rule only. Simplicity is only meaningful relative to the set of primitives out of which the hypotheses are generated, and in that sense any simplicity metric is indeed arbitrary. But this is true for any computational model (in which the primitives define the space of possible behavior) or, indeed, any theory (where the primitives assumed about the learner or the task implicitly define the solutions that can be found). This doesn't mean we should all retire and go into management consulting or stock-car racing; it means that we should be aware that one's choice of primitives, whether implicit or explicit, counts among the assumptions made by the theory. Whether those are the *appropriate* primitives is an empirical question, and interesting theoretical work might be done showing that with a different choice of primitives, very different results occur. But for the most part, unless the primitives are *very* different, the assignment of simplicity will often produce broadly similar results across the board. And either way, it offers a solution to the No Negative Evidence problem.

Q: Any learner is only as good as its assumptions; there is no such thing as a free lunch. What assumptions does this sort of Bayesian reasoning depend on, and in what situations will those assumptions lead the learner astray?

Implicit in all of the models in this thesis, and most Bayesian models more generally, is an assumption that data is generated based on a process in which each datapoint is sampled independently of one another. This is not an in-principle assumption that Bayesian models *must* make – one could make any well-defined sampling assumptions one wished – but assuming independence makes the mathematics considerably easier. If the data is not actually independently generated, a Bayesian learner assuming that it is may infer the wrong model.

A similar situation would arise if for some reason data were systematically (rather than randomly) excluded from the observed evidence in the world (e.g., due to performance limitations not captured in the assumed generative process). All other things being equal, a Bayesian learner faced with this dataset might (wrongly) infer that the generalization that would produce the missing data is not permitted.

In both of these cases, the problem emerges because the assumptions made about the generative process are not accurate. In general, the performance of a Bayesian learner depends in an important way on these assumptions. Since the generative process also defines the hypothesis space, if the correct solution does not lie in the hypothesis space, the learner will not arrive at it. This point may appear trivial, but it highlights the importance of defining the generative process, hypothesis space, and representational capacity with care. (Of course, even if one's hypothesis space excludes the "correct" hypothesis – as is probably true for all of the analyses presented in this thesis, due to their simplicity – the analysis can still be a useful way of comparing the theories that *are* in the hypothesis space).

Ideal learner, real dataset

Q: What is the point of exploring the performance of an ideal learner, given that we're interested in actual human cognition?

This is a common issue, and I address it throughout this thesis, especially in Chapters 1, 2, and 3. Ideal learnability arguments focus on the question of whether something is learnable in principle, given the data. By describing the characteristics of an ideal learner and then evaluating whether it could acquire the knowledge in question, these analyses suggest a series of testable questions and are a positive way of guiding further inquiry. To what extent do humans really have the hypothesized characteristics? Are the assumptions made about the data accurate? Do the ways in which the analysis departs from reality prove critically important? If so, why? If not, why not?

Whether or not people are ideal learners themselves, understanding how humans *do* reason can often be made if one can identify the ways in which people depart from the ideal: this is approximately the methodology by which Kahneman and Tversky derived many of their famous heuristics and biases. Bayesian modelling is an implementation of scientific inquiry that operates on Marr's third (computational) level, which seeks to understand cognition based on what its goal is, why that goal would be appropriate, and the constraints on achieving that goal, rather than precisely how it is implemented algorithmically (Marr, 1982). Understanding at this level is important because the nature of the reasoning may often depend more on the learner's goals and constraints than it does on the particular implementation. This approach 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.

Q: In what way is Bayesian learning "optimal", anyway? That sounds like a value judgment rather than a scientific claim.

As I note in Chapter 2, Bayesian inference is optimal in the sense that a non-Bayesian reasoner attempting to predict the future will always be out-predicted by a Bayesian

reasoner in the long run (de Finetti, 1937). It is essentially an extension of deductive logic to the case where propositions have degrees of truth or falsity (and is identical to deductive logic if we know all the propositions 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 (Jaynes, 2003).

If we were to try to come up with a set of desiderata that a system of "proper reasoning" should meet, they might include things like consistency and qualitative correspondence with common sense – if you see some data supporting a new proposition A, you should conclude that A is more plausible rather than less; the more you think A is true, the less you should think it is false; if a conclusion can be reasoned 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). Bayesian inference has also been shown to correspond to common-sense reasoning and the scientific method (Jeffreys, 1931, 1939; de Finetti, 1974; Jaynes, 2003).

Q: Okay, you've convinced me that an ideal learning analysis is worthwhile. But what's the point of studying realistic datasets? It seems like you lose your ability to make any general claims if you do that.

Yes, there is an inherent tradeoff here. One of the real virtues of ideal learnability analyses on ideal datasets, as we saw in Chapter 1, is that they are more general and not tied to the vagaries or idiosyncracies of any particular dataset. At the same time, they are only a first step (at best) for anybody interested in human cognition. If the ultimate goal of cognitive science is to understand how realistic learners use realistic data, then – given that we don't yet understand how realistic learners do their thing – exploring how an ideal learner can use realistic data is an important next step.

This is not just a step forward; it also reframes the questions in a useful way. Instead of asking what is learnable in the limit (or not), we can focus on what learners actually *could* learn – and how different representational or other assumptions affect that. As an example, in Chapter 1 I mentioned research showing that making

the assumption that grammars are probabilistic does not actually expand the class of learnable languages (Angluin, 1988; P. Niyogi, 2006). Positive learnability results emerge only if the probability distribution on rules μ comes from a family of approximately uniformly computable distributions. The probability distribution associated with the class of context-free grammars follows an exponential-decay distribution; a learner who (correctly) assumes that the distribution is of this form will converge to the correct context-free grammar (Horning, 1969), but a learner assuming an arbitrary distribution may not. Since it is fairly clear that natural languages are more expressive than context-free grammars, but that the actual family of distributions (if there is one) corresponding to natural language grammars is unknown, this has widely been taken as further evidence for the in-principle unlearnability of natural languages without further assumptions.

The analysis in Chapter 3 appears on the surface to contradict this result by suggesting that some aspects of syntactic knowledge previously argued to be innate may actually be learnable. These findings are reconcilable because I used an actual, realistic dataset. The Bayesian learner there actually *does* assume the exponential-decay distribution of context-free grammars – and, even though the resulting CFG only roughly approximated English, enough was learned to support abstract inferences previously argued to be impossible. This result depending on performing computations over sentence types rather than tokens, but this is not a drawback – it is a prediction, and offers a new way of conceptualizing the learning problem. Even if learning entire families of languages is impossible without knowing the full nature of the distribution in advance, it may nevertheless be true that making certain distributional assumptions results in interesting learning of *actual* natural languages.[1]

This example illustrates that in a sense, ideal learnability analyses on ideal datasets

[1]As an aside, I can't resist noting that it is also possible that people *do* make the correct distributional assumption about the grammars underlying human language, and that is why human languages are learnable. This distribution μ, whatever it is, might not be exponential-decay (or something else corresponding to the classes of grammars picked out by formal syntactic theories), but it might neatly pick out exactly the class of natural languages. Indeed, some explanation along these lines is almost inevitable – cf. Terence Deacon (1997), if people were not making the assumptions necessary to learn human languages, languages would not have been learned and passed on through the generations.

run into the danger of "letting the perfect be the enemy of the good." Focusing on realistic datasets allows us to explore what *is* learnable from the sort of data that people actually are presented with, and what assumptions guide that learning.

Q: These datasets don't actually seem very realistic to me, though.

At our current state of understanding, it is impossible to have a completely realistic dataset – even videotaped evidence needs to be encoded in some sort of usable form, and may not reflect exactly what the child was attending to or noticing. All of the choices about encoding and preprocessing, as I discuss in Chapter 2, constitute lapses from perfect realism; but making these sorts of choices, simplifying the dataset so that models can do something sensible and interpretable with it, is something that every modeler must do. In fact, the point of modelling *is* to simplify, and to hope that we have done so in such a way that it adds to our understanding rather than detracts.

The important question is whether the preprocessing choices I made are such that they render the conclusions invalid or uninteresting. In each of Chapters 3, 4, and 5, I devoted some time to evaluating the datasets with respect to this criterion. In each case, while the preprocessing of the datasets limits the range and generality of some of the results, the choices made were not indefensible, nor did they presuppose the conclusions that I do draw.

Q: Ultimately, of course, we would like to understand how real learners, not just ideal ones, might deal with real data. How can we move from the analyses in this thesis toward this direction?

This is a very good question. I can foresee two major steps, and many other researchers are already moving in this direction. One would be to combine this sort of ideal learning model with experiments on adults and children, with the aim of exploring in what ways people diverge from optimality. Done properly, this sort of analysis requires experimentation and modelling to be tightly yoked, since it is most interesting to not simply note where this divergence occurs, but to be able to explain

it in modelling terms. For instance, what if adult learners *do* perform differently on the datasets in Chapter 4 based on whether they are given category information or not? This behavior is not what is found in the ideal learning model, and the next interesting question would be to characterize exactly why. Do people have memory or attentional biases or deficits? Do they have perceptual or cognitive biases that make them systematically ignore some possibilities, or overweight others? Do they use heuristics (like, say, weighting features by salience) that drive their behavior?

This ties into the other step, which is modifying these ideal learning models to be less ideal in systematic ways. It would not be difficult to impose memory or processing constraints on Bayesian models, either by adding noise to the input or "forgetting" it in consistent ways. Different assumptions about feature salience, biases that exclude certain possibilities from being considered, and other heuristics can also be added. The advantage of starting with an optimal learner and adding limitations is that it makes clear what aspects of the behavior are due to the limitations. By yoking this sort of modelling with experimental work on subjects, we can significantly broaden the set of questions we can address.

Learning on multiple levels of abstraction

Q: You keep mentioning learning higher-order constraints even before the lower-level specific knowledge. I don't understand how this is possible.

To be clear, the claim is not that the higher-order constraints must necessarily be learned before any specific lower-level knowledge; it is that we cannot assume that they must be learned later. Depending on the nature of the task and the dataset, either may be faster than the other (Kemp et al., 2007). In Chapter 3 we saw a situation in which the higher-order abstraction T (that the corpus of child-directed speech is better captured by a grammar with hierarchical phrase structure) was acquired before honing on the precisely correct lower-level grammar G: at both *Epochs* and *Levels* prior to the full corpus, CFGs were preferred even though the favored specific grammar was different than the best one on the full corpus. Though that model 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 fact, in Chapter 4, we *do* see an instance where higher-order inferences act to constrain inferences at a lower level. That is what the shape bias is. And, as we discovered, this higher-order generalization about feature variability emerges as soon as there is enough lower-level information to form categories. It is not the case that the model must wait until acquiring all of the (lower-level) category information before it is capable of forming generalizations on a higher level.

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. If there are many fewer possibilities at a higher level than at a lower level, less data may be required to draw conclusions at the higher level. For instance, 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 higher levels of abstraction consist of inferences about an entire hypothesis, while lower levels may be relevant to only a small part, there is in a sense much more data available about these higher levels of abstraction. For instance, the higher level T of grammatical knowledge affects the grammar of the language as a whole while any component of G affects only a small subset of the language produced. A single sentence like *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. In a similar way, every item in the world provides some evidence about whether categories in general are organized by shape or not, but only items from a specific category provide evidence about that categories. Higher-order generalizations may thus be learned faster simply because there is much more evidence relevant to them.

Because of the possibility of this sort of learning, it is important to remember that children learn individual phenomena as a part of a system of knowledge. As with auxiliary fronting, most PoS arguments consider some isolated linguistic phenomenon that children appear to master and conclude that because there is not enough evidence for that phenomenon in isolation, it must be innate. I have suggested here that even when the data does not appear to explain an isolated inference, there may be enough evidence to learn a larger system of linguistic knowledge – a whole grammar – of which the isolated inference is a part. 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. As long as higher-level learning is possible, it becomes necessary to explore what kinds of system-wide inferences can be acquired.

Q: Even if higher-level learning is possible, it doesn't seem to have especially interesting implications. After all, you haven't removed the need for innate knowledge; you've just moved it up a level.

It is definitely true that none of this analysis removes the need for innate knowledge; I am not trying to argue against that. But "moving it up a level" accomplishes several things. First, as the knowledge gets increasingly abstract, it also gets increasingly simple and general. So this framework does not do away with the need for innate

knowledge, but it may change the *nature* of that knowledge. It raises the possibility that a learner could make domain-specific inferences on higher levels of abstraction, based on innate knowledge on even higher levels that is vague or general enough that it is plausibly domain general. The work is not meant to show that this is always true, but it illustrates that the simple existence of an observed higher-level inference is not reason to conclude that it is innate; what *is* innate may be quite different.

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 knowledge, 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, this work offers an alternative, by providing just such a learning mechanism in the form of hierarchical Bayesian models. If an abstract generalization can be acquired very early and can function as a constraint on later development of specific knowledge, 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.

Q: Are you making a general argument against innateness here?

No.

Q: Can you expand on that answer?

As I mentioned in the first few pages of this thesis, the simplistic distinction between nature and nurture is not a very constructive or interesting distinction from a scientific point of view. Very few things in cognition are entirely due to one or the other; indeed, feedback loops and non-linear interactions over the course of development make the distinction almost meaningless. Even if that weren't true, though, the interesting question often turns on the question of domain specificity rather than innateness. Chomsky's argument from the Poverty of the Stimulus, considered in Chapter 3, was used to argue for the presence of a particular kind of innate language-specific knowledge about grammar; as we saw in Chapter 5, Baker's Paradox was used by Pinker to argue for the presence of particular innate language-specific semantic linking rules and classes; and the debate in Chapter 4 critically turned on the question of what language-specific assumptions children make about word labels.

Throughout this thesis, I formalized these debates in Bayesian terms, with the goal of clarifying what innate knowledge must be assumed, and what form that knowledge must take. In Chapter 3 I argued that the particular domains-specific knowledge that has been argued to be innate – the knowledge that language has hierarchical phrase structure – need not be: it might, in fact, be derived from more general-purpose representational capacities and inductive biases. The learner I considered incorporates a mechanism that trades off simplicity and goodness-of-fit when evaluating hypotheses, performs inference over type rather than token data, and encodes sentences according to their syntactic categories. The first two items are domain-general biases, and the last is either learnable via domain-general mechanisms or is at least not the precise language-specific assumption that was previously argued to be innate.

Chapter 5 was similar in many ways. There I argued, in contrast to Pinker (1989), that innate semantic categories and linking assumptions are not necessary to solve Baker's Paradox. My model does make certain assumptions, most critically in the preprocessing, since it "knows" automatically what syntactic forms correspond to verb constructions (e.g., that NP_1 V NP_3 NP_2 is a construction). This is domain-

specific, but – as before – could plausibly have been learned, and in any case is not the precise language-specific assumption previously argued to be necessary.

One finding in Chapter 4 is that making strong innate assumptions about the role of words does not yield differences in the nature or extent of second-order generalization. This does not necessarily mean that those strong innate assumptions might not be present for other reasons. It is also possible that a less ideal learner might need to make certain innate assumptions about words, even if they are not necessary for an ideal learner.

In all of these cases, I'm not trying to make a claim that there is no innate knowledge, nor even that there is no innate domain-specific knowledge; I'm simply interested in using this paradigm to be able to specify precisely what happens when we make different assumptions about what is built in (or not), or domain-specific (or not).

Q: How do you identify what is domain-specific and what isn't, and why does this matter?

The reason it matters is that few if any cognitive scientists believe that there is literally *nothing* that is innate; many debates about innateness tend to boil down to the question of whether the capacity in question is domain-general or not. For instance, one way of framing the issue of the learnability of hierarchical phrase structure discussed in Chapter 3 is that it turns on the question of whether the innate knowledge consists of (a) a domain-general learning mechanism (Bayesian inference) in combination with domain-general biases (to evaluate theories on the basis of type-based input) and representational capacities (grammars), or (b) a language-specific constraint that eliminates non-hierarchical grammars from consideration.

Any answer to this question will have two components: first, identifying which abilities or propensities are necessary to explain acquisition; and second, identifying whether those abilities or propensities are domain-general or not. The latter component is more difficult than one might think: while some are obvious, others are not. Is the bias to evaluate grammars on the basis of type-based input domain-general or

domain-specific? I suggest here that it may be domain-general, since one can imagine may domains where it would be valuable to assume a two-component generative process analogous to that captured by the adaptor grammar framework, in which one component captures the "deep knowledge", and the other captures the more superficial or performance-based factors that affect how that deep knowledge is used to generate data. However, it is possible that this bias is domain-specific. Ultimately, many of these questions are only resolvable empirically. In many ways the resolution of the question of domain specificity is far less important than the identification of the bias in the first place – what matters is that it exists (or not) – and I am mostly concerned with the latter.

The Bayesian paradigm

Q: What's the big deal about Bayesian learning?

It's not that Bayesian learning is the be-all and end-all of cognitive science, but the paradigm is uniquely well-suited to address questions of learnability. As discussed in Chapter 2 and depicted schematically in Figure 6-1, it offers a unique way to study how two fundamental questions in cognitive science interact. The question of whether human learners have (innate) language-specific knowledge is logically separable from the question of whether and to what extent human linguistic knowledge is based on structured representations. In practice, however, these issues are often conflated. Within cognitive science, recent computational models of how language might be learned have usually assumed that domain-general learning operates on representations without explicit structure (e.g., Elman et al., 1996; Rumelhart & McClelland, 1986; Reali & Christiansen, 2005). The main proponents of innate domain-specific factors, on the other hand, have typically assumed that the representations involved are structured (e.g., Chomsky, 1965, 1980; Pinker, 1984). Few cognitive scientists have explored the possibility that explicitly structured mental representations might be learned via domain-general mechanisms. This framework offers a way to explore this relatively uncharted territory in the context of language acquisition.

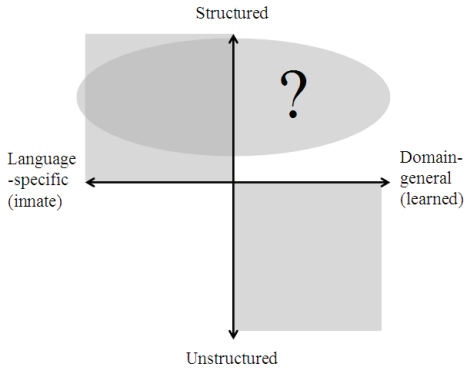


Figure 6-1: A schematic representation of the theoretical landscape for language acquisition in cognitive science. The vertical axis reflects the nature of the representation. The horizontal axis reflects the source of inductive bias: "innate" and "learned" are in parentheses because they are often conflated with "language-specific" and "domain-general", which I suggest is closer to the real issue. The two most prominent views are represented by the two opposite shaded quadrants. I explore another area, represented by the shaded oval. (The oval covers both sides of the horizontal axis because this approach explores both: it is in principle possible that it could yield results suggesting that the particular innate language-specific bias in question is necessary).

In addition, Bayesian modelling offers a normative framework for rational inference. 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 questions considered here. 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, it 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.

Q: How can a statistical learner perform inference as rapidly as these models do, on the basis of relatively few examples? I thought statistical learning was slow and gradual.

This is a common confusion. Statistical learners like associationist or connectionist models, which operate through a process of incremental updating in response to data, generally produce a pattern of learning that is more gradual or slow. But this pattern is a result of the update rule, not the nature of statistical reasoning in general. As we saw in Chapter 4, a Bayesian learner can actually make relatively strong inferences on the basis of only a few datapoints. This is a byproduct of measuring likelihood with respect to the size of the hypothesis space. Since each datapoint is presumed to be independent, the likelihood of seeing n particular datapoints increases exponentially by the size of the space. Statistical learning isn't always slow.

Q: Doesn't the Bayesian paradigm build the correct answer into the hypothesis space? How can this be considered learning? Isn't it actually a pretty strong innate assumption?

I focus on this issue in some detail in Chapter 2, and I will not reiterate the entire answer since it verges on territory that is kind of complicated and deep. The essential point centers on the distinction between giving the model the answer and having the answer be latent in the hypothesis space. As an example, imagine I asked you to guess an integer. At this point you haven't been given the answer in any meaningful sense, even though you've been given the hypothesis space: the set of all integers. In the sense that the answer is latent in the hypothesis space, you've been given it, but I consider it uncontroversial that a learner should be able to represent and get to the hypotheses it ultimately should believe. The interesting part of learning is how the learner figures out which hypothesis is correct, out of the (potentially infinite) hypotheses it *could* represent.

This just touches the surface of a more complicated and interesting discussion, and for that I refer you to Chapter 2.

Q: You're being very rah-rah about the Bayesian approach. Is there anything it *doesn't* do well?

As I mentioned earlier, I am certainly not arguing that the Bayesian framework should supplant other methods in cognitive science. It is an essential tool in the toolbox, but there are many other tools as well.

Trivially, one thing the Bayesian approach doesn't do well is tell us what actual humans do. For that, nothing can beat experimental and observational studies. I've argued that yoking modelling efforts with experimental data can be especially useful, but experimental and observational work is valuable, period.

As for other computational methods, different methods have different strengths and weaknesses. I have argued that Bayesian models are particularly appropriate for learnability issues, for all of the reasons discussed before. But if you're interested in, for instance, studying how different regions of the brain (like the hippocampus and the neocortex) might play complementary roles in memory and learning, then connectionist models might be more appropriate (e.g., McClelland, McNaughton, & O'Reilly, 1995). They can also be useful for studying how semantic information can be stored in a distributed manner. Dynamical systems models are appropriate for studying feedback loops and non-linearities over development. Which sort of approach you should use depends on what question you're asking.

Conclusion

Q: You've talked about future directions for some of the specific topics in their respective chapters. But what do you think are the most interesting unexplored areas that lie ahead in a more general sense?

One very exciting topic, which I've mentioned before at some length (and so will not belabor too much here) is the possibility of extending Bayesian models to capture different ways of being non-ideal. Although I think there are still many, many questions that pure ideal learning analyses can tell us a lot about, I mention this direction

because I think it is qualitatively different, and opens up a lot of future directions.

Another exciting direction is that with Bayesian models I think we can finally start to address one of the deepest, most unresolved questions in all of cognitive science: where we get our hypotheses and representations in the first place. I discuss this question at length from a more theoretical perspective in Chapter 2, but I think recent developments in nonparametric Bayesian models can help us address them computationally as well. These models, called "infinite" because they assume that the true underlying dimensionality is unbounded, can dynamically learn to create additional features as learning progresses (e.g., Ghahramani, Griffiths, & Sollich, 2006). Can this sort of model be used to explain how people themselves infer which features are important for categorization or reasoning? Might we even extend the basic idea to explore how hypothesis spaces themselves can grow?

Q: What do you think are the biggest limitations of your analyses?

I knew you'd ask that. Here is a handy list:

Chapter 3: The computational problem of searching the space of context-free grammars (and, to some extent, regular grammars) is so difficult that I was only able to approximate a full search. I used many different methods for doing this approximation – creating linguistically motivated hand-designed grammars, performing a local search using them as the starting point, and performing an automatic search from scratch – but until we can develop a procedure that we can be sure is adequately searching the entire space, my conclusions can only be preliminary.

Chapter 4: It is extraordinarily difficult to know how to characterize the nature of objects in categories in the world, and therefore how to map the datasets I considered there onto the input children receive. I addressed this difficulty by systematically varying the coherence of the features and the number of items and categories in the datasets, but much remains to be done here. What if the organization of features is entirely different? What about larger datasets? What

if the choice to represent items as sets of features itself is completely off base? We have to start from where we are – and most models and theories of categorization have similar difficulties – but this is nevertheless a major shortcoming.

Chapter 5: The major problem here is similar to the problem in Chapter 4: the limitations of the representation. Representing knowledge about verbs as counts of occurrences in different syntactic constructions has one advantage, because it demonstrates how many phenomena in verb learning can be captured by such a simple model; but it is nevertheless quite limited. Many of the interesting questions in the verb acquisition literature center on the nature of the semantic representations, which this model can say very little on (although it is somewhat interesting that, again, it can qualitatively replicate some phenomena even making these sparse assumptions). Future work in this area will probably depend on expanding its representational capacity in some way.

Q: What do you think are the most important take-home messages?

This work demonstrates that one of the learnability problems, the No Negative Evidence problem, is solved by a Bayesian learner – or, indeed, by any learner that trades off between simplicity and goodness-of-fit. In Chapter 2 I explained in abstract terms why this occurs, and in the three subsequent chapters – particularly Chapter 5 – we saw how this principle plays out in several different domains.

One common thread running through all of the chapters is that higher-level abstract constraints may be learned at least as rapidly as lower-level specific information, early enough to act to constrain generalization at the lower-level. This was especially apparent in Chapters 3 and 4, and as I noted then, may in general imply that simply noting the presence of an early higher-order constraint is not sufficient motivation to assume that it is innate.

A related implication is that abstract knowledge or higher-order constraints may be supported by an entire *system* of data, not just individual items. This may serve as a cautionary warning of the perils of concluding that there is no evidence for some higher-level inference, based on the observed lack of some individual type of data; it

might be that system-wide statistics of the entire dataset would support the inference, as we see occur in Chapter 3.

In general, this research offers a new perspective about what we can and can't conclude from standard learnability arguments, and provides a framework that allows for an increasingly subtle, detailed, and rigorous analysis of claims about what must or must not be innate (as well as what that even means). The framework allows us to evaluate the bounds of "optimal" reasoning while also being able to vary with precision how this might depend on the data, the nature of the learner's mental representation, and both domain-specific as well as domain-general biases (whether due to the attentional, perceptual, memory-based, or learning mechanisms of the child). I suggest that this approach provides an essential tool as we move toward constructing a full and accurate picture of the human mind.

Appendix: Model details

Model details from Chapter 3

Searching the space of grammars

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

Search was inspired by work by Stolcke and Omohundro (1994), in which a space of grammars is searched via successive merging of productions; some sample merges are shown in Table 1. 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.

Table 1: Sample merges for context-free and regular grammars. Identical merges for right-hand side items were also considered.

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	

Prior probabilities

Non-terminals, productions, and items

The probabilities of the number of non-terminals $p(n)$, productions $p(P)$, and items $p(N_i)$ are modelled 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). 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 1 and illustrated in Figure -2.

$$p(1-p)^{n-1}. \tag{1}$$

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.

Production-probability parameters

Because each θ_k corresponds to the production-probability parameters for non-terminal k, the individual parameters θ_1,...,θ_m in each vector θ_k should sum to one. As is stan-

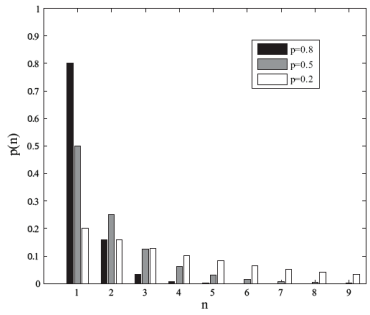


Figure -2: A geometric distribution, with three possible values of p.

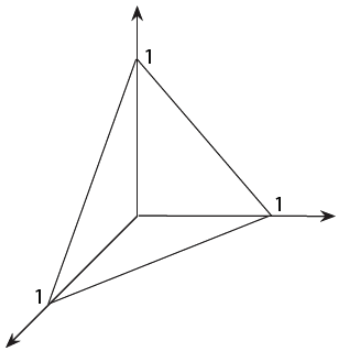


Figure -3: The unit simplex for $m_k = 3$ (a triangle), corresponding to the Dirichlet distribution with $\alpha = 1$ on a θ_k vector of production-probability parameters with three productions.

dard 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 θ_1,...,θ_m, given that each production has been used $\alpha - 1$ times. We set $\alpha = 1$, which is equivalent to having never observed any sentences and not assuming *a priori* that any one sentence or derivation is more likely than another. This therefore puts a uniform distribution on production-probability parameters and captures the assumption that any set of parameters is as likely as any other set. In general, drawing samples from a Dirichlet distribution with $\alpha = 1$ is equivalent to drawing samples uniformly at random from the $m_k - 1$ unit simplex; the simplex (distribution) for $m_k = 3$ is shown in Figure -3.

The Dirichlet distribution is continuous, which means that the probability of any

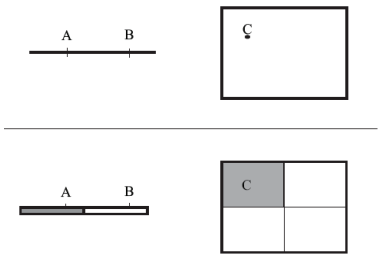


Figure -4: Top: one cannot compare the probability of continuous points A and C with different dimensionality. Bottom: when A and C correspond to discrete points, comparison is possible.

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 -4. 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 θ_1,...,θ_m, even though the Dirichlet distribution is continuous.

However, one cannot validly compare the relative probability of choosing points from sets with different dimensions, as in A and C in Figure -4. 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 -4, 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 θ_1,...,θ_m from a Dirichlet distribution is equivalent to drawing samples from the $m_k - 1$ unit simplex, we calculate their probability by dividing the simplex into identically-sized pieces. Any $m - 1$ simplex can be subdivided into g^{m-1} simplices of the same volume, where g is the number of subdivisions (grids) along each dimension (Edelsbrunner & Grayson, 2000). If $\alpha = 1$, all grids are *a priori* equally probable; thus, $p(\theta_k)$ is given by the volume of one grid divided by the volume of the entire simplex, that is, $\frac{1}{g^{m-1}}$. 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. This tradeoff is captured by scoring g like we do other choices. We assign a possible distribution of grid sizes over g by assuming that $ln(g)$ is distributed geometrically with parameter $p = 0.5$. Thus, smaller g has higher prior probability, and we can select the grid size that best maximizes the tradeoff between simplicity and goodness-of-fit. We evaluated each grammar with g = 1, 10, 100, 1000, and 10000. The results reported use $g = 1000$ because that is the value that maximizes the posterior probability for all grammars; the hierarchical grammar type was preferred for all values of g.

Additional complexities involved in scoring prior probability

Depending on the type of the grammar, some specific probabilities vary. The flat grammar has no non-terminals (aside from S) and thus its $p(n)$ is always equal to 1.0. Both the regular and context-free grammars, written in Chomsky Normal Form to

conform to standard linguistic usage, are constrained to either have one or two items on the right hand side. The regular grammars have further type-specific restrictions on what kind of item (terminal or non-terminal) may appear where, which effectively increase their prior probability relative to context-free grammars. These restrictions affect $p(N_i)$ as well as the effective vocabulary size V for specific items. For example, the first item on the right-hand side of productions in a regular grammar is constrained to be a terminal item; the effective V at that location is therefore smaller. A context-free grammar has no such restrictions.

Model details from Chapter 4

Model L2: Learns overhypotheses at Level 2

Model L2, also specified in Kemp et al. (2007), is known to statisticians as a Dirichlet-Multinomial model (A. Gelman et al., 2004), and can be written as:

$$\begin{aligned}
\alpha &\sim \text{Exponential}(\lambda) \\
\boldsymbol{\beta} &\sim \text{Dirichlet}(\boldsymbol{\mu}) \\
\boldsymbol{\theta^i} &\sim \text{Dirichlet}(\alpha\boldsymbol{\beta}) \\
\boldsymbol{y^i} \,|\, n^i &\sim \text{Multinomial}(\boldsymbol{\theta^i})
\end{aligned}$$

where n^i is the number of observations for category i. Because the model is learning only Level 2 knowledge, I specify the Level 3 knowledge by setting parameters $\lambda = 1$ and $\boldsymbol{\mu} = \mathbf{1}$, which indicates weak prior knowledge over both α and $\boldsymbol{\beta}$.

This notation is ambiguous about whether categories may have multiple features or not. If there are multiple features, inference proceeds exactly as before except that each α and $\boldsymbol{\beta}$ is learned separately for each feature. The posterior probability for the full model is therefore the product of the probabilities along each feature, which corresponds to the assumption that each feature is independently generated. When working with multiple features, I will use α to refer to the collection of α values along

all features, and $\boldsymbol{\beta}$ for the set of all $\boldsymbol{\beta}$ counts along all features.

Inference is performed by computing posterior distributions over the unknown knowledge at the higher levels. For instance, the posterior distribution $P(\alpha, \boldsymbol{\beta}|\boldsymbol{y})$ represents a belief given the data $\boldsymbol{y}$ (the categories seen so far). I formally instantiate this model by denoting the true distribution over features of category i as $\boldsymbol{\theta^i}$; thus, if (as in Figure 4-1) the true distribution of features means the category occurs 30% of the time with a round shape and 70% of the time with a square shape, then $\boldsymbol{\theta^i}$ = [0.3 0.7]. If we have observed that one item in the category is round and four are square, then $\boldsymbol{y^i}$ = [1 4]. I assume that $\boldsymbol{y^i}$ is drawn from a multinomial distribution with parameter $\boldsymbol{\theta^i}$, which means that the observations of the category are drawn independently at random from the true distribution of category i. The vectors $\boldsymbol{\theta^i}$ are drawn from a Dirichlet distribution parameterized by a scalar α and a vector $\boldsymbol{\beta}$: α determines the extent to which each category tends to be associated with only one feature, and $\boldsymbol{\beta}$ represents the distribution of features across all categories.

To fit the model to data I assume that counts $\boldsymbol{y}$ are observed for one or more categories. The goal is to compute the posterior distribution $P(\alpha, \boldsymbol{\beta}, \{\boldsymbol{\theta^i}\}|\boldsymbol{y})$. Inferences about α and $\boldsymbol{\beta}$ can be made by drawing a sample from $P(\alpha, \boldsymbol{\beta}|\boldsymbol{y})$ – the posterior distribution on $(\alpha, \boldsymbol{\beta})$ given the observed categories. Inferences about $\boldsymbol{\theta^i}$, the distribution of features for category i, can be made by integrating out α and $\boldsymbol{\beta}$:

$$P(\boldsymbol{\theta^i}|\boldsymbol{y}) = \int_{\alpha,\boldsymbol{\beta}} p(\boldsymbol{\theta^i}|\alpha, \boldsymbol{\beta}, \boldsymbol{y}) p(\alpha, \boldsymbol{\beta}|\boldsymbol{y}) d\alpha d\boldsymbol{\beta} \tag{2}$$

For most of the analyses in this thesis (in both Chapters 4 and 5), this is estimated using numerical integration via a Markov Chain Monte Carlo (MCMC) scheme. The sampler uses Gaussian proposals on $\log(\alpha)$, and proposals for $\boldsymbol{\beta}$ are drawn from a Dirichlet distribution with the current $\boldsymbol{\beta}$ as its mean. Four different runs are averaged to produce the graphs.

In order to speed up the runs in Study 2 and Extensions B and C of Study 1 in Chapter 4, $P(\boldsymbol{\theta^i}|\boldsymbol{y})$ there is estimated using a form of non-stochastic numerical integration in which the space of parameter values is discretized and evaluated directly. Because the datasets themselves, being simulated, are more variable than in other

sections, these graphs reflect the averages of 24 different runs rather than four.

Model extension: Learning category assignments

To add the ability to discover categories to Model L2, instead of assuming that the model is given data corresponding to counts of occurrences of categories $\boldsymbol{y}$, I assume that the model is given data corresponding to individual items j, from which it must infer the proper assignment of categories $\boldsymbol{y}$. As before, the true distribution over features of category i is denoted as $\boldsymbol{\theta^i}$, and the data corresponding to category i is $\boldsymbol{y^i}$. The difference now is that the model is only given data corresponding to individual items in the world, j, and it must decide which items belong in each category $\boldsymbol{y^i}$.

I assume that each category $\boldsymbol{y^i}$ corresponds to a partition of items, and represent this partition by a vector $\boldsymbol{z}$. The partition of the twelve items in Figure 4-1, where the first four items are cups, the next five are balls, and the last three are cats, would be represented by the vector [1 1 1 1 2 2 2 2 2 3 3 3]. The prior distribution on $\boldsymbol{z}$ is induced by the Chinese Restaurant Process:

$$P(z_j = c|z_1, \ldots, z_{j-1}) = \begin{cases} \frac{n^i}{j-1+\gamma} & n^i > 0 \\ \frac{\gamma}{j-1+\gamma} & i \text{ is a new category} \end{cases} \tag{3}$$

where z_j is the category assignment for item j, n^i is the number of items previously assigned to category i, and γ is a hyperparameter which captures the degree to which the process favors simpler category assignments (I set $\gamma = 1$). The Chinese Restaurant Process prefers to assign items to categories that already have many members, and therefore tends to prefer partitions with fewer categories.

If $\boldsymbol{z}$ is known, this model reduces to several independent versions of Model L2, and predictions can be computed using the techniques described in the last section. Since $\boldsymbol{z}$ is unknown, we integrate over each of the possible category partitions $\boldsymbol{z}$:

$$P(\boldsymbol{\theta^i}|\boldsymbol{y}) = \sum_{\boldsymbol{z}} P(\boldsymbol{\theta^i}|\boldsymbol{y}, \boldsymbol{z}) P(\boldsymbol{z}|\boldsymbol{y}) \tag{4}$$

where $P(\boldsymbol{z}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{z})P(\boldsymbol{z})$ and $P(\boldsymbol{z})$ is the prior induced by the CRP pro-

cess. For Model L2, computing $P(\boldsymbol{y}|\boldsymbol{z})$ reduces to the problem of computing several marginal likelihoods

$$P(\boldsymbol{y\prime}) = \int_{\alpha,\boldsymbol{\beta}} P(\boldsymbol{y\prime}|\alpha,\boldsymbol{\beta})P(\alpha,\boldsymbol{\beta})d\alpha d\boldsymbol{\beta} \tag{5}$$

which is estimated by drawing 10,000 samples from the prior $P(\alpha,\boldsymbol{\beta})$.

Model details from Chapter 5

This chapter considered two models. Model L2 is the model introduced in Chapter 4 and specified in Kemp et al. (2007) and earlier in this appendix. L3 performs inference over another level of overhypotheses, learning not only overhypotheses α and $\boldsymbol{\beta}$, but also overhypotheses on em those overhypotheses, denoted by parameters λ and μ, respectively.

Model L3: Learns overhypotheses at Level 2 and 3

The *Level 3* model is quite similar to the *Level 2* model, except that instead of assuming that λ and μ are known, we learn those as well. In statistical notation, it can be written[2]:

$$\begin{aligned}
\lambda &\sim \text{Exponential}(1) \\
\mu &\sim \text{Exponential}(1) \\
\alpha &\sim \text{Exponential}(\lambda) \\
\boldsymbol{\beta} &\sim \text{Dirichlet}(\boldsymbol{\mu}) \\
\boldsymbol{\theta^i} &\sim \text{Dirichlet}(\alpha\boldsymbol{\beta}) \\
\boldsymbol{y^i} \,|\, n^i &\sim \text{Multinomial}(\boldsymbol{\theta^i})
\end{aligned}$$

where n^i is the number of observations for verb i.

[2]Note that μ is a scalar, but in order for it to be a proper hyperparameter for the vector $\boldsymbol{\beta}$, it is vectorized: $\boldsymbol{\mu} = \mu\mathbf{1}$.

As before, inference is performed by computing posterior distributions over the unknown knowledge at the higher levels. The only difference is that the posterior distribution is given by $P(\lambda, \mu, \alpha, \boldsymbol{\beta}, \{\boldsymbol{\theta^i}\}|\boldsymbol{y})$. Inferences about λ,μ,α, and $\boldsymbol{\beta}$ can be made by drawing a sample from $P(\alpha, \boldsymbol{\beta}, \lambda, \mu|\boldsymbol{y})$, which is given by:

$$P(\alpha, \boldsymbol{\beta}, \lambda, \mu|\boldsymbol{y}) \propto P(\boldsymbol{y}|\alpha, \boldsymbol{\beta})P(\alpha|\lambda)P(\boldsymbol{\beta}|\mu)P(\lambda)P(\mu) \tag{6}$$

Inferences about $\boldsymbol{\theta^i}$, which in this case correspond to the distribution of constructions for verb i, can be made by integrating out α, $\boldsymbol{\beta}$, λ, and μ:

$$P(\boldsymbol{\theta^i}|\boldsymbol{y}) = \int_{\alpha,\boldsymbol{\beta},\lambda,\mu} P(\boldsymbol{\theta^i}|\alpha, \boldsymbol{\beta}, \boldsymbol{y})P(\alpha, \boldsymbol{\beta}, \lambda, \mu|\boldsymbol{y})d\alpha d\boldsymbol{\beta} d\lambda d\mu \tag{7}$$

This is estimated using numerical integration via a Markov Chain Monte Carlo (MCMC) scheme. The sampler uses Gaussian proposals on $\log(\alpha)$, $\log(\lambda)$, and $\log(\mu)$; as before, proposals for $\boldsymbol{\beta}$ are drawn from a Dirichlet distribution with the current $\boldsymbol{\beta}$ as its mean. Four different runs are averaged to produce the graphs, except in the case of Study 3 where, since the results were more variable, eight runs are included.

Model extension: Learning verb classes

The procedure for discovering verb classes is somewhat analogous to the procedure for learning item categories, introduced in Chapter 4 and explained elsewhere in this appendix. The difference is that learning item categories involves clustering on the level of specific items (Level 1), whereas learning verb classes (or, equivalently, ontological kinds) involves clustering on a more abstract level, Level 2. To add the ability to discover verb classes to both Models L2 and L3, I assume that verbs may be grouped into classes, each of which is associated with its own hyperparameters.[3] For Model L2, this means that there is a separate α^c and $\boldsymbol{\beta^c}$ for each class c inferred by the model; for Model L3, there is a separate α^c, $\boldsymbol{\beta^c}$, λ^c, and μ^c for each class c. In both cases, the model partitions the verbs into one or more classes. Each possible

[3]This extension for Model L2 is presented in Kemp et al. (2007); though it is referred to learning ontological kinds rather than learning verb classes, the model itself is equivalent. Here I extend the same idea to Model L3.

partition can be represented by a vector $\boldsymbol{z}$. A partition of six verbs in which the first three verbs are in one class and the last three were in another can be represented by the vector [1 1 1 2 2 2]. The prior distribution on $\boldsymbol{z}$ is induced by the Chinese Restaurant Process:

$$P(z_i = c|z_1, \ldots, z_{i-1}) = \begin{cases} \frac{n_c}{i-1+\gamma} & n_c > 0 \\ \frac{\gamma}{i-1+\gamma} & c \text{ is a new class} \end{cases} \tag{8}$$

where z_i is the class assignment for verb i, n_c is the number of verbs previously assigned to class c, and γ is a hyperparameter which captures the degree to which the process favors simpler class assignments (I set $\gamma = 1$). The Chinese Restaurant Process prefers to assign verbs to classes that already have many members, and therefore tends to prefer partitions with fewer classes.

The extension for Model L3 can now be written as follows:

$$\begin{aligned} \boldsymbol{z} &\sim \text{CRP}(\gamma) \\ \lambda^c &\sim \text{Exponential}(1) \\ \mu^c &\sim \text{Exponential}(1) \\ \alpha^c &\sim \text{Exponential}(\lambda^c) \\ \boldsymbol{\beta^c} &\sim \text{Dirichlet}(\boldsymbol{\mu^c}) \\ \boldsymbol{\theta^i} &\sim \text{Dirichlet}(\alpha^{c_i}\boldsymbol{\beta^{c_i}}) \\ \boldsymbol{y^i} \,|\, n^i &\sim \text{Multinomial}(\boldsymbol{\theta^i}) \end{aligned}$$

The equivalent extension for Model L2 is trivially derivable from this.

If $\boldsymbol{z}$ is known, the extended model reduces to several independent versions of the basic (L2 or L3) model, and predictions can be computed using the techniques described in the last section. Since $\boldsymbol{z}$ is unknown, we must integrate over each of the possible class partitions $\boldsymbol{z}$:

$$P(\boldsymbol{\theta^i}|\boldsymbol{y}) = \sum_{\boldsymbol{z}} P(\boldsymbol{\theta^i}|\boldsymbol{y}, \boldsymbol{z}) P(\boldsymbol{z}|\boldsymbol{y}) \tag{9}$$

where $P(\boldsymbol{z}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{z})P(\boldsymbol{z})$ and $P(\boldsymbol{z})$ is the prior induced by the CRP process. For Model L2, computing $P(\boldsymbol{y}|\boldsymbol{z})$ reduces to the problem of computing several marginal likelihoods

$$P(\boldsymbol{y\prime}) = \int_{\alpha,\boldsymbol{\beta}} P(\boldsymbol{y\prime}|\alpha,\boldsymbol{\beta})P(\alpha,\boldsymbol{\beta})d\alpha d\boldsymbol{\beta} \tag{10}$$

which I estimate by drawing 10,000 samples from the prior $P(\alpha,\boldsymbol{\beta})$.

For model L3, computing $P(\boldsymbol{y}|\boldsymbol{z})$ reduces to computing

$$P(\boldsymbol{y\prime}) = \int_{\alpha,\boldsymbol{\beta},\lambda,\mu} P(\boldsymbol{y\prime}|\alpha,\boldsymbol{\beta})P(\alpha,\boldsymbol{\beta}|\lambda,\mu)P(\lambda,\mu)d\alpha d\boldsymbol{\beta}d\lambda d\mu \tag{11}$$

which is also estimated by drawing 10,000 samples, this time from the joint prior $P(\alpha,\boldsymbol{\beta}|\lambda,\mu)P(\lambda,\mu)$.

Corpus of dative verbs

The data of verb counts is collected from the sentences spoken by adults in the Adam corpus (Brown, 1973) of the CHILDES database (MacWhinney, 2000), and consists of all instances of each of the dative verbs listed in Levin (1993), including the number of occurrences in each construction (PD and DOD). *Epochs* correspond to the counts for verbs in subsections of the corpus of 55 files, split by age (*Epoch 1* is the first 11 files, *Epoch 2* is the first 22, and so on).

Table 2: Number of times each verb appears in each construction (Adam corpus).

	Epoch 0		Epoch 1		Epoch 2		Epoch 3		Epoch 4		Full corpus	
Verb	DOD	PD	DOD	PD	DOD	PD	DOD	PD	DOD	PD	DOD	PD
take			0	5	0	9	0	11	0	16	0	16
say			0	3	0	4	0	6	0	6	0	6
explain					0	1	0	1	0	1	0	1
send							0	1	0	1	0	2
sell							0	1	0	1	0	1
mail							0	1	0	1	0	1
throw			1	0	1	2	1	2	1	2	1	2
read	1	1	2	5	2	11	3	12	3	13	3	16
give	2	1	15	18	39	27	62	31	82	33	106	44
show	2	1	10	5	23	9	27	11	31	15	36	17
bring			2	1	4	3	6	3	9	4	11	5
tell			1	1	8	1	14	1	17	1	22	1
call			3	0	9	0	24	0	32	0	46	0
sing					1	0	1	0	1	0	1	0
pay									2	0	2	0
serve							2	0	2	0	2	0
find											2	0
ask					2	0	3	0	3	0	4	0
make					1	0	5	0	6	0	11	0

References

Aguiar, A., & Baillargeon, R. (1999). 2.5-month-old infants' reasoning about when objects should and should not be occluded. *Cognitive Psychology*, *39*, 116–157.

Alishahi, A., & Stevenson, S. (2005). A probabilistic model of early argument structure acquisition. *Proceedings of the 27th Annual Conference of the Cognitive Science Society.*

Alishahi, A., & Stevenson, S. (2008). A computational model for early argument structure acquisition. *Cognitive Science*, *32*.

Allen, J., & Seidenberg, M. (1999). The emergence of grammaticality in connectionist networks. In B. MacWhinney (Ed.), *Emergence of language.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Ambridge, B., Pine, J., Rowland, C., & Young, C. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. *Cognition*, *106*, 87–129.

Ambridge, B., Rowland, C., & Pine, J. (2005). 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. Rep. No. RR-614). Yale University.

Baker, C. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, *10*, 533–581.

Baker, C., Tenenbaum, J., & Saxe, R. (2007). Goal inference as inverse planning. *Proceedings of the 29th Annual Conference of the Cognitive Science Society.*

Balaban, M., & Waxman, S. (1996). Words may facilitate categorization in 9-month-old infants. *Journal of Experimental Child Psychology*, *64*, 3–26.

Berwick, R. (1982). *Locality principles and the acquisition of syntactic knowledge.* Unpublished doctoral dissertation.

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.

Bloom, P. (2000). *How children learn the meanings of words.* Cambridge, MA: MIT Press.

Bloom, P., & Markson, L. (1998). Intention and analogy in children's naming of pictorial representations. *Psychological Science*, *9*, 200–204.

Bohannon, J., & Stanowicz, L. (1988). The issue of negative evidence: Adult responses to children's language errors. *Developmental Psychology*, *24*, 684–689.

Booth, A., & Waxman, S. (2002). Word learning is 'smart': Evidence that conceptual information affects preschoolers' extension of novel words. *Cognition*, *84*, B11-B22.

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.

Booth, A., & Waxman, S. (2006). Deja vu all over again: Re-revisiting the conceptual status of early word learning: comment on Smith and Samuelson (2006). *Developmental Psychology*, *42*(6), 1344–1346.

Booth, A., Waxman, S., & Huang, Y. T. (2005). Conceptual information permeates word learning in infancy. *Developmental Psychology*, *41*(3), 491–505.

Bowerman, M. (1982). Evaluating competing linguistic models with language acquisition data: Implications of developmental errors with causative verbs. *Quaderni di Semantica*, *3*, 5–66.

Bowerman, M. (1988). The no negative evidence problem: How do children avoid constructing an overly general grammar? In J. Hawkins (Ed.), *Explaining language*

universals. Oxford: Basil Blackwell.

Bowerman, M. (1990). Mapping thematic roles onto syntactic functions: Are children helped by innate linking rules? *Linguistics*, *28*(6), 1253–1289.

Braine, M. (1971). On two types of models of the internalization of grammars. In D. Slobin (Ed.), *The ontogenesis of grammar: A theoretical symposium.* New York, NY: Academic Press.

Braine, M., & Brooks, P. (1995). Verb argument structure and the problem of avoiding an overgeneral grammar. In M. Tomasello & W. Merriman (Eds.), *Beyond names of things: Young children's acquisition of verbs* (pp. 353–376). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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

Brooks, P., & Tomasello, M. (1999). How children constrain their argument structure constructions. *Language*, *75*, 720–781.

Brooks, P., Tomasello, M., Dodson, K., & Lewis, L. (1999). Young children's overgeneralizations with fixed transitivity verbs. *Child Development*, *70*, 1325–1337.

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

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. Hayes (Ed.), *Cognition and the development of language.* New York, NY: Wiley.

Carey, S., & Spelke, E. (1996). Science and core knowledge. *Philosophy of Science*, *63*, 515–533.

Chang, N. (2004). Putting meaning into grammar learning. *Proceedings of the First Workshop on Psycho-computational Models of Human Language Acquisition*, 17–24.

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., & 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.

Chouinard, M., & Clark, E. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language*, *30*, 637-669.

Clark, A., & Eyraud, R. (2006). Learning auxiliary fronting with grammatical inference. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.

Clark, E. (1982). The young word maker: a case study of innovation in the child's lexicon. In E. Wanner & L. Gleitman (Eds.), *Language acquisition: The state of the art*. New York, NY: Cambridge University Press.

Collins, M. (1999). *Head-driven statistical models for natural language parsing*. Unpublished doctoral dissertation, University of Pennsylvania.

Colunga, E., & Smith, L. (2004). Dumb mechanisms make smart concepts. *Proceedings of the 26th Annual Conference of the Cognitive Science Society*.

Colunga, E., & Smith, L. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review*, *112*(2), 347–382.

Conwell, E., & Demuth, K. (2007). Early syntactic productivity: Evidence from dative shift. *Cognition*, *103*(2), 163–179.

Cox, R. (1946). Probability, frequency, and reasonable expectation. *American Journal*

of Physics, *14*, 1–13.

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

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

Dantzig, D. van. (1957). Statistical priesthood (Savage on personal probabilities). *Statistica Neerlandica*, *2*, 1–16.

Davidson, N., & Gelman, S. (1990). Inductions from novel categories: The role of language and conceptual structure. *Cognitive Development*, *5*, 151–176.

Deacon, T. (1997). *The symbolic species: The co-evolution of language and the brain.* W. W. Norton & Co.

de Finetti, B. (1937). Prevision, its logical laws, its subjective sources. In H. Kyburg & H. Smokler (Eds.), *In studies in subjective probability* (2 ed.). New York: J. Wiley and Sons.

de Finetti, B. (1974). *Theory of probability* (2 ed.). New York: J. Wiley and Sons.

Demetras, M., Post, K., & Snow, C. (1986). Feedback to first language learners: The role of repetitions and clarification questions. *Journal of Child Language*, *13*, 275–292.

Deneve, S. (2004). Bayesian inference in spiking neurons. *Advances in Neural Information Processing Systems*, *17*.

Desai, R. (2002). Bootstrapping in miniature language acquisition. *Proceedings of the 4th International Conference on Cognitive Modelling.*

Diesendruck, G., & Bloom, P. (2003). How specific is the shape bias? *Child Development*, *74*(1), 168–178.

Dominey, P. (2003). Learning grammatical constructions in a miniature language from narrated video events. *Proceedings of the 25th Annual Conference of the Cognitive Science Society.*

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*.

Doya, K., Ishii, S., Pouget, A., & Rao, R. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. Cambridge, MA: MIT Press.

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. (1972). Some decidability results on grammatical inference and complexity. *Information and Control*, *20*(3), 244–262.

Feldman, N., & Griffiths, T. (2007). A rational account of the perceptual magnet effect. *Proceedings of the 29th Annual Conference of the Cognitive Science Society*.

Fisher, C. (1996). Structural limits on verb mapping: The role of analogy in children's interpretations of sentences. *Cognitive Psychology*, *31*(1), 41–81.

Fisher, C., Gleitman, H., & Gleitman, L. (1991). On the semantic content of subcategorization frames. *Cognitive Psychology*, *23*, 331–392.

Fisher, R. (1933). Probability, likelihood, and quantity of information in the logic of uncertain inference. *Proceedings of the Royal Society*, *146*, 1–8.

Fodor, J. (1975). *The language of thought*. New York, NY: Thomas Y. Crowell Company.

Fodor, J. (1981). *Representations: philosophical essays on the foundations of cognitive science*. Cambridge, MA: MIT Press.

Fulkerson, A., & Waxman, S. (2007). Words (but not tones) facilitate object catego-

rization: Evidence from 6- and 12-month-olds. *Cognition*, *105*(1), 218–228.

Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, *2*, 183–192.

Funahashi, K. (1998). Multilayer neural networks and bayes decision theory. *Neural Networks*, *11*(2), 209–213.

Gazzaniga, M., Ivry, R., & Mangun, G. (2002). *Cognitive neuroscience: The biology of the mind* (2 ed.). New York, NY: W.W. Norton & Company.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis* (2 ed.). Chapman & Hall.

Gelman, S., & Ebeling, K. (1998). Shape and representational status in children's early naming. *Cognition*, *66*, B35–B47.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naive theory of rational action. *Trends in Cognitive Science*, *7*(7), 287–292.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, *32*, 249–268.

Ghahramani, Z., Griffiths, T., & Sollich, P. (2006). Bayesian nonparametric latent feature models. *ISBA 8th World Meeting on Bayesian Statistics*.

Gilks, W., Richardson, S., & Spiegelhalter, D. (1996). *Markov chain monte carlo in practice*. Chapman & Hall.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1991). Human simulations of vocabulary learning. *Cognition*, *73*, 153–176.

Gleitman, L. (1990). The structural sources of word learning. *Language Acquisition*, *1*, 3–55.

Gleitman, L., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. (2005). Hard words. *Language Learning and Development*, *1*, 23–64.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*(5), 447-474.

Goldberg, A. (1995). *A construction grammar approach to argument structure*. Chicago, IL: University of Chicago Press.

Goldstone, R., Steyvers, M., Spencer-Smith, J., & Kersten, A. (2000). Interactions

between perceptual and conceptual learning. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 191–228). Lawrence Erlbaum.

Goldwater, S., & Griffiths, T. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. *Proceedings of the 45th Annual Conference of the Association for Computational Linguistics*.

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

Goodman, N. (1955). Cambridge, MA: Harvard University Press.

Goodman, N., Griffiths, T., Feldman, J., & Tenenbaum, J. (2007). A rational analysis of rule-based concept learning. *Proceedings of the 29th Annual Conference of the Cognitive Science Society*.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*(1), 1–30.

Gordon, P. (1990). Learnability and feedback. *Developmental Psychology*, *26*(2), 217–220.

Green, G. (1974). *Semantics and syntactic regularity*. Bloomington, IN: Indiana University Press.

Griffiths, T., Steyvers, M., Blei, D., & Tenenbaum, J. (2005). Integrating topics and syntax. *Advances in Neural Information Processing Systems*, *17*.

Griffiths, T., Steyvers, M., & Tenenbaum, J. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Gropen, J., Pinker, S., Hollander, M., & Goldberg, R. (1991). Affectedness and direct objects: The role of lexical semantics in the acquisition of verb argument structure. *Cognition*, *41*, 153–195.

Gropen, J., Pinker, S., Hollander, M., Goldberg, R., & Wilson, R. (1989). The learnability and acquisition of the dative alternation in English. *Language*, *65*.

Grünwald, P., Myung, J., & Pitt, M. (2005). *Advances in minimum description*

length: Theory and applications. Cambridge, MA: MIT Press.

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.

Heibeck, T., & Markman, E. (1987). Word learning in children: an examination of fast mapping. *Child Development*, *58*, 1021–1024.

Hertz, J., Krogh, A., & Palmer, R. (1991). *Introduction to the theory of neural computation: Santa Fe institute studies in the science of complexity* (Vol. 1). Reading, MA: Perseus Books.

Hirsh-Pasek, K., Treiman, R., & Schneiderman, M. (1984). Brown and Hanlon revisited: Mothers' sensitivity to ungrammatical forms. *Journal of Child Language*, *11*, 81–88.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*, 359–366.

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

Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach* (2 ed.). Open Court Publishing Company.

Hudson Kam, C., & Newport, E. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*, 151–195.

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

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

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

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

Johnson, M., Griffiths, T., & Goldwater, S. (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.

Jones, S., & Smith, L. (2002). How children know the relevant properties for generalizing object names. *Developmental Science*, *5*(2), 219–232.

Juliano, C., & Tanenhaus, M. (1993). Contingent frequency effects in syntactic ambiguity resolution. *Proceedings of the 15th Annual Conference of the Cognitive Science Society*.

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

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., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Kemp, C., Tenenbaum, J., Griffiths, T., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. *Proceedings of the 21st National Conference on Artificial Intelligence*.

Klein, D., & Manning, C. (2004). Corpus-based induction of syntactic structure: Models of dependency and constituency. *Proceedings of the 42nd Annual Conference of the Association for Computational Linguistics*, 479–486.

Kolmogorov, A. (1965). Three approaches to the quantitative definition of information. *Problems of Information and Transmission*, *1*(1), 1–7.

Landau, B., & Gleitman, L. (1985). *Language and experience: Evidence from a blind child*. Cambridge, MA: Harvard University Press.

Landau, B., Smith, L., & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*, 299–321.

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.

Lee, T., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, *20*, 1434–1448.

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

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

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.

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

Lindley, D. (1956). On a measure of the information provided by an experiment. *Annals of Mathematics*, *27*, 986–1005.

Ma, W., Beck, J., Latham, P., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

Macario, J. F. (1991). Young children's use of color in classification: Foods as canonically colored objects. *Cognitive Development*, *6*, 17–46.

MacKay, D. (1995). Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, *6*, 469–505.

MacKay, D. (2004). *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.

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

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

Marcus, G. (1993). Negative evidence in language acquisition. *Cognition*, *46*, 53–85.

Marcus, G., Pinker, S., Ullman, M., Hollander, M., Rosen, T., Xu, F., et al. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57*, 1–178.

Markman, E. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*, 57–77.

Markson, L., Diesendruck, G., & Bloom, P. (2008). The shape of thought. *Developmental Science*, *11*(2), 204–208.

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

Matthews, D., Lieven, E., Theakston, A., & Tomasello, M. (2005). The role of frequency in the acquisition of the english word. *Cognitive Development*, *201*, 121–136.

Mazurkewich, I., & White, L. (1984). The acquisition of the dative alternation: Unlearning overgeneralizations. *Cognition*, *16*, 261–283.

McClelland, J. (1998). Connectionist models and Bayesian inference. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 21–53). Oxford: Oxford University Press.

McClelland, J., McNaughton, B., & O'Reilly, R. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419–457.

McNeill, D. (1966). Developmental psycholinguistics. In F. Smith & G. Miller (Eds.), *The genesis of language.* Cambridge, MA: MIT Press.

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

Mitchell, D. (1987). Lexical guidance in human parsing: Locus and processing characteristics. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 601–681). Hillsdale, NJ: Erlbaum.

Morgan, J., & Demuth, K. (1996). *Signal to syntax: Bootstrapping from speech to*

grammar in early acquistition. Mahwah, NJ: Lawrence Erlbaum Associates.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, *17*, 357–374.

Navarro, D., & Griffiths, T. (2007). A nonparametric Bayesian method for inferring features from similarity judgments. *Advances in Neural Information Processing Systems*, *19*.

Neal, R. (1994). *Priors for infinite networks* (Tech. Rep. No. CRG-TR-94-1). University of Toronto.

Neal, R. (1996). *Bayesian learning for neural networks*. Springer.

Needham, A., & Baillargeon, R. (2000). Infants' use of featural and experiential information in segregating and individuating objects: a reply to xu, carey, and welch (1999). *Cognition*, *74*, 255–284.

Newport, E., Gleitman, H., & Gleitman, L. (1977). Mother, id rather do it myself: Some effects and non-effects of maternal speech style. In C. Snow & C. Ferguson (Eds.), *Talking to children: Language input and acquisition*. Cambridge: Cambridge University Press.

Niyogi, P. (2006). *The computational nature of language learning and evolution*. Cambridge, MA: MIT Press.

Niyogi, S. (2002). Bayesian learning at the syntax-semantics interface. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*.

Onnis, L., Roberts, M., & Chater, N. (2002). Simplicity: A cure for overregularizations in language acquisition? *Proceedings of the 24th Annual Conference of the Cognitive Science Society*.

Osherson, D., Stob, M., & Weinstein, S. (1986). *Systems that learn*. Cambridge, MA: MIT Press.

Papafragou, A., Cassidy, K., & Gleitman, L. (2007). What we think about thinking: The acquisition of belief verbs. *Cognition*, *105*, 125–165.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Perfors, A., Tenenbaum, J., Gibson, E., & Regier, T. (submitted). How recursive is

language? A Bayesian exploration. *Linguistic Review*.

Perfors, A., Tenenbaum, J., & Regier, T. (2006). Poverty of the stimulus? A rational approach. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.

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

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Pinker, S. (1994). How could a child use verb syntax to learn verb semantics? *Lingua*, *92*, 377-410.

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

Pouget, A., Dayan, P., & Zemel, R. (2003). Inference and computation with population codes. *Annual Reviews in Neuroscience*, *26*, 381–410.

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

Quine, W. (1960). *Word and object*. Cambridge, MA: MIT Press.

Rao, R. (2004). Bayesian computation in recurrent neural circuits. *Neural Computation*, *16*, 1–38.

Rao, R. (2007). Neural models of Bayesian belief propagation. In K. Doya, S. Ishii, A. Pouget, & R. Rao (Eds.), *Bayesian brain: Probabilistic approaches to neural coding* (pp. 239–267). Cambridge, MA: MIT Press.

Rao, R., Olshausen, B., & Lewicki, M. (2002). Probabilistic models of the brain: Perception and neural function.

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 pow-

erful 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.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471.

Rissanen, J., & Ristad, E. (1992). Language acquisition in the mdl framwork. In E. Ristad (Ed.), *Language computations.* American Mathematical Society (DIMACS series).

Robinson, C., Timbrook, C., & Sloutsky, V. (2006). Auditory overshadowing and categorization: When decreased visual processing facilitates categorization. *28th Annual Conference of the Cognitive Science Society.*

Rogers, T., Garrard, P., McClelland, J., Ralph, M., Bozeat, S., Hodges, J., et al. (2004). Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychology Review*, *111*, 205–235.

Rogers, T., & McClelland, J. (2004). *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.

Roy, D., Kemp, C., Mansinghka, V., & Tenenbaum, J. (2006). Learning annotated hierarchies from relational data. *Advances in Neural Information Processing Systems*, *19.*

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

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month old infants. *Science*, *274*, 1926–1928.

Samuelson, L. (2002). Statistical regularities in vocabulary guide language acquisition in connectionist models and 15-20 month olds. *Developmental Psychology*, *38*, 1016–1037.

Samuelson, L., & Smith, L. (1999). Early noun vocabularies: Do ontology, category organization, and syntax correspond? *Cognition*, *73*, 1–33.

Samuelson, L., & Smith, L. (2000). Children's attention to rigid and deformable shape in naming and non-naming tasks. *Child Development*, *71*(6), 1555–1570.

Savage, L. (1954). *Foundations of statistics*. New York, NY: J. Wiley & Sons.

Saxton, M. (1997). The contrast theory of negative input. *Journal of Child Language*, *24*, 139–161.

Schmidt, L., Kemp, C., & Tenenbaum, J. (2006). Nonsense and sensibility: Inferring unseen possibilities. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*.

Schütze, H. (1995). Distributional part-of-speech tagging. *Proceedings of the 7th conference of the European Chapter of the Association for Computational Linguistics*.

Shafto, P., Kemp, C., Mansinghka, V., Gordon, M., & Tenenbaum, J. (2006). Learning cross-cutting systems of categories. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.

Siegler, R. (2004). U-shaped interest in U-shaped development – and what it means. *Journal of Cognition and Development*, *5*(1), 1–10.

Sloutsky, V., & Robinson, C. (2008). The role of words and sounds in infants' visual processing: From overshadowing to attentional tuning. *Cognitive Science*, *32*, 354–377.

Smith, L. (2005). Shape: A developmental product. In L. Carlson & E. VanderZee (Eds.), *Functional features in language and space* (pp. 235–255). Oxford University Press.

Smith, L., Jones, S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

Smith, L., Jones, S., Yoshida, H., & Colunga, E. (2003). Whose DAM account? attentional learning explains Booth and Waxman. *Cognition*, *87*, 209–213.

Smith, L., & Samuelson, L. (2005). An attentional learning account of the shape bias: Reply to Cimpian and Markman (2005) and Booth, Waxman, and Huang (2005). *Developmental Psychology*, *42*(6), 1339–1343.

Snedeker, J., & Trueswell, J. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence

processing. *Cognitive Psychology*, *49*, 238–299.

Soja, N., Carey, S., & Spelke, E. (1991). Ontological categories guide young children's inductions of word meaning: object terms and substance terms. *Cognition*, *38*, 179–211.

Solomonoff, R. (1964). A formal theory of inductive inference, parts 1 and 2. *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.

Spelke, E. (1990). Principles of object perception. *Cognitive Science*, *14*, 29–56.

Spelke, E., & Kinzler, K. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96.

Spelke, E., Phillips, A., & Woodward, A. (1995). Infants' knowledge of object motion and human action. In *Causal cognition: A multidisciplinary debate* (pp. 44–78). Oxford: Oxford University Press.

Steyvers, M., Tenenbaum, J., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453–489.

Stinchcombe, M., & White, H. (1989). Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions. *Proceedings of the International Joint Conference on Neural Networks*, 607–611.

Stolcke, A., & Omohundro, S. (1994). Introducing probabilistic grammars by Bayesian model merging. *Proceedings of the 2nd International Colloquium on Grammatical Inference*.

Tenenbaum, J., & Griffiths, T. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–641.

Theakston, A. (2004). The role of entrenchment in childrens and adults performance limitations on grammaticality judgment tasks. *Cognitive Development*, *19*, 15–34.

Tomasello, M. (2000). The item based nature of children's early syntactic development. *Trends in Cognitive Sciences*, *4*, 156–163.

Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech: How

much is enough? *Journal of Child Language*, *31*, 101–121.

Trueswell, J., Tanenhaus, M., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 528–553.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *135*, 1124–1131.

Valiant, L. (1984). A theory of the learnable. *ACM*, *27*(11), 1134–1142.

Vapnik, V., & Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, *16*, 264–280.

Verma, D., & Rao, R. (2006). Goal-based imitation as probabilistic inference over graphical models. *Advances in Neural Information Processing Systems*, *18*.

Vitànyi, P., & Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory*, *46*(2), 446–464.

Waxman, S. (2004). Everything had a name, and each name gave birth to a new thought: Links between early word-learning and conceptual organization. In D. Hall & S. Waxman (Eds.), *From many strands: Weaving a lexicon.* Cambridge, MA: MIT Press.

Wexler, K., & Culicover, P. (1980). *Formal principles of language acquisition.* Cambridge, MA: MIT Press.

Wharton, R. (1974). Approximate language identification. *Information and Control*, *26*, 236–255.

Wonnacott, E., Newport, E., & Tanenhaus, M. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, *56*, 165–209.

Woodward, A. (1999). Infants' ability to distinguish between purposeful and non-purposeful behaviors. *Infant Behavior and Development*, *22*(2), 145–160.

Xu, F. (1999). Object individuation and object identity in infancy: The role of spatiotemporal information, object property information, and language. *Acta*

Psychologica, *102*, 113–136.

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, *85*, 223–250.

Xu, F., Carey, S., & Quint, N. (2004). The emergence of kind-based object individuation in infancy. *Cognitive Psychology*, *49*, 155–190.

Xu, F., & Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review*. (in press)

Zemel, R., Huys, Q., Natarajan, R., & Dayan, P. (2005). Probabilistic computation in spiking populations. *Advances in Neural Information Processing Systems*, *17*.

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