John Benjamins Publishing Company

This is a contribution from *Experience, Variation and Generalization. Learning a first language*.
Edited by Inbal Arnon and Eve V. Clark.
© 2011. John Benjamins Publishing Company

Bayesian modeling of sources of constraint in language acquisition

Amy Perfors & Elizabeth Wonnacott
University of Adelaide, School of Psychology/University of Oxford,
Department of Experimental Psychology

Theories of language acquisition must address the role of constraints in children's learning. Are they language-specific or domain-general? Do they come from the learner or are do they result from external factors like the nature of the data? In this chapter we describe how Bayesian modeling may be used to explore this issue. The Bayesian framework has been useful for determining what an ideal learner might be able to learn given a certain set of specific constraints and a certain type of input. It also provides a natural way to compare the effect of different constraints, and to grow towards increasingly cognitively natural models by altering those constraints.

Keywords: Learning constraints; Bayesian modeling

1. Introduction

A central question for theories of acquisition is how children are able to make linguistically appropriate generalizations on the basis of the available data. To what extent does this rely on inbuilt, specifically linguistic knowledge? To what extent is it based on general cognitive abilities and experience? Is it some interesting combination of the two? Other sections of this book center on the role of linguistic and cognitive experience – reviewing empirical work and theories focusing on individual differences, extracting regularities, and using multiple cues in learning. Any theory must also address the role of *constraints* in children's learning: constraints on the learner – such as constraints on representational capacity, limitations on processing, or memory – as well as constraints imposed by the structure of the input and the nature of the task.

In this chapter, we describe how Bayesian modeling may be used to explore constraints on learning. The Bayesian framework forces us to be explicit about every assumption made, allows us to combine representational flexibility with powerful domain-general statistical learning mechanisms, and enables us to evaluate what can be learned given different types of input. As a result of these factors, Bayesian modeling has been useful for determining the "bounds of the possible" – what an ideal learner might be able to learn given a certain set of specific constraints and a certain type of input. In addition, it provides a natural way to compare the effect of different constraints, and to grow towards increasingly cognitively natural models by altering those constraints.

The chapter has the following structure. First, we describe, in a fairly non-technical way, how Bayesian modeling works: the assumptions common to all models and how it can be used to address questions of interest to language researchers.[1] The bulk of the chapter explores some specific Bayesian models in areas ranging from word learning to grammatical acquisition to sentence processing. The goal is to illustrate the variety of ways that Bayesian modeling can be used to explore how constraints shape learning. We conclude with a brief discussion of the limitations of Bayesian models.

2. Basics of Bayesian modeling

Bayesian inference is a general-purpose computational framework for understanding and exploring how people might learn from limited evidence. The fundamental question the Bayesian modeling framework addresses is how to update beliefs and make new inferences in light of new data – how to make the sort of inductive leaps and generalizations necessary to acquire human language. Central to this framework is the assumption that a learner's degree of belief in a given theory or hypothesis can be represented as a probability. Making such an assumption allows us to use the mathematics of probability theory to conduct inference. This mathematics is, in essence, an extension of deductive logic to the case where propositions have degrees of truth or falsity; just as formal logic describes a deductively correct way of thinking, Bayesian probability theory describes an inductively correct way of thinking. This means that 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 – for instance, if you see some data supporting a new proposition *A*, you should conclude that *A* is more

[1] For a more in-depth tutorial introduction to Bayesian models in cognitive development, please see Perfors, et al. (in press), from which some of this chapter is adapted.

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 optimal common-sense reasoning and the scientific method (Jeffreys 1939; Jaynes 2003).

Because the mathematical theory is used to calculate optimal inference, Bayesian models typically operate from an "ideal learner" perspective centered at Marr's computational level (Marr 1982). This means they are generally most concerned with understanding cognition by trying to specify its goal, why that goal would be appropriate, and what the constraints on achieving that goal would be; the question of how it is implemented algorithmically is not usually addressed. As a result, the Bayesian framework is well-suited to questions about learnability – what can be learned and how this depends on the data and the constraints inherent in the learner or the task. Bayesian models help 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, one may conclude either that actual children couldn't learn it, or that the assumptions underlying the model are wrong. It is then possible to identify what those assumptions are and what precisely 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 for complicated problems, Bayesian modeling is a useful tool for this sort of problem. Inference involves comparing the different hypotheses in the hypothesis space, ultimately preferring those with the highest probability.

The techniques involved in making Bayesian models work involve ensuring that the space of hypotheses has been searched (or approximated) accurately; traditionally, the details of *how* that search is done are of less immediate interest for the modeler. However, Bayesian models can also be used to explore the nature of the constraints that a human learner might operate under. There are two ways that this is generally done. One, the most common, is simply by analyzing the assumptions inherent in the choices of representation, priors, and data to explore what constraints those set. A second, which we briefly discuss at the end of this chapter, is still in its infancy, and contrasts with the traditional assumption that the approximation made by the model is always correct; the idea is to evaluate the effects of "hobbling" the model, either by limiting the search of the hypothesis space, or presenting the model with highly constrained and/or error-filled input. Both of these ways allow us to explore the issue of how constraints on learning affect what can be learned in principle.

Before exploring any specific models, it will be useful to describe some general properties of Bayesian modeling common to all of them. The goal of inference is to select from among a set of theories, or hypotheses, which one(s) best explain the observed data. The Bayesian framework does this 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 hypotheses that offer a balance between the two.

Data is assumed to have been 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 might be generated from a mental lexicon; and both may be partially affected by social and pragmatic factors as well. The job of the learner/model is to evaluate different hypotheses about the underlying nature of that process, and to make predictions based on the most likely hypotheses. Construction of the model involves making explicit assumptions about that generative process, the nature of the resulting set of hypotheses, and the constraints that might affect the learner. The goal is to evaluate what can be learned (i.e. which hypotheses are favored) based on different types and amounts of realistic input.

Hypotheses are compared using Bayes' Rule (Equation 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 *a priori* probability that the hypothesis is true (the prior).

(1) $P(H|D) \propto p(D|H)\, p(H)$

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 1 illustrates this graphically: in it, data points can be generated by sampling uniformly from different areas in the 2D space, and the rectangles shown represent hypotheses about what areas exist. Based on Figure 1 we might conclude that hypothesis *A* looks too simple, hypothesis *C* seems too complex, and hypothesis *B* is "just right." Bayesian models capture this intuition using techniques sometimes known as the Bayesian Occam's Razor (MacKay 2003).

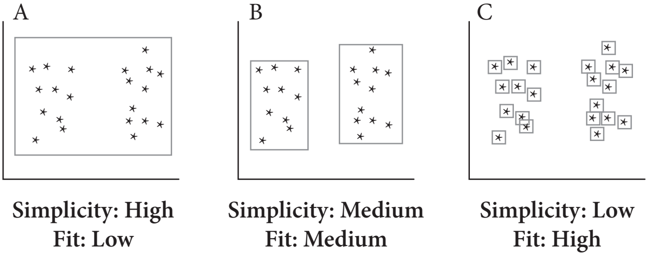

Figure 1. Hypothesis *A* is too simple, *C* is too complex, and *B* is "just right." Hypothesis *A* is quite simple, but fits the observed data poorly: *C* fits closely but is highly complicated. The best description of the data should optimize a tradeoff between complexity and fit, as in *B*

How well a hypothesis fits the data is captured by the likelihood, given by $P(D|H)$: 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 intuitively. For instance, hypothesis *C* in Figure 1 clearly has a high likelihood: if the hypothesis is true – that is, if the data is truly generated by twenty-one distinct underlying processes corresponding to the twenty-one rectangles of *C* – then the data points 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 data points are found where they are rather than in any of the other locations within the rectangle. The ratio of the observed data points 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; thus, hypotheses that make specific predictions – those with more explanatory power – are favored in the likelihood.

The prior probability of a hypothesis, or $P(H)$, may be set arbitrarily to capture whatever *a priori* assumptions one might wish. It is therefore useful in principle for exploring the effect of making different assumptions. In practice, is it usually calculated in such a way that hypotheses that are simpler – that is, hypotheses with fewer parameters, or that require less information to specify – have higher probability. This emerges as a natural result of the generative assumptions underlying the Bayesian framework, in which hypotheses are themselves generated from a space of candidate hypotheses. For instance, the hypotheses in Figure 1 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 center of the rectangle (x and y), one for its length (l), and one for its width (w). By contrast, hypothesis *C* contains twenty-one distinct rectangular regions, and therefore requires 84 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 easy it is to mess up – in other words, the more choices a hypothesis requires, the more likely it is that those choices could have been made in a different way. There are many fewer ways to have one rectangle than to have twenty-one. This conception of a prior probability has some dependence upon the nature of the underlying generative model, but less than one might imagine. For instance, suppose that we assumed that the data points in Figure 1 were generated by circular regions rather than rectangles. Then each region would require three choice points rather than four (the *x* and *y* coordinates of the center of the circle, plus its radius). Despite the different generative model, the logic favoring simple hypotheses would be the same: multiple regions would still be *a priori* less likely than a few. The particular 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*. It can also be thought of as the range of hypotheses the model can entertain, or the set of things that the model can learn. This is a natural reflection of the space of possibilities inherent in the problem: for instance, the hypothesis space corresponding to Figure 1 consists of all possible combinations of rectangles. Note that hypothesis spaces are incorporated into every learning model, whether connectionist, Bayesian, or verbally specified, since every learning model has some things it can learn and some things it cannot. Put another way, every model (at least implicitly) assumes that the learner is entertaining some hypotheses and not others. This is simply made more explicit in the Bayesian framework than in others.

For most of the rest of this chapter, we focus on some specific results within the language acquisition literature. These do not reflect all of the Bayesian approaches to topics in language acquisition, but are chosen to reflect the diversity of topics as well as to highlight the ways that the Bayesian modeling framework can be used to explore the issue of constraints in language learning.

Constraints on language learning may be external or internal to the learner: constraints that come from the input are clearly external and the constraints that come from the prior assumptions or representational capacities of the learner are clearly internal. These sources of constraint must be specified when creating any Bayesian model, and thus every Bayesian model is potentially informative about them. However, the contribution of the input or learner may also be more subtle. Bayesian models have explored the following additional sources of constraint: (a) constraints caused by mutual learning of multiple phenomena at once;

(b) constraints caused by making different assumptions about how data is sampled; (c) constraints caused by being able to make abstract inferences at the same time as specific ones; or (d) constraints caused by limitations on the learner. The examples below are organized according to the ways in which they provide these constraints. Since *all* of them incorporate constraints based on the learner's representational capacities and prior assumptions, we will highlight these as we go rather than discussing them in a separate section of their own.

3. Constraints caused by mutual learning

There are several examples elsewhere in this book exploring the issue of learning multiple pieces of linguistic knowledge simultaneously. The basic idea is much like bootstrapping, which suggests that knowing a small amount about topic A might provide enough to learn a bit about related topic B, which in turn makes it easier to learn about topic A, and so forth until one fully understands both A and B. This is an appealing theory, especially because it appears to provide a way to solve classic chicken-and-egg problems like word learning.

Models that are both explicit about what is built in and incorporate powerful learning machinery allow us to form testable theories and determine whether it is possible for such a "mutual constraint" explanation to work. What assumptions must the learner make about how the two topics interact? What biases must exist to get the earliest learning off the ground? For instance, consider the problem of phoneme category learning. During the first 12 months of life, infants lose the ability to discriminate non-native speech sound contrasts (e.g. Werker & Tees 1984), and also begin to successfully segment words from fluent speech (e.g. Jusczyk & Aslin 1995; Jusczyk, Houston & Newsome 1999). Although it is sometimes assumed that word segmentation follows phonetic category acquisition, it may also be possible for the two to mutually reinforce each other; for instance, one of the indications that a phoneme contrast exists in a language is for two words to differ only on that contrast (e.g. minimal word pairs like "race" and "lace" are an indication that the *r*/*l* contrast exists in English). Conversely, knowing which phonetic category a sound belongs to can be helpful for determining which word it is a part of.

Recent work by Feldman et al. (2009) uses a Bayesian model to explore the hypothesis that information about the nature of one's lexicon can provide a useful source of information for phonetic category acquisition. The model contains two components: one acquires phonetic categories (specifically vowel categories) from raw acoustic input using a Mixture of Gaussians approach, which assumes that phonetic categories are learned on the basis of the distribution of individual exemplars in phonetic space (an assumption which has some experimental support: e.g. Maye, Werker, & Gerken 2002). The Mixture of Gaussians approach (and the corresponding assumptions about how phonemes are represented by the learner) is standard among computational models of phonetic category acquisition (e.g. Boer & Kuhl 2003; Vallabha et al. 2007; McMurray, Aslin & Toscano 2009). The difference with the Feldman et al. (2009) work is that this approach is incorporated into a framework that allows the model to combine phonemic knowledge with the second component of the model: a rudimentary lexicon. The lexicon is composed of words that correspond to sequences of the phonemes (each of which is an exemplar from one of the learned phonetic categories). The goal is to simultaneously acquire both the correct lexicon and the set of phonemes in the language; because the lexicon is made up of those phonemes, they mutually constrain each other.

Input consists of data generated from an artificial lexicon composed of words consisting of sequences of phonemes generated from phonetic categories based on the Hillenbrand et al. (1995) vowel formant data. Feldman et al. (2009) found that learning phonemes from distributional information was facilitated by being able to acquire a lexicon at the same time, assuming that the learner has a prior preference for fewer phonetic categories. Interestingly, the model performed best when data was generated from a lexicon containing *non*-minimal pairs because words differing in a single segment (i.e. minimal pairs) could lead the model to hypothesize that they represented the same word, whereas non-minimal pairs helped the learner notice the difference between similar-sounding phonemes based on other differences in the word. This provides an explanation for empirical work demonstrating that 15-month-olds in a word learning task might not be able to distinguish between similar-sounding object labels, but could succeed if familiarized with non-minimal pairs containing the same sounds (Thiessen 2007). It demonstrates how the nature of the constraints provided by the structure of the input may change in the context of different assumptions about the learning system and the types of representations being computed.

Other models within the Bayesian framework have been used to explore the idea of mutual constraint. For instance, Frank, Goodman and Tenenbaum (2009) presented a model of word learning that combined learning about speakers' intentions with acquiring the meaning of individual words. The model assumes that the words produced by a speaker depend both on their intention – the referential goal of their utterance – and what the words mean. The model incorporates information about which objects are in the scene at the time of the utterance, representing both them and words as units; it also represents possible speaker intentions, which consist of distributions over the sort of things that speaker tends to talk about. The model, evaluated on annotated corpus data, inferred better pairings between words and object concepts than comparison models which only learned from cross-situational statistics of how words were used and did not also try to learn about intentions. This work also explains other phenomena in the word-learning literature; for instance, mutual exclusivity – proposed by some theories to be an innate constraint on the learner – arises naturally out of the mathematics of probability theory. A lexicon in which one object can be labeled by multiple words will have a lower likelihood, because "multiple words" are like hypothesis *A* in Figure 1; such a lexicon make a less specific prediction than a lexicon in which there is a one-to-one mapping.

4. Constraints due to different assumptions about how data is sampled

Another source of constraints is often overlooked, but can have a profound effect on what is ultimately learned: these are assumptions about how data is sampled from the environment. This was implicit even in the previous models we discussed – for instance, the phoneme-category-learning model gains some of its explanatory power from the assumption that phonemes are generated as parts of words. However, assumptions about sampling (which every statistical model must make, whether explicitly or implicitly) can be even simpler and more basic. Certain statistical assumptions may even explain the basis on which children are willing to make strong inferences based on apparently little data.

One example of this concerns work focusing on the problem of learning in the absence of negative evidence. This learnability puzzle centers on the question of how children know that some unobserved pattern is ungrammatical, rather than simply not heard. It arises in many contexts, but one common situation is in the acquisition of verb constructions (Baker 1979; Pinker 1989). Verbs that take arguments follow certain underlying regularities. For instance, many verbs occur in both the prepositional dative (PD) and direct object dative (DOD) constructions. However, the generalization that all verbs that occur in one should be grammatical in another does not seem to apply; as an example, *confess* is grammatical in PD syntax ('Carrie confessed the truth to Katharine') but not in DOD (*'Carrie confessed Katharine the truth'). Despite never receiving explicit instruction about the grammaticality of *confess* – and even though a near-synonym, *tell*, is grammatical in both – fluent speakers of English have no trouble avoiding the correct form. How is this possible?

There have been many suggestions about how language learners resolve the problem of negative evidence (Braine, 1971; Pinker 1989; Braine & Brooks 1995; Goldberg 1995; MacWhinney 2004; Wonnacott, Newport & Tanenhaus 2008). One influential explanation suggests that learners take indirect evidence into account; as time goes by and certain constructions are not used, particularly if other constructions are used when they could be, this is indirect evidence that those constructions are probably ungrammatical (Braine 1971; Braine & Brooks 1995; Goldberg 1995; Wonnacott et al. 2008). Again Bayesian modeling allows us to formally investigate this explanation: under what assumptions about how data was generated would this sort of inference be justified? Precisely how much data would constitute sufficient negative evidence, and why?

Several Bayesian models of the acquisition of verb argument constructions in the absence of negative evidence exist, and all provide an answer to these questions (Dowman 2000; Onnis, Roberts, & Chater 2002; Chater & Vitànyi 2007; Alishahi & Stevenson 2008; Hsu & Griffiths 2009; Perfors, Wonnacott & Tenenbaum 2010). Although all of the models must by nature assume that the learner is capable of representing verb constructions (or their features) in some way, the models otherwise differ widely in terms of their representational assumptions. Notably, however, all solve the problem of no negative evidence in the same way. This demonstrates that the answer does not arise because of idiosyncratic properties of an individual model, but rather is the result of a general property of optimal inference. In particular, all of the models show that as long as a learner assumes that linguistic data is *strongly sampled* from the environment – that individual utterances are independently drawn from the set of grammatical utterances (perhaps with some error) – then an ideal learner will realize, as evidence accumulates, that not having seen a particular utterance is increasing evidence that it is ungrammatical.

We can understand how this emerges as a general property of optimal inference by making reference to Figure 1. Here, each dot represents a single verb usage, and the rectangles represent the underlying process (i.e. the underlying verb knowledge, perhaps instantiated by a probabilistic rule of some sort). The job of a learner is to evaluate hypotheses about what that underlying verb knowledge is; as we saw earlier, the complexity of the hypothesis (captured by the prior) and the degree of fit to the data (captured by the likelihood) are traded off by the model. A rational statistical learner will generally therefore favor neither the most over-general explanation (hypothesis *A*) nor the most specific one (hypothesis *C*). Furthermore, as evidence accumulates, a distinctive pattern of reasoning emerges. When there are few data points, the simpler theories are favored, resulting in a tendency toward overgeneralization. As the number of data points 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.

Sampling assumptions – assumptions about how data is generated – can explain patterns of inference in other areas as well. For instance, there is evidence that children make more narrow generalizations on the basis of three data points than they do from a single data point when learning the name of a novel word (Xu & Tenenbaum 2007). This can also be explained if children assume that the words are "sampled" independently from the possible extensions of the concept: if they are, then three instances provide stronger evidence about the extent of the concept than one does. Other work has shown that different sampling assumptions license different patterns of inference (Navarro et al. 2008). If people assume pedagogical sampling – that data was generated by somebody explicitly trying to teach them about a concept – then they make different inferences than if they assume that data was generated randomly from the concept (Shafto & Goodman 2008). This is explained by a Bayesian model which also predicts children's behavior in a causal-learning situation (Bonawitz et al. 2009).

5. Constraints imposed by learning abstract knowledge

Another constraint on learning emerges if a learner has access to, or a propensity to acquire, higher-level, abstract knowledge about the thing to be learned. For instance, the issue of determining the referent of a word is famously underconstrained (Quine 1960). As we have already seen, to some extent this problem may be solved based on the constraints imposed by other problems that must be solved simultaneously, like the problem of inferring speaker's intentions. Other solutions involve hypothesizing the existence of abstract constraints, such as the whole object constraint in word learning, which suggests that children assume that words apply to whole objects rather than parts (Heibeck & Markman 1987). While some of the constraints like this are theorized to be innate, others are clearly learned. By the age of 24 months old, English-learning children assume that count nouns are organized by shape; given a novel object with a novel label, they tend to generalize the label to items that are similar in shape but not color or texture (Landau, Smith & Jones 1988; Soja, Carey & Spelke 1991). There are many reasons to believe that this bias is learned. 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 acquire the bias earlier and also results in faster learning of other, non-taught words (Smith et al. 2002).

The shape bias is a constraint that says that, in general, categories of objects associated with a given label will tend to share the shape properties. This generalization emerges given the statistical structure of count nouns (at least in English), as connectionist modeling work has demonstrated (Smith et al. 2003). More recently, however, hierarchical Bayesian models have additionally explained on a computational level *why* such a bias might emerge, especially on the basis of as little input as children require (as in Smith et al. 2002). There are a range of hierarchical Bayesian models of category learning, each capable of acquiring higher-level constraints about the general nature of category organization at the same time as specific knowledge about particular categories (Navarro 2006; Kemp, Perfors & Tenenbaum 2007; Griffiths et al. 2008; Heller, Sanborn & Chater 2009). All of them assume that the learner is capable of representing features such as shape (or substance, or material, or whatever); given this, all suggest that not only can higher-level constraints about which features are important be learned, but sometimes these constraints can also be acquired more rapidly than the specific knowledge. This is because of several factors, but one of the most important is that, whereas any particular piece of data is generally only relevant to one specific category (e.g. a ball labeled "ball" does not inform a child about what dogs are labeled), all pieces of data are relevant about higher-order constraints (learning that balls are ball-shaped provides some evidence for the hypothesis that all categories are organized by shape). Bayesian models are useful for demonstrating how higher-level constraints can emerge, and why they do so extremely rapidly.

Higher-level constraints exist in other areas besides word learning. For example, we previously discussed the usage of indirect evidence to solve the problem of no negative evidence in domains such as the acquisition of verb constructions. There is some experimental evidence that the extent to which learners are influenced by the usage of a verb in one construction, and its corresponding non-usage in another, depends upon their previous experience of how likely a verb encountered in one construction is to continue in the same construction (Wonnacott et al. 2008). This is captured by a hierarchical Bayesian model capable of learning about the relationships between verbs and constructions on different levels simultaneously (Perfors et al. 2010).

Other types of syntactic generalizations are also governed by abstract knowledge: for instance, knowing that language contains hierarchical phrase structure, and that linguistic rules are consistent with that structure, may underpin children's ability to correctly form polar interrogatives in English without making certain types of incorrect generalizations (Chomsky 1968, 1980). A recent Bayesian model of syntax acquisition suggests that this knowledge may be learnable on the basis of typical child-directed input given certain reasonable assumptions about representation (e.g. that the child is capable of representing grammars both with and without hierarchical phrase structure) and data generation (Perfors, Tenenbaum & Regier 2006, 2011). In fact, this research suggests that it may be possible to acquire this abstract principle on the basis of less input than is necessary to identify the correct specific grammar. This is another example of a higher-level constraint that may be learnable before more specific knowledge is acquired.

6. Constraints imposed by imperfect approximations to the ideal

So far all of the examples of Bayesian models considered have been “ideal learning” models, as explained at the beginning of the chapter. They have been useful for explaining how, in principle, different sorts of constraints can guide different patterns of learning: constraints imposed by acquiring multiple pieces of knowledge at once can explain how the learning problem can be simplified; constraints imposed by making assumptions about data can explain how inferences are guided; and constraints imposed by learned higher-level knowledge can explain certain types of generalizations.

However, children are *not* ideal learners, and this in itself can impose a constraint of its own. Some theories, in fact, propose that some of children’s success in language learning is due to their poorer cognitive capacities (e.g. Newport 1990; Hudson-Kam & Newport 2005). Beyond this, it is evident that a learner with poor memory, processing speed, metacognition, or other cognitive capacities will make different generalizations and form different inferences than a learner without such deficits. Can Bayesian models tell us anything about these sorts of constraints?

One new area of research focuses on this very question. This is called the “rational process” approach, and the basic idea explores what happens when inference within a Bayesian framework is only approximate, rather than ideal, subject to resource limitations. As described earlier, typical Bayesian models use search techniques that are guaranteed to accurately approximate the entire hypothesis space and identify the best hypothesis. It is possible to “hobble” these techniques, either by stopping the search early, or limiting the amount of resources it can command (see Doucet et al. 2001 for technical details on one of the most common approach for doing this, known as particle filters). Hobbling ideal learning models allows us to explore to what extent human behavior can be explained as *approximations* to the ideal; or, seen another way, as ideal learning given built-in constraints on memory or processing capacity. Rational process models have been applied to non-linguistic learning problems including visual attention and object tracking (Vul et al. 2009), decision making (Yi et al. 2009) and categorization (Sanborn et al. 2006, 2010).

Although there are few examples of these models in language acquisition (though see Pearl et al. 2010), their utility has been demonstrated in explaining adult language processing. Levy et al. (2008) present a limited-memory Bayesian model of speech processing that reproduces classic results in online sentence processing, including an explanation for garden-path sentences, as well as an account explaining why the preferred alternative in a syntactic ambiguity appears to grow more attractive over time even when there is no strong disambiguating information. The memory constraints built into the model play a key explanatory role.

7. Limitations of Bayesian models

We have argued that Bayesian models provide a powerful tool for exploring problems in language acquisition, due to their combination of representational flexibility and domain-general statistical learning mechanisms. Nevertheless, there are limitations on what can be accomplished within the Bayesian framework, both currently and in principle. Most importantly, Bayesian models are inapplicable to problems that cannot be framed or understood as problems involving induction. Although many problems in cognitive science and language acquisition can be so construed (including the problems discussed in this chapter as well as issues related to categorization, causal reasoning, and social inference), not all important problems are problems of induction. For instance, many scientists are concerned with understanding how different cognitive characteristics are related to each other (for instance, IQ and attention), and how that changes over time. As another example, Bayesian models might be fairly unilluminating in areas such as emotional regulation or most types of psychopathology.

Another limitation is that because Bayesian models offer explanations largely on Marr's computational level, they may not be appropriate in situations in which behavior arises as a result of the specific details of the implementation. Rational process models may ameliorate this problem somewhat, but it is currently unclear how far they will go. Even if we can be sure that approximating optimal performance is what people are doing, there are always many different ways to approximate it, and ours may be incorrect. For instance, if behavior in a certain situation emerges because of the particular architecture of the brain or the way in which action potentials are propagated, a rational process model may not capture that behavior. Of course, this limitation is an example of a general type of limitation that afflicts all cognitive models: if reality is not expressible in the model, the model will not be an appropriate simplification of reality. (In a similar way, connectionist models may fall short in cases where particular assumptions about architecture, learning rate or form, etc are incorrect in ways that affect the nature of the outcome.) Since we do not know what the situations in which this might occur are (or even if they exist), it is hard to say how constraining this limitation will be in practice. The best way to identify those situations is to attempt to model them and note if and how they fail to match behavior.

A final limitation exists more in practice than in principle. As Bayesian models get increasingly complex, their computational tractability decreases dramatically. Currently, no Bayesian model exists that can deal with a quantity of input that even approximates the amount seen by a child – the search space is simply too large. Moore's Law and the development of improvements in machine learning mean that this limitation will grow less important over time; however, for now, it does mean that generating precise predictions on the basis of large amounts of data, or even searching effectively through very high-dimensional hypothesis spaces, is difficult. Although this limitation is at least in part a reflection of our current technology, to the extent that it arises simply because these problems are enormously complex, it is a limitation that all models share.

8. Summary

This chapter has explored some of the ways that Bayesian models can be used to explore how different constraints can have an affect on what is learned during the process of language acquisition. This modeling approach – especially the use of rational process models to approximate ideal learners – is still relatively new. As a result, many research questions remain open, and the complete utility of the approach is yet to be demonstrated. One direction for future research is to reconcile computational-level "ideal learner" Bayesian models with cognitively plausible models of language learning and processing. Approaches that approximate ideal learners go some of the way towards this. We would ultimately like an account in which higher level rational learning naturally "falls out" of lower level processes. Correctly specifying and understanding that higher-level learning is a vital first step to that end. For now, Bayesian models provide a precise method for exploring assumptions about learnability under particular conditions, and thus provide an important tool for exploring theories of acquisition.

References

Alishahi, A. & Stevenson, S. 2008. A computational model for early argument structure acquisition. *Cognitive Science* 32: 789–834.

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

de Boer, B. & Kuhl, P.K. 2003. Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online* 4: 129–134.

Bonawitz, E., Shafto, P., Gweon, H., Chang, I., Katz, S. & Schulz, L. 2009. The double-edged sword of pedagogy: Modeling the effect of pedagogical contexts on preschoolers' exploratory play. In *Proceedings of the* 31st *Annual Conference of the Cognitive Science Society*, N. Taatgen, H. van Rijn, L. Schomaker & J. Nerbonne (eds), 1575–1580. Austin TX: Cognitive Science Society.

Braine, M.D.S. 1971. On two types of models of the internalization of grammars. In *The Ontogenesis of Grammar: A Theoretical Symposium*, D. Slobin (ed.), 153–186. New York NY: Academic Press.

Braine, M.D.S. & Brooks, P. 1995. Verb argument structure and the problem of avoiding an overgeneral grammar. In *Beyond Names of Things: Young Children's Acquisition of Verbs*, M. Tomasello & W. Merriman (eds), 353–376. Hillsdale NJ: Lawrence Erlbaum Associates.

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

Chomsky, N. 1968. *Language and Mind*. New York NY: Harcourt, Brace, Jovanovitch.

Chomsky, N. 1980. *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*, M. Piatelli-Palmarini (ed.). Cambridge MA: Harvard University Press.

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.

Doucet, A., de Freitas, N. & Gordon, N. 2001. *Sequential Monte Carlo in Practice.* Berlin: Springer.

Dowman, M. 2000. Addressing the learnability of verb subcategorizations with Bayesian inference. In *Proceedings of the* 22nd *Annual Conference of the Cognitive Science Society*, L. Gleitman & A. Joshi (eds), 107–112. Austin TX: Cognitive Science Society.

Feldman, N., Griffiths, T.L. & Morgan, J. 2009. Learning phonetic categories by learning a lexicon. In *Proceedings of the* 31st *Annual Conference of the Cognitive Science Society*, N. Taatgen, H. van Rijn, L. Schomaker & J. Nerbonne (eds), 2208–2213. Austin TX: Cognitive Science Society.

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

Goldberg, A. 1995. *A Construction Grammar Approach to Argument Structure*. Chicago IL: University of Chicago Press.

Griffiths, T.L., Sanborn, A., Canini, K. & Navarro, D. 2008. Categorization as non-parametric Bayesian density estimation. In *The Probabilistic Mind: Prospects for Bayesian Cognitive Science,* M. Oaksford & N. Chater (eds), 303–328. Oxford: OUP.

Heibeck, T. & Markman, E. 1987. Word learning in children: an examination of fast mapping. *Child Development* 58: 1021–1024.

Heller, K., Sanborn, A. & Chater, N. 2009. Hierarchical learning of dimensional biases in human categorization. In *Advances in Neural Information Processing Systems 22,* Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams & A. Culotta (eds), 727–735. Cambridge MA: The MIT Press.

Hillenbrand, J., Getty, L.A., Clark, M.J. & Wheeler, K. 1995. Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America* 97: 3099–3111.

Hudson Kam, C. & Newport, E. 2005. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning & Development* 1: 151–195.

Hsu, A. & Griffiths, T.L. 2009. Differential use of implicit negative evidence in generative and discriminative language learning. In *Advances in Neural Information Processing Systems 22,* Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams & A. Culotta (eds), 754–762. Cambridge MA: The MIT Press.

Jaynes, E. 2003. *Probability Theory: The Logic of Science.* Cambridge: CUP.

Jeffreys, H. 1939. *Theory of Probability.* Oxford: Clarendon Press.

Jusczyk, P.W. & Aslin, R.N. 1995. Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology* 29: 1–23.

Jusczyk, P.W., Houston, D.M. & Newsome, M. 1999. The beginnings of word segmentation in English-learning infants. *Cognitive Psychology* 39: 159–207.

Kemp, C., Perfors, A. & Tenenbaum, J.B. 2007. Learning overhypotheses with hierarchical Bayesian models. *Developmental Science* 10: 307–321.

Landau, B., Smith, L. & Jones, S. 1988. The importance of shape in early lexical learning. *Cognitive Development* 3: 299–321.

Levy, R., Reali, F. & Griffiths, T. 2008. Modeling the effects of memory on human online sentence processing with particle filters. In *Advances in Neural Information Processing Systems 21,* D. Koller, D. Schuurmans, Y. Bengio & L. Bottou (eds), 937–944. Cambridge MA: The MIT Press.

MacKay, D. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge: CUP.

MacWhinney, B. 2004. A multiple process solution to the logical problem of language acquisition. *Journal of Child Language* 31: 883–914.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York NY: Henry Holt & Company.

Maye, J., Werker, J.F. & Gerken, L. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82: B101-B111.

McMurray, B., Aslin, R.N. & Toscano, J.C. 2009. Statistical learning of phonetic categories: Computational insights and limitations. *Developmental Science* 12: 369–378.

Navarro, D. 2006. From natural kinds to complex categories. In *Proceedings of the* 28th *Annual Conference of the Cognitive Science Society,* R. Sun & N. Miyake (eds), 621–626. Austin TX: Cognitive Science Society.

Navarro, D., Lee, M., Dry, M. & Schultz, B. 2008. Extending and testing the Bayesian theory of generalization. In *Proceedings of the* 30th *Annual Conference of the Cognitive Science Society,* V. Sloutsky, B. Love & K. McRae (eds), 1746–1751. Austin TX: Cognitive Science Society.

Newport, E. 1990. Maturational constraints on language learning. *Cognitive Science* 14: 11–28.

Onnis, L., Roberts, M. & Chater, N. 2002. Simplicity: A cure for over-regularizations in language acquisition? In *Proceedings of the* 24th *Annual Conference of the Cognitive Science Society,* W. Gray & C. Schunn (eds), 720–725. Austin TX: Cognitive Science Society.

Pearl, L., Goldwater, S. & Steyvers, M. 2010. How ideal are we? Incorporating human limitations into Bayesian models of word segmentation. In *Proceedings of the* 34th *annual Boston University Conference on Child Language Development*, K. Franich, K. Iserman & L. Keil (eds), 315–326. Somerville MA: Cascadilla Press.

Perfors, et al. In press. A tutorial introduction to Bayesian models of cognitive development. *Cognition.*

Perfors, A., Tenenbaum, J. & Regier, T. 2006. Poverty of the stimulus? A rational approach. In *Proceedings of the* 28th *Annual Conference of the Cognitive Science Society*, R. Sun & N. Miyake (eds), 663–668. Austin TX: Cognitive Science Society.

Perfors, A., Tenenbaum, J.B. & Regier, T. 2011. The learnability of abstract syntactic principles. *Cognition* 118: 306–338.

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

Pinker, S. 1989. *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge MA: The MIT Press.

Quine, W. 1960. *Word and Object.* Cambridge MA: The MIT Press.

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: 1555–1570.

Sanborn, A., Griffiths, T. & Navarro, D. 2006. A more rational model of categorization. In *Proceedings of the* 28th *Annual Conference of the Cognitive Science Society*, R. Sun & N. Miyake (eds), 726–731. Austin TX: Cognitive Science Society.

Sanborn, A.N., Griffiths, T.L. & Navarro, D.J. 2010. Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review* 117: 1144–1167.

Shafto, P. & Goodman, N. 2008. Teaching games: Statistical sampling assumptions for pedagogical situations. In *Proceedings of the* 30th *Annual Conference of the Cognitive Science Society*, V. Sloutsky, B. Love, & K. McRae (eds), 1632–1637. Austin TX: Cognitive Science Society.

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: 13–19.

Smith, L., Jones, S., Yoshida, H. & Colunga, E. 2003. Whose DAM account? Attentional learning explains Booth and Waxman. *Cognition* 87: 209–213.

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.

Thiessen, E.D. 2007. The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory & Language* 56: 16–34.

Vallabha, G.K., McClelland, J.L., Pons, F., Werker, J.F. & Amano, S. 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences* 104: 13273–13278.

Vul, E., Frank, M.C., Alvarez, G.A. & Tenenbaum, J.B. 2009. Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In *Advances in Neural Information Processing Systems 22,* Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams & A. Culotta (eds), 1955–1963. Cambridge MA: The MIT Press.

Werker, J.F. & Tees, R.C. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development* 7: 49–63.

Wonnacott, E., Newport, E. & Tanenhaus, M. 2008. Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology* 56: 165–209.

Xu, F. & Tenenbaum, J.B. 2007. Word learning as Bayesian inference. *Psychological Review* 114: 245–272.

Yi, S.K.M., Steyvers, M. & Lee, M.D. 2009. Modeling human performance on restless bandit problems using particle filters. *Journal of Problem Solving* 2: 33–53.