"Statistical Learning, Inductive Bias, and Bayesian Inference in Language Acquisition"

Lisa Pearl & Sharon Goldwater

Language acquisition is a problem of induction: the child learner is faced with a set of specific linguistic examples and must infer some abstract linguistic knowledge that allows the child to generalize beyond the observed data, i.e., to both understand and generate new examples. Many different generalizations are logically possible given any particular set of input data, yet different children within a linguistic community end up with the same adult grammars. This fact suggests that children are biased towards making certain kinds of generalizations rather than others. The nature and extent of children's inductive bias for language is highly controversial, with some researchers assuming that it is detailed and domain-specific (e.g., Chomsky 1973, Baker 1978, Chomsky 1981, Huang 1982, Fodor 1983, Bickerton 1984, Lasnik & Saito 1984, Gleitman & Newport 1995) and others claiming that domain-general constraints on memory and processing are sufficient to explain the consistent acquisition of language (e.g., Elman, Bates, Johnson, Karmiloff-Smith, Parisi, & Plunkett 1996, Sampson 2005). In this chapter, we discuss the contribution of an emerging theoretical framework called Bayesian learning that can be used to investigate the inductive bias needed for language acquisition.[1]

In the Bayesian view of learning, inductive bias consists of a combination of hard and soft constraints. Hard constraints make certain grammars[2] impossible for any human to acquire; in the language of Bayesian modeling, these impossible grammars are outside the learner's *hypothesis space*. Grammars inside the hypothesis space are learnable given the right input data, but they may not all be equally easy to learn. Soft constraints, implemented in the form of a probability distribution over the hypothesis space, mean that the learner will be biased towards certain of these grammars more than others. A "difficult" (low-probability) grammar can be learned, but will require more evidence (input data favoring this grammar) in order to be learned. In the absence of such evidence, the child will instead acquire a high-probability grammar that is also compatible with the input.

Under this view of learning, the central question of language acquisition is to determine what the hard and soft constraints are. A key assumption is that learners have access to domain-general statistical learning mechanisms that closely approximate the rules of probability theory. Given a particular set of input data, these probabilistic learning mechanisms then allow learners to converge on a grammar that is both compatible with the data and has high probability in the hypothesis space. In this sense, the grammar is the *optimal* choice, given the data. This notion of optimization arises from the ties between Bayesian modeling and the tradition of *rational analysis* in cognitive science (Chater & Oaksford, 1999), which focuses on the adaptation of the

[1] Bayesian models themselves are not a new idea, and have long been used in mathematics and computer science for both statistical analysis and machine learning (Bishop 2006, Duda, Hart, & Stork 2000, Gelman, Carlin, Stern & Rubin 2003), but their use in cognitive modeling, particularly in the area of language acquisition, is much newer.

[2] We use the term "grammar" very broadly to mean any kind of abstract linguistic knowledge used for generalization.

organism to its environment, with the resulting implication that cognitive processes are in some sense optimized to the task. Probability theory plays a central role in Bayesian modeling precisely because it is a mathematical tool for optimizing behavior under uncertainty. We discuss all of these ideas and their implications further in Section 2.

Although many of these ideas may be new to the reader, some aspects of Bayesian modeling may be familiar from other approaches to learning. For example, many nativist linguists (particularly those in the Chomskyan tradition) also have a fundamental research goal of explicitly defining the hypothesis space of possible grammars (as part of Universal Grammar). However, unlike most models of learning based on Chomskyan theories, Bayesian models of learning are inherently probabilistic, both in defining a probability distribution over the hypothesis space and in the way the learner is assumed to incorporate information from linguistic data. These qualities allow Bayesian models to be more robust to noisy data and also to avoid some of the classic learnability problems faced by non-statistical learners, such as the *subset* problem, which is a specific instance of the *no negative evidence* problem (see Tenenbaum & Griffiths (2001) for a review).[3] Unlike deterministic learners, probabilistic learners can accumulate *indirect negative evidence* against a (probabilistic) grammar if a structure that is licensed by the grammar occurs in the input significantly less often than expected under the grammar.

Although some other recent proposals incorporate probabilistic learning (e.g., Yang 2002, Legate & Yang 2007, Pearl 2011), they don't include the idea of optimization discussed above. In addition, many learning models in the Chomskyan tradition assume not only that the hypothesis space itself is defined using domain-specific concepts, but that the learning algorithm makes reference to these concepts, so that it too is domain-specific (see Sakas (this volume) for a more detailed discussion). Bayesian models, in contrast, assume that the learner's use of statistical information is entirely domain-general, and domain-specificity (if any) is restricted to the nature of the hypothesis space.

Of course, the Bayesian approach is not the only statistically-grounded theoretical framework for studying language acquisition—the *connectionist* approach is similarly committed to domain-general statistical learning mechanisms (e.g., Elman et al. 1996, Prince & Smolensky 2004, Rumelhart & McClelland 1986, Smolensky & Legendre 2006). Like Bayesian models, connectionist models incorporate a notion of optimization (minimizing prediction error); nevertheless, the two approaches differ in important ways. For example, a defining feature of connectionism is the use of distributed representations. Although Bayesian models could in principle be developed using distributed representations, these are given no special status, and in fact symbolic representations (e.g., rules and categories) are typically used because they make it easier to understand and define the space of hypotheses. For this reason, Bayesian models may be more attractive to linguists who are used to symbolic representations. Another major difference between the two approaches is that Bayesian models are *declarative*, defining the learner's constraints and associated hypothesis space explicitly using mathematical equations, whereas connectionist models are *procedural*, imposing constraints only implicitly through the choice of network architectures and learning algorithms. We

[3] We note that the ability to solve the subset problem does not negate the poverty of the stimulus argument de facto, as discussed later on in the introduction.

elaborate on this distinction below, noting here only that the use of explicitly defined constraints can make it easier to understand the assumptions built into the learner and how they relate to linguistic theory or domain-general cognitive principles (i.e., whether the constraints are domain-specific or not).

As implied by the previous paragraphs, there is nothing inherent in the Bayesian approach that either favors or disfavors domain-specific constraints, and Bayesian researchers hold different views about their necessity.[4] Moreover, although Bayesian learners have certain advantages over non-statistical learners, we are not claiming that these advantages are sufficient to overcome the problem of the poverty of the stimulus (PoS) on their own, nor (we think) would other Bayesians (e.g., see Regier & Gahl (2004), Pearl & Lidz (2009), and Pearl & Mis (submitted) who also discuss the necessity of additional constraints). The PoS problem is much broader than any particular learnability problem such as the subset problem—it is a claim that the data children encounter are compatible with multiple generalizations. Even if Bayesian learning can help solve the subset problem, multiple generalizations may still be possible given the positive and indirect negative evidence available. The question, in our view, is not whether there is a PoS problem (there clearly is), but rather what kinds of constraints are needed in order to overcome it. The traditional argument from the PoS claims that the necessary constraints come from innate, domain-specific knowledge (Chomsky 1981). While the use of Bayesian learning does not automatically negate the need for domain-specific constraints, the ability to obtain information through indirect negative evidence and other properties of statistical learning may mean that a Bayesian learner is able to acquire the correct generalizations with *less* domain-specific prior knowledge than the PoS argument normally assumes. Whether *less* means *none* or simply *less detailed* is an open question that can be evaluated empirically using Bayesian modeling. For example, one way to argue in favor of a less constrained hypothesis space is to show that a Bayesian statistical learner operating within that hypothesis space is capable of acquiring the linguistic generalization of interest, i.e., that additional constraints are not needed. We can then consider whether the constraints being used are domain-general or domain-specific, and whether they are necessarily innate or could be derived from previous linguistic experience. Researchers have applied this approach to problems such as the acquisition of English anaphoric *one* (Regier & Gahl 2004, Foraker et al. 2009, Pearl & Lidz 2009, Pearl & Mis 2011), the structure-dependence of syntactic rules (Perfors, Tenenbaum, & Regier 2011), and the type of syntactic rules that account for recursion (Perfors et al. 2010).

Examples like those above show how Bayesian modeling can be used to argue that certain linguistic generalizations are learnable in principle. However, if the Bayesian framework is to be taken seriously as a way of modeling actual language acquisition, it is also important to show that children's behavior is consistent both with the assumptions of the framework and the predictions of specific models. In the remainder of this chapter,

---

[4] Some might think this is a difference from connectionism, which is often associated with an anti-nativist viewpoint (Rumelhart and McClelland 1986, Elman et al. 1996). However, the defining characteristics of the connectionist approach—e.g., distributed representations, parallel processing, and statistical learning mechanisms (see, for example, http://cognet.mit.edu/library/erefs/mitecs/smolensky.html: *Connectionist Approaches to Language*, by Paul Smolensky)—are also agnostic regarding domain-specificity. One well-known connectionist proposal that incorporates strong domain-specific constraints is Harmonic Grammar (Legendre et al. 1990, Smolensky et al. 1992, Smolensky & Legendre 2006).

we aim to do just that. We begin with the most basic assumption of the framework, namely that learning is based on statistical properties of the input data. We first review some of the wide-ranging behavioral evidence suggesting that children are indeed able to extract useful generalizations from statistical information, and can do so in a range of situations and using different types of statistics (Section 1). We then formalize the details of the Bayesian approach, expanding upon the key features mentioned above, and discuss some additional pros and cons of this approach (Section 2). Finally, we present several case studies to illustrate the ideas we have introduced and to show how Bayesian models can be applied to problems of interest in language acquisition (Section 3).

1. Experimental studies of statistical learning abilities

Although statistical properties of language were widely studied by the structuralist linguists of the 1950s (e.g., Harris 1954) , research in this area declined sharply with the rise of generative linguistics in the following decade, and only began to reemerge in the 1990s as an important topic in language acquisition. Domain-general processes of statistical learning were long recognized as part of the acquisition process even under generative theories (Chomsky 1955, Hayes & Clark 1970, Wolff 1977, Pinker 1984, Goodsitt, Morgan, & Kuhl 1993, among others), but these processes by themselves were believed to be incapable of accounting for the acquisition of complex linguistic phenomena (e.g., syntactic or phonological structure, the syntax-semantics interface) without an accompanying structured hypothesis space for those linguistic phenomena. This does not mean that researchers did not investigate the nature of the learning procedure by which the child uses the input data to disambiguate between different hypotheses and attain the correct grammar – for example, see Wexler & Culicover (1980), Dresher & Kaye (1990), Gibson & Wexler (1994), and Niyogi & Berwick (1996). However, the learning procedure was often only interesting as a tool to support the validity of a particular hypothesis space, such as the parameters hypothesis space of Chomsky (1981): a learning procedure, such as statistical learning, could demonstrate that it was possible for the child to converge on the correct hypothesis in the specified hypothesis space, given the available input data. Under this view of acquisition, the truly interesting question was about defining the child's hypothesis space appropriately – and so domain-general statistical learning was largely ignored as a research topic.

Saffran, Aslin, & Newport (1996) was an important study in this respect since it considered the nature of children's statistical learning abilities to be a question worth pursuing. Though this study was aimed at the process of word segmentation (identifying words in a fluent stream of speech) rather than more abstract knowledge acquisition at the phonological, syntactic, or semantic level, it successfully demonstrated that very young children have "powerful mechanisms for the computation of statistical properties of language input" (Saffran et al. 1996). In particular, it showed that 8-month-old infants were able to track statistical cues between syllables, and so segment novel words out from a stream of artificial language speech where the statistical information was the only cue to where word boundaries were. Saffran et al. hypothesized, and Aslin, Saffran, & Newport (1998) later confirmed, that the cue the infants were using is what they called "transitional probability". The transitional probability between syllables X and Y (e.g., "pre", "tty") is the probability that Y will occur following X, computed as the frequency

of XY ("pretty") divided by the frequency of X ("pre").[5] Pelucchi, Hay, & Saffran (2009a) later showed that infants can track transitional probability in realistic child-directed speech, as well as the artificial language stimuli Saffran et al. and Aslin et al. used.

With respect to word segmentation in natural language, Saffran et al. believed transitional probability would be a reliable cue to word boundaries, since the transitional probability of syllables spanning a word boundary would be low while the transitional probability of syllables within a word would be high. For example, in the sequence "pretty baby", the transitional probabilities between (1) "pre" and "tty" and (2) "ba" and "by" would be higher than the transitional probability between "tty" and "ba". Because of this property, they assumed that infants' ability to track transitional probability would be very useful for word segmentation in real languages (as opposed to the artificial language stimuli used in their study). Interestingly, later studies discovered that transitional probability is perhaps a less useful cue to segmentation in English child-directed speech than originally assumed (Brent 1999, Yang 2004, Gambell & Yang 2006). The precise way in which infants might use transitional probability information (if at all) for realistic language data therefore remains an open question.

Notably, however, the broader claim of Saffran et al. (1996) was not tied to transitional probability, but instead was that some aspects of acquisition may be "best characterized as resulting from innately biased statistical learning mechanisms rather than innate knowledge" that explicitly constrains the hypothesis space. Tracking syllable transitional probability is clearly one kind of statistical learning mechanism, but it need not be the only one. This led to a revitalized interest in characterizing the statistical learning abilities of children, and what types of acquisition problems could be solved by these abilities. Subsequent research has investigated a number of questions raised by these initial studies, particularly the following:

1. What kinds of statistical patterns are human language learners sensitive to?
2. To what extent are these statistical learning abilities specific to the domain of language, or even to humans?
3. What kinds of knowledge can be learned from the statistical information available?

The first question addresses the kinds of biases that are present in the human language learning mechanism, while the second question is important for understanding whether our linguistic abilities fall out from other cognitive abilities, or are better viewed as a cognitively distinct mechanism. The third question explores what can be gained if humans can capitalize on the distributional information available in the data.

Many studies have attempted to ascertain the statistical patterns humans are sensitive to. Thiessen & Saffran (2003) discovered that 7-month-olds prefer syllable transitional probability cues over language-specific stress cues when segmenting words, while 9-month-olds show the reverse preference. Graf Estes, Evans, Alibali, & Saffran (2007) found that word-like units that are segmented using transitional probability are viewed by 17-month-olds as better candidates for labels of objects, highlighting the potential utility of transitional probability both for word segmentation and subsequent

[5] This statistic is more standardly known in probability theory as the conditional probability of Y given X.

word-meaning mappings. Moving beyond the realm of word segmentation, Gómez & Gerken (1999) discovered that one-year-olds could learn both specific information about word ordering, and more abstract information about grammatical categories in an artificial language, based on the statistical cues in the input. Thompson & Newport (2007) discovered that adults can use transitional probability between grammatical categories to identify word sequences that are in the same phrase, a precursor to more complex syntactic knowledge.

It is worth pointing out that although most of the experiments described above have focused on transitional probability as the statistic of interest, researchers have begun to examine a wider range of statistical cues. These include other simple statistics involving relationships of adjacent units to one another, such as backward transitional probability (Perruchet & Desaulty 2008, Pelucchi, Hay, & Saffran 2009b) and mutual information (Swingley 2005).

Another line of work focuses on non-adjacent dependencies, and when these are noticed and used for learning. Newport & Aslin (2004) showed that learners were sensitive to non-adjacent statistical dependencies between consonants and between vowels, using either of these to successfully segment an artificial speech stream (though see Bonatti et al. (2005) and Mehler et al. (2006), who only found a preference for statistical dependencies between consonants rather than for both consonants and vowels). Additionally, learners were unsuccessful when the non-adjacent dependencies were between entire syllables, suggesting a bias in either perceptual or learning abilities. Work by Gómez (2002) has shown that learners are able to identify non-adjacent dependencies between words, but only when there is sufficient variation in the intervening word. This idea is similar to the concept of *frequent frames* introduced by Mintz (2002). A frequent frame is an ordered pair of words that frequently co-occur with one word position intervening. For example, *the___one* is a frame that could occur with *big*, *other*, *pretty*, etc.). Mintz suggests that frequent frames could be used by human learners to categorize words because they tend to surround a particular syntactic category (e.g., *the___one* tends to frame adjectives). Mintz (2002, 2006) demonstrated that both adults and infants are able to categorize novel words based on the frames in which those novel words appear.

In addition, recent experimental studies in learning mappings between words and meanings (Yu & Smith 2007, Xu & Tenenbaum 2007, Smith & Yu 2008) suggest that humans are capable of extracting more sophisticated types of statistics from their input. Specifically, the experimental evidence suggests that humans can combine statistical information across multiple situations (though see Medina et al. (2011) for some evidence that learners do not prefer to combine information across situations), and that the statistics they use cannot always be characterized as just transitional probabilities or frequent frames.

Yu & Smith (2007) and Smith & Yu (2008) examined the human ability to track probabilities of word-meaning associations across multiple trials where any specific word within a given trial was ambiguous as to its meaning. Importantly, only if human learners were able to combine information across trials could a word-meaning mapping could be determined. Both adults (Yu & Smith 2007) and 12 and 14-month-old infants (Smith & Yu 2008) were able to combine probabilistic information across trials. So, both adults and infants can learn the appropriate word-meaning mappings, given data that are uninformative within a trial but informative when combined across trials.

Xu & Tenenbaum (2007) investigated how humans learn the appropriate set of referents for basic (*cat*), subordinate (*tabby*), and superordinate (*animal*) words, a task that has traditionally been considered a major challenge for early word learning (e.g., Markman 1989, Waxman 1990) because these words overlap in the referents they apply to (a tabby is a cat, which is an animal)—an example of the subset problem. Previously, it was assumed that children had an innate bias to prefer the "basic level" in order to explain children's behavior (Markman 1989). One sophisticated statistical inference that can help with this problem is related to what Xu & Tenenbaum call a *suspicious coincidence*, and is tied to how well the observed data accord with a learner's prior expectations about word-meaning mappings. For example, suppose we have a novel word *blick*, and we encounter three examples of *blicks*, each of which is a cat. The learner at this point might (implicitly) have two hypotheses (*blick = animal, blick = cat*), and expectations associated with these two hypotheses. Specifically, if *blick = animal*, other kinds of animals besides cats should be labeled *blicks* sometimes because the set of *blicks* is larger than just the set of cats. In the language of our introduction, the fact that three *blicks* were labeled and all of them were cats provides *indirect negative evidence* against the hypothesis that *blick* means *animal*. Or, in Xu & Tenenbaum's terminology, it is a suspicious coincidence to see three cats if *blick* really means *animal*. Instead, it is more likely that *blick* is a "basic" label that is more specific, in this case *cat*. Xu & Tenenbaum (2007) discovered that both adults and children between the ages of 3 and 5 can use suspicious coincidences like this to infer the appropriate meaning of a novel word like *blick*. This suggests that humans are indeed able to perform this sophisticated statistical inference.

Turning to the question of domain-specificity for human statistical learning abilities, Saffran et al. (1999) showed that both infants and adults can segment non-linguistic auditory sequences (musical tones) based on the same kind of transitional probability cues that were used in the original syllable-based studies. Similar results have been obtained in the visual domain using both temporally ordered sequences of stimuli (Kirkham et al., 2002) and spatially organized visual "scenes" (Fiser & Aslin, 2002). Conway & Christiansen (2005) adapted the grammar from Gómez & Gerken's (1999) experiments to explore learning in different modalities: auditory, visual, and tactile. They showed that adults could learn grammatical generalizations in all three modalities, although there was a quantitative benefit to the auditory modality, as well as some qualitative differences in learning. These results (particularly those in the tactile modality, which is not used in natural languages) support the idea that the kinds of statistical learning seen in the earlier artificial language studies are highly domain-general, showing robustness across modalities and presentation formats.

Another way of investigating whether particular learning abilities could in principle be specific to language is by comparing learning across species. If non-human animals are able to learn the same kinds of generalizations as humans, then whatever cognitive mechanism is responsible must not be a linguistic one. To this end, Hauser et al. (2001) exposed cotton-top tamarins to the same kind of artificial speech stimuli used in the original Saffran et al. (1996) segmentation experiments, and found that the monkeys were able to perform the task as well as infants. Saffran et al. (2008) later found that tamarins could also learn some simple grammatical structures based on statistical information, but were unable to learn patterns as complex as those learned by infants. This suggests that

infants' abilities to extract information from statistical patterns are more powerful than those of other animals. Additional evidence is provided by the experiments of Toro & Trobalon (2005), who showed that rats were able to segment a speech stream based on syllable co-occurrence frequency (similar to the mutual information explored in Swingley (2005)), but not transition probability alone. The rats also showed no evidence of learning generalizations from non-adjacent dependencies such as those in the Gómez (2002) experiments, or abstract rules as in Marcus et al. (1999).

The main lesson from the experimental evidence reviewed in this section is that children do seem capable of using statistical information in their language input, from tracking simple statistical cues like transitional probability to making sophisticated inferences that combine ambiguous information from multiple data sources. To learn more about the abilities and biases of human learners, researchers continue to investigate the statistical information humans are sensitive to, and what kinds of generalizations are learned from them. In addition, experiments using other modalities, domains, and species can help to shed light on the question of whether these abilities are domain-specific or domain-general.

This kind of experimental research is undoubtedly important for our understanding of the role of statistical learning in language acquisition. However, the third question of what knowledge can be learned from the statistical information available can be addressed more easily, or in a complementary fashion, through other research methodologies such as Bayesian modeling, which we turn to in the next section.

2. An introduction to the Bayesian modeling framework

As noted in the introduction, Bayesian modeling offers a concrete way to examine what knowledge is required for acquisition, without committing *a priori* to a particular view about the nature of that knowledge. It also addresses the question of whether human language learners can be viewed as being *optimal* learners in a sense that will become clearer below once we formalize the approach. We expand on both of these points in this section, starting with a conceptual introduction to the Bayesian approach that explains the kinds of questions it can answer and how these differ from the questions addressed by other approaches. Next, we describe the formal implementation of a Bayesian model, in particular how it operates over an explicitly defined hypothesis space. We then highlight some attractive features of Bayesian models with respect to the kind of hypothesis spaces they can operate over, and conclude with a brief discussion of some of the algorithms commonly used in Bayesian modeling.

2.1. Bayesian modeling as a computational-level approach

Most models of language acquisition are *procedural*: they hypothesize specific procedures or algorithms that can be applied to the input and/or grammar in order to produce linguistically meaningful generalizations. For example, learners might segment words by identifying syllable sequences with high frequency and mutual information (Swingley, 2005), create a grammatical category by grouping together words that share a frequent frame (Mintz 2003, Wang & Mintz 2008, Chemla et al. 2009), use back-propagation to change the set of weights in a neural network (Elman 1990, Elman 1993), or demote the ranking of a constraint if it causes an error in parsing the input (Boersma & Hayes 2001, Prince & Tesar 2004, Tesar & Smolensky 2000). These kinds of models

provide what Marr (1982) calls *algorithmic-level* explanations, focusing on the question of *how* learners generalize from their input. In contrast, the Bayesian approach investigates the problem of language acquisition at Marr's (1982) *computational level* of analysis, seeking answers to the questions of *what* computational problem is being solved and *why* the learner ends up with a particular solution. This kind of investigation calls for a *declarative* (rather than procedural) model of the learner. That is, in designing a Bayesian model, the researcher considers what the nature of the learning task is (i.e., what does the learner need to achieve), what sources of information are available, and what the inductive biases of the learner are (i.e., what kinds of generalizations/grammars are easy, difficult, or impossible to learn). It is then possible to ask what will be learned, given particular assumptions about these aspects of the problem and also assuming that the learner behaves optimally under those assumptions. This kind of approach is often referred to as an *ideal observer* (or *ideal learner*) analysis, since it explores the solutions that would be found by an idealized optimal learner capable of extracting the necessary statistical information from the input. The idea of optimality leads naturally to the use of probability theory for defining Bayesian models, because probability theory is a tool for determining optimal behavior under uncertainty.

Some readers may not be comfortable with the idea of humans as optimal statistical learners, especially since well-known early studies in other areas of cognition suggested just the opposite (Cascells, Schoenbeger, & Grayboys 1978). However, the rational analysis view of cognition (Chater & Oaksford, 1999; Anderson, 1990) has countered by arguing that human behavior is adapted to our natural environment and the tasks we must achieve there – thus, "optimal" behavior must be interpreted within that context, rather than within the context of a laboratory experiment. Behavioral and modeling work has supported the idea of humans as optimal learners in areas such as numerical cognition (Tenenbaum 1996), causal induction (Griffiths & Tenenbaum 2005), and categorization (Kemp, Perfors, & Tenenbaum 2007). More recently, evidence has begun to accumulate in language acquisition as well (Feldman et al. 2009a, Xu & Tenenbaum 2007).

Whether or not humans behave optimally in all situations, the kind of ideal learner analysis provided by a Bayesian model is still useful for answering two kinds of questions. First, the question of learnability: what is possible to learn from the available input, given particular assumptions about the learner's inductive biases? Second, once we have identified the optimal solution to the problem as defined by the model, we can ask whether human behavioral data is consistent with the model's predictions. If so, then we have helped to explain why humans behave in this way -- it is the optimal response to the data they are exposed to. If not, then we can begin to investigate how and why humans might differ from the optimal behavior (Goldwater et al. 2009, Frank et al. 2010).

Although these are worthwhile questions to investigate, some researchers still find the Bayesian approach unsatisfying because of its focus on computational-level explanations. In particular, Bayesian models often do not address how the learner might perform the computations required to achieve the optimal solution to the learning problem, even if such a solution is achieved. Rather, they simply state that if human behavior accords with the predictions of the model, then humans must be performing *some* computation (possibly a very heuristic one) that allows them to identify the same optimal solution that the model did. We discuss this issue further, including some responses to it, in section 2.4.

2.2. Formalizing the Bayesian approach

Bayesian models assume the learner comes to the task with some space of hypotheses $\mathcal{H}$, each of which represents a possible explanation of the process that generated the input data. The hypothesis space could be discrete (e.g., a finite or infinite set of symbolic grammars) or continuous (e.g., a set of real-valued parameters representing the tongue positions necessary to produce a particular set of vowels). Given the observed data d, the learner's goal is to determine the probability of each possible hypothesis h, i.e., to estimate $P(h|d)$, the *posterior distribution* over hypotheses. A correct estimate of the posterior distribution will allow the learner to behave optimally in the future, i.e., to have the best chance of interpreting and/or generating new data in accordance with the true hypothesis (the one that actually generated the observed data).

Rather than estimating $P(h|d)$ directly, we first apply Bayes' Theorem, derived from the axioms of probability theory, to reformulate it as in (1):

(1) Bayes' Theorem

$$P(h \mid d) = \frac{P(d \mid h)P(h)}{P(d)}$$

where $P(d|h)$, the *likelihood*, expresses how well the hypothesis explains the data, and $P(h)$, the *prior*, expresses how plausible the hypothesis is regardless of any data. $P(d)$, the *evidence,* is a constant normalizing factor that ensures that $P(h|d)$ is a proper probability distribution, summing to 1 over all values of h. Often we only care about the *relative* probabilities of different hypotheses, in which case we can ignore the denominator and simply write Bayes' Theorem as a proportionality, as in (2):

(2) $P(h \mid d) \propto P(d \mid h)P(h)$

Defining a Bayesian model usually involves three steps:

(1) Defining the hypothesis space: Which hypotheses does the learner consider?
(2) Defining the prior distribution over hypotheses: Which hypotheses is the learner biased towards or against?
(3) Defining the likelihood function: How is the observed data generated under a given hypothesis?

A simple example, adapted from Griffiths & Yuille (2006), should help to clarify these ideas. Suppose you are given three coins, and told that two of them are fair, and one produces heads with probability 0.9. You choose one of the coins at random and must determine whether it is fair or not, i.e., whether θ (the probability of heads) is 0.5 or 0.9. Thus, the hypothesis space contains two hypotheses: h_0 ($\theta = 0.5$) and h_1 ($\theta = 0.9$), with $P(h_0) = 2/3$ and $P(h_1) = 1/3$. Data is obtained by flipping the coin, with the probability of a particular sequence d of flips containing s heads and t tails being dependent on θ, as $P(d|\theta) = \theta^s(1-\theta)^t$. For example, if $\theta = 0.9$, then the probability of the sequence HHTTHTHHHT is 0.0000531. If $\theta = 0.5$, then the same sequence has

probability 0.000978. To determine which hypothesis is more plausible given that particular sequence, we can compute the *posterior odds ratio* as in (3):

(3) Posterior Odds Ratio

$$\frac{P(h_1 \mid d)}{P(h_0 \mid d)} = \frac{\frac{P(d \mid h_1)P(h_1)}{P(d)}}{\frac{P(d \mid h_0)P(h_0)}{P(d)}} = \frac{P(d \mid h_1)P(h_1)}{P(d \mid h_0)P(h_0)} = \frac{(0.0000531)(1/3)}{(0.000978)(2/3)} \approx \frac{1}{37}$$

This tells us that the odds in favor of h_0 are roughly 37:1. Note that the $P(d)$ (*evidence*) term cancels, so we do not need to compute it.

This very simple example illustrates how to compare the plausibility of two different hypotheses, but in general the same principles can be applied to much larger and more complex hypothesis spaces (including countably infinite spaces), such as might arise in language acquisition. With minor modifications, we can also use similar methods to compare hypotheses in a continuous (uncountably infinite) space (see Griffiths & Yuille (2006) for a more explicit description of the modifications required). Such a space might occur in a syntax-learning scenario if we suppose that the hypotheses under consideration consist of probabilistic context-free grammars (PCFGs), with different grammars varying both in the rules they contain, and the probabilities assigned to the rules.[6] The input data in this situation could be a corpus of sentences in the language, with $P(d|h)$ determined by the rules for computing string probabilities under a PCFG (Chater & Manning 2006). $P(h)$ could incorporate various assumptions about which grammars the learner might be biased towards -- for example, grammars with fewer rules, or grammars that incorporate linguistically universal principles. See section 3 below for more detailed examples of how these ideas can be applied to language acquisition.

2.3. Bayesian hypothesis spaces

As mentioned above, a Bayesian learner can operate over a variety of hypothesis spaces (discrete, continuous, countably infinite, uncountably infinite, etc.), without changing the underlying principles of a Bayesian learner. Another useful property of Bayesian models is that the hypothesis space can be highly structured, supporting multiple levels of linguistic representation simultaneously. For example, the word segmentation model of Goldwater et al. (2006, 2009) contains two levels of representation -- words and phonemes -- though only one of these (words) is unobserved in the input and must be learned. However, Bayesian models can in principle learn multiple levels of latent structure simultaneously, and doing so can even improve their performance. For example, Johnson (2008) showed that learning both syllable structure and words from unsegmented phonemic input improved word segmentation in a Bayesian model similar to that of Goldwater et al. Feldman, Griffiths, & Morgan (2009a) compared two Bayesian models of phonetic category acquisition to demonstrate that simultaneously

[6] Since probabilities are represented using real numbers, the hypothesis space is continuous; if the learner is assumed to acquire a non-probabilistic grammar, then the hypothesis space consists of a discrete set of grammars.

learning phonetic categories and the lexical items containing those categories led to more successful categorization than learning phonetic categories alone. Dillon, Dunbar, & Idsardi (2011) also compared two Bayesian models of phonetic category learning: one that first learned phonetic categories and would later identify allophones and phonological rules based on those phonetic categories, and one that learned all the information (phonetic categories, allophones, and rules) at once. Again, the joint learner was more successful. By allowing us to build such joint models and compare them to staged learning models, the Bayesian approach is helpful for understanding the process of *bootstrapping* -- using preliminary or uncertain information in one part of the grammar to help constrain learning in another part of the grammar, and vice versa.

In addition to including multiple levels of structure, the predefined hypothesis space of a Bayesian learner can be instantiated very abstractly, which should appeal to generative linguists who believe abstract linguistic parameters determine much of the constrained variation observed in the world's languages (Chomsky 1981). Kemp, Perfors, & Tenenbaum (2007) and Kemp & Tenenbaum (2008) discuss *overhypotheses* in Bayesian modeling, where overhypotheses refer to strong inductive constraints on possible hypotheses in the hypothesis space (Goodman 1955). This idea is intuitively similar to the classic notion of a linguistic parameter as an abstract (structural) property that constrains the hypothesis space of the learner. To see how, consider first a very simple example illustrating the idea of an overhypothesis, taken from Goodman (1955) and presented in Kemp, Perfors, & Tenenbaum (2007). Suppose a learner is presented with several bags of marbles, where marbles can be either black or white. During training, the learner is allowed to examine all of the marbles in each bag, and finds that each bag contains either all black or all white marbles. During testing, the learner draws only a single marble from a bag and must predict the color distribution in the bag. Possible *hypotheses* are that the bag contains all black marbles, all white marbles, 70% black and 30% white, or any other combination. Possible *overhypotheses* are that all bags contain a uniform color distribution, all bags contain the same distribution, all bags contain a mixture of colors, etc. By observing (during training) several different bags that all have uniform color distributions, the learner learns to assign high probability to the overhypothesis of uniform color distribution. This overhypothesis in turn constrains the hypotheses for individual new bags observed – high probability is given to "all black" and "all white" before ever observing a marble from the bag, while low probability is given to hypotheses like "70% black and 30% white". This example demonstrates how information can be indirectly used (i.e. at a very abstract level) to make predictions, e.g., observing all black bags and all white bags allows the prediction that a bag with mixed black and white marbles has low probability of occurring.

Translating this to a linguistic example, suppose the marble bags are individual sentences. Some sentences contain verbs and objects, and whenever there is an object, suppose it appears after the verb (e.g., *see the penguin,* rather than *the penguin see*). Other sentences contain modal verbs and nonfinite main verbs, and whenever both occur, the modal precedes the main verb (e.g., *could see*, rather than *see could*). A learner could be aware of the shared structure of these sentences – specifically that these observable forms can be characterized as the head of a phrase appearing before its complements (specifically $[_{VP}$ V NP] and $[_{IP}$ Aux VP]) – and could encode this "head-first" knowledge at the level that describes sentences in general, akin to a hierarchical Bayesian learner's

overhypothesis. This learner would then infer that sentences in the language generally have head-first structure. If this learner then saw a sentence with a preposition surrounded by NPs (e.g. *penguins on icebergs are cute*), it would infer that the preposition should precede its object ([$_{NP}$ *penguins* [$_{PP}$ *on icebergs*]] and not [$_{NP}$ [$_{PP}$ *penguins on*] *icebergs*]), even if it had never seen a preposition used before with an object. In this way, the notion of a head-directionality parameter can be encoded in a Bayesian learner at the level of an overhypothesis. In particular, inferences (e.g., about prepositional phrase internal structure) can be made on the basis of examples that bear on the general property being learned (e.g., head directionality), even if those examples are not examples of the exact inference to be made (e.g., verb phrase examples).

2.4. Algorithms

It is worth reiterating that, unlike neural networks and other algorithmic-level models such as those of Mintz (2003), Swingley (2005), and Wang & Mintz (2008), Bayesian models are intended to provide a declarative description of what is being learned, not necessarily how the learning is implemented. Bayesian models predict a particular posterior distribution over hypotheses given a set of data, and can also be used to make predictions about future data based on the posterior distribution. If human subjects' performance in a task is consistent with the predictions of the model, then we can consider the model successful in explaining what has been learned and which sources of information were used in learning. However, we do not necessarily assume that the particular algorithm used by the model to identify the posterior distribution is the same as the algorithm used by the humans. We only assume that the human mind implements some type of algorithm (as mentioned previously, perhaps a very heuristic one) that is able to approximately identify the posterior distribution over hypotheses.

In practice, most Bayesian models of language acquisition have used Markov chain Monte Carlo algorithms such as Gibbs sampling to obtain samples from the posterior distribution (Gilks et al., 1996; Geman & Geman, 1984; also see Resnik & Hardisty (2009) for an accessible tutorial). These are batch algorithms, which operate over the entire data set simultaneously. This is clearly an unrealistic assumption about human learners, who must process each data point as it is encountered, and presumably do not revisit or reanalyze the data at a later time (or at most, are able to do so only to a very limited degree). If humans are indeed behaving as predicted by Bayesian models, they must be using a very different algorithm to identify the posterior distribution over hypotheses – an algorithm about which most Bayesian models have nothing to say.

Researchers who are particularly concerned with the mental mechanisms of learning often find the Bayesian approach unsatisfactory precisely because in its most basic form, it does not address the question of mechanisms (e.g., see McClelland et al. (2010), and Griffiths et al. (2010) for a reply). In response to this kind of critique, a recent line of work has begun to address the question of how learners might implement Bayesian predictions in a more cognitively plausible way. These kinds of models are sometimes called *rational process models*, since they are models of rational learners that are concerned with implementing the process of approximating Bayesian inference. For example, Shi, Griffiths, Feldman, & Sanborn (2010) discuss how exemplar models may provide a possible mechanism for implementing Bayesian inference, since these models allow an approximation process called importance sampling. Other examples include the

work of Bonawitz et al. (2011), who discuss how a simple sequential algorithm can be used to approximate Bayesian inference in a basic causal learning task, and that of Pearl, Goldwater, and Steyvers (2011), who (as described in section 3) investigated various online algorithms for Bayesian models of word segmentation. See also McClelland (1998) for a discussion of how neural network architectures can be used to approximate optimal Bayesian inference (again emphasizing that the connectionist and Bayesian frameworks are not so much in opposition as they are addressing different aspects of the learning problem, with one focusing on the description of the task and the other focusing on the implementation).

3. Specific example studies

This section surveys a few representative studies in different areas of language acquisition in order to illustrate how Bayesian modeling can be applied within each domain. For each study, we review the problem faced by the learner, describe the hypothesis space assumed by the model and how Bayesian inference operates within it, and discuss the results with reference to relevant behavioral data.

3.1. Phonetics and phonology

Feldman, Griffiths, & Morgan (2009b), Feldman (2011), and Feldman, Griffiths, Goldwater, & Morgan (submitted) address the question of phonetic category acquisition, specifically the acquisition of vowel categories. This is a difficult problem because of the variation in acoustic properties between different tokens of the same vowel, even when spoken by the same speaker. Although the mean formant values of different vowel categories are different, the distribution of values overlaps considerably, e.g., a particular token of /e/ may sound exactly like a token of /ɛ/, even if spoken by the same individual. Figure 1 illustrates this variation in men's vowels sounds.

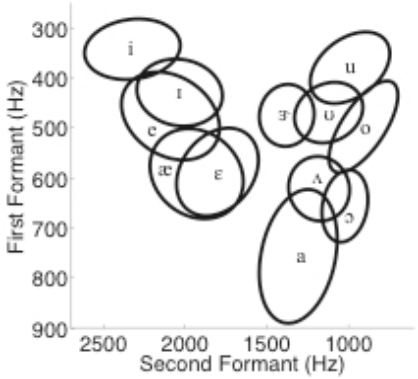

Figure 1. (Reproduced from Feldman et al. (2009b)). Example distribution of men's vowel sounds. Many vowel sounds have overlapping distributions, such as /e/ and / ɛ/.

Experimental studies suggest that infants are able to learn separate phonetic categories for speech sounds that occur with a clear bimodal distribution (Maye, Werker, & Gerken, 2002, Maye & Weiss, 2003), but the extent of overlap between phonetic categories in real speech suggests that some categories might be difficult to distinguish in this way. Instead, Feldman et al. hypothesize that learners must make use of an additional source of information beyond the acoustic properties of individual sounds;

specifically, they take into account the words those sounds occur in (an idea also advocated by Swingley (2009)). Of course, young infants who are still learning the phonology of their language have very little lexical knowledge. Indeed, Feldman et al. review experimental studies suggesting that phonetic categorization and word segmentation (a precursor to word-meaning mapping) occur in parallel, between the ages of 6-12 months. So, rather than assuming either that phonetic categories are acquired first and then used to learn words, or that words are acquired first and then used to disambiguate phonetic categories, Feldman et al. propose a joint model in which phonetic categories and word forms are learned simultaneously. They compare this model to a simpler baseline model in which phonetic categories alone are learned. We describe each of these models briefly before reviewing the results.

Feldman et al.'s baseline model is a distributional model of categorization: it assumes that phonetic categories can be identified based on the distribution of sounds in the data. In particular, it assumes that the tokens in each phonetic category have a Gaussian (normal) distribution, and the goal of the learner is to identify how many categories there are, and which sounds belong to which categories. Since the number of categories is unknown, Feldman et al. use a *Dirichlet process* prior (Ferguson, 1973), a distribution over categories that does not require the number of categories to be known in advance. The Dirichlet process favors categorizations that contain a smaller number of categories, unless the distributional evidence suggests otherwise. In other words, if there is good reason to assume that a set of sounds are produced from two different categories (e.g. because they have a strongly bimodal distribution, leading to a low likelihood if collapsed into a single Gaussian category), then the model will split the sounds into two categories; otherwise it will assign them to a single category.

Feldman et al.'s second model is a lexical-distributional model, which assumes that the input consists of acoustically variable word forms rather than just sequences of phonetic tokens (i.e., that the child recognizes that the phonetic tokens are part of larger units). The learner now has two goals: to find phonetic categories (as in the distributional learner) and also to recognize acoustically distinct word forms as variants of the same lexical item, grouping together tokens that contain the same sequence of phones. Note that these two tasks are interdependent. On the one hand, the categorization of phonetic tokens affects which word tokens are considered to be the same lexical item. On the other hand, if two word tokens are assigned to the same lexical item, then the acoustic tokens comprising them should belong to the same phonetic categories. The hypothesis space for this model thus consists of pairs of categorizations (of acoustic tokens into phonetic categories, and word forms into lexical items). Since the lexical learning task is also viewed as categorization, it is modeled using another Dirichlet process, which prefers lexicons containing fewer items when possible.

Using a small hand-constructed data set, Feldman et al. show that the lexical-distributional model makes a counterintuitive prediction about minimal pairs (i.e., words that differ by a single phoneme). Specifically, if a pair of sounds (say, B and C) only occur within minimal pairs (say, lexical items AB, AC, DB, DC), then they are likely to be categorized as a single phoneme if they are acoustically similar, since this would reduce the size of the lexicon, replacing four words with two (AX, DX). On the other hand, if B and C occur in different contexts (say, AB and DC only), then they are more likely to be categorized as separate phonemes. This is because the lexical-distributional

learner can use phonemes A and D to recognize that AB and DC are different words, and then use this information to recognize that the distribution of B and C are actually slightly different. This prediction is interesting for two reasons. First, it means that the lack of minimal pairs in early vocabularies (e.g., see Dietrich, Swingley, & Werker 2007) may actually be helpful. Secondly, recent experiments by Thiessen (2007) seem to bear out the model's prediction in a word learning task with 15-month-olds: infants are better at discriminating similar-sounding object labels (e.g., *daw* vs. *taw*) after being familiarized with non-minimal pairs containing the same sounds (*dawbow, tawgoo*).

In a second simulation, Feldman et al. compared the performance of their distributional model, lexical-distribution model, and a second distributional model (Vallabha et al. 2007) on a larger corpus containing 5000 word tokens from a hypothetical set of lexical items containing only vowels (e.g., "aei" - vowel-only words were necessary because the model can only learn vowel categories). Both of the distributional models identified too few phonetic categories, collapsing highly overlapping categories into one category. In contrast, the lexical-distributional learner was much more successful in distinguishing between very similar categories. Although these results are preliminary and still need to be extended to more realistic lexicons, they provide intriguing evidence that simultaneously learning linguistic generalizations at multiple levels (phonetic categories and word forms) can actually make the learning problem easier than learning in sequence.

Dillon, Dunbar, & Idsardi (2011) have also recently studied the acquisition of phonemes and phonological rules from acoustic data, using a Bayesian model. Like Feldman et al. (2009b, submitted), they recognize that word forms are comprised of phonetic categories. However, they also note that a phoneme is a more abstract representation that may relate multiple phonetic categories across word forms (known as *allophones* of the phoneme). For example, in Spanish, there is a single phoneme /b/ that is realized in distinct ways, depending on the surrounding linguistic context: between two vowels, it is pronounced as the fricative /β/, while in all other contexts, it is pronounced as the stop /b/. These two pronunciations are distinct phonetic categories that appear in different lexical items. The Lexical-Distributional Model of Feldman et al. would likely recognize these as separate phonetic categories precisely because they appear in different linguistic contexts. However, it would not recognize them as being allophones of the phoneme /b/, related by a phonological rule that is conditioned on the surrounding phonemes; instead, learning that mapping from two phonetic categories to a single phoneme would occur in a subsequent stage of learning. While this is a reasonable model of acquisition, it nonetheless implies a learning sequence where learning phonetic categories must happen before learning phonemes and phonological rules. Dillon et al. (2011) explore whether relaxing this assumption could lead to better learning.

In particular, Dillon et al. (2011) investigate the vowel system of Inuktitut, which has three phonemes with two allophones each (for a total of six phonetic categories). This kind of vowel system is not uncommon in the world's languages (e.g., it is shared by Quechua and many dialects of Arabic), and so represents a realistic learning problem. They compare a learner that attempts to first identify the phonetic categories from the acoustic data (and would only later hypothesize phonemes and phonological rules) to a learner that learns phonetic categories, phonemes, and phonological rules simultaneously. Their findings are similar to Feldman et al's more general finding: Learning multiple

levels of representation simultaneously can be a better strategy than trying to learn them in sequence. In particular, Dillon et al (2011) found that the learner who only identifies phonetic categories will converge on phonetic categories that make it much harder to formulate the correct phonological rule (and so define the correct phonemes). This problem occurs because the learner disregards the linguistic context when identifying its categories, and uses only the acoustic information. In contrast, if the learner is trying to identify context-sensitive phonological rules at the same time that it is identifying phonetic categories, then it views the linguistic context as informative. This learner identifies phonetic categories that are conducive to formulating phonological rules based on linguistic context; it can then find the correct phonemes (and allophones) for the language.

An interesting acquisition trajectory prediction that comes from Dillon et al's single-stage model is that children should have some knowledge of phonemes even while they're learning the phonetic categories of their language, as opposed to passing through a preliminary stage where they have solid knowledge of phonetic categories but little knowledge of phonological rules and phonemes. To our knowledge, the infant perceptual experimental literature does not currently distinguish between these possibilities, which suggests an area of future research.

3.2. Word segmentation

There have been a number of recent papers on Bayesian modeling of word segmentation. These are all based on the models presented in Goldwater (2006) and Goldwater, Griffiths, & Johnson (2009), which make the simplifying assumption (shared by most other computational models of word segmentation) that the input to the learner consists of a sequence of phonemes, with each word represented consistently using the same sequence of phonemes each time it occurs. Between-utterance pauses are represented as spaces (known word boundaries) in the input data, but other word boundaries are not represented. So, the input corresponding to the two utterances "see the kitty? look at the kitty!", transcribed using the phonemic representation used by Goldwater et al., would be **`siD6kIti lUk&tD6kIti`** (or, represented orthographically for readability, *seethekitty lookatthekitty*).

The hypothesis space considered by the learner consists of all possible segmentations of the data (e.g., *seethekitty lookatthekitty, s e e t h e k i t t y l o o k a t t h e k i t t y, seet he k itty loo k att he k itty, see the kitty look at the kitty,* etc.). In this model, $P(d|h)$ is 1 for all of these segmentations because they are all completely consistent with the unsegmented data (in the sense that concatenating the words together produces the input data).[7] Consequently, the segmentation preferred by the model is the one with the highest prior probability. The prior is defined, as in the Feldman et al. (2009) models, using a Dirichlet process, which assigns higher probability to segmentations that contain relatively few word types, each of which occurs frequently and contains only a few

[7] In fact, the full hypothesis space for the model consists of all possible sequences of potential words, including those that are inconsistent with the observed data, such as *have some pizza* and *gix blotter po nzm.* However since these sequences are inconsistent with the data, $P(d|h) = 0$, and these hypotheses can be disregarded.

phonemes. In other words, the model prefers segmentations that produce smaller lexicons with shorter words.

Goldwater et al.'s (2009) computational studies were purely theoretical, with the aim of examining what kinds of segmentations would be preferred by a learner making the assumptions above, as well as one of two additional assumptions: either that words are statistically independent units (a *unigram* model), or that words are units that predict each other (implemented in this case using a *bigram* model). While it is clear that the second of these assumptions holds in natural language, the first assumption is simpler (because the learner only needs to track individual words, rather than dependencies between words). So, if infants' ability to track word-to-word dependencies is limited, then it is worth knowing whether the simpler model might allow them to achieve successful word segmentation anyway. Goldwater et al. found that the optimal segmentation for their unigram model (in fact for any reasonable unigram model) is one that severely undersegments the input data -- the word boundaries it finds tend to be very accurate, but it does not find as many boundaries as actually exist. Thus, it produces "chunks" that contain more than one word. The bigram model is nearly as precise when postulating boundaries, but identifies far more boundaries overall, leading to a more accurate segmentation. This study is a good example of an ideal observer analysis, demonstrating the behavior of optimal statistical learners given the available input and certain assumptions about the capabilities of the learners (i.e., whether the learner knows to track word-to-word dependencies or not). Although there is some evidence that children do undersegment when learning words (Peters 1983), it is not clear whether they do so to the same degree as the unigram model, whether their segmentations are more similar to the bigram model, or neither. Thus this study by itself does not tell us whether human behavior is actually consistent with either of the proposed ideal learners, or in what situations, or how more limited (non-ideal) learners might differ from the ideal. Follow-up work by Goldwater and colleagues has begun to address these questions through experimental and computational studies.

In the work of Frank et al. (2007, 2010), the authors examine the predictions of Goldwater et al.'s unigram word segmentation model, as well as that of several other models, and compare these predictions to human performance in several experiments.[8] The experiments are modeled on those of Saffran et al. (1996), and involve segmenting words from an artificial language based on exposure to utterances containing no pauses or other acoustic cues to word boundaries. Frank et al. performed three experiments, manipulating either the number of word tokens in each utterance (1-24 words), the total number of utterances (thus, word tokens) heard in the training phase (48-1200 words), or the number of lexical items in the vocabulary (3-9 lexical items).

In the experiment that manipulated the length of utterances, Frank et al. found that humans had more difficulty with the segmentation task as the utterance length increased, with a steep drop-off in performance between one and four words, and a more gradual decrease thereafter. Several of the models captured the general decreasing trend, but the Bayesian model correlated better with the human results than all other models tested. The Bayesian model's results can be interpreted as a competition effect: longer utterances have more possible segmentations, so there is a larger hypothesis space for the model to

[8] The unigram model was used because in these experiments, words really are almost statistically independent, so the bigram model would have provided little or no benefit.

consider. Although most hypotheses have very low posterior probability, nevertheless as the hypothesis space increases, the total probability mass assigned to all the incorrect hypotheses begins to grow.

In the experiment that manipulated the amount of exposure, subjects' performance improved as exposure increased, but again there was a non-linear effect, with greater improvement initially followed by a more gradual improvement later on. Again, the Bayesian model captured this effect better than the other models. The Bayesian model incorporates a notion of statistical evidence (more data leads to more certainty in conclusions), while many of the other models do not. For example, Frank et al. tested a transitional probability model and found that its performance changes very little over time because it only requires a few utterances to correctly estimate the transitional probabilities between syllables, after which the transitional probabilities do not change with more data.

Taken together, these two experiments show that the Bayesian model incorporates notions of competition and accumulating evidence in ways that predict human segmentation behavior more effectively than other models, at least with respect to the effects of utterance length and exposure time. This suggests that in some ways humans do behave like an optimal statistical learner. However, the third experiment, which manipulated vocabulary size, showed that in other ways, humans are not like an ideal learner (or for that matter, like other proposed statistical learners). In this experiment, subjects found languages with larger vocabularies more difficult to segment than those with smaller vocabularies. Although this finding was not surprising - intuitively, larger vocabularies impose greater memory demands - all of the models predicted exactly the opposite result. The models have perfect memory, so storing a larger vocabulary poses no difficulty. At the same time, a larger vocabulary makes the sequences of syllables that are true words more statistically distinct from the sequences that are not words. For example, with a three-word vocabulary (words A, B, C), an incorrect segmentation where the hypothesized words are all the possible two-word combinations (AB, AC, BA, BC, CA, CB) scores not much differently from the correct segmentation under the Bayesian model -- one hypothesis has three words in the vocabulary, whereas the other has six. In contrast, if there are nine words in the vocabulary, then the analogous incorrect segmentation would require 72 vocabulary items, a much bigger difference from nine. Similarly, in a transitional probability model, transitions across words in a three-word language have relatively high probability, whereas transitions across words in a nine-word language have much lower probability, making them more distinct from within-word transitions. Thus, although humans performed most similarly to the Bayesian ideal learner model in the first two experiments, the third experiment provides an example where human performance differs from the statistically optimal solution assuming perfect memory.

The above discussion suggests that in order to successfully model human behavior in some language acquisition tasks, it is necessary to account for human memory limitations. Frank et al. present several possible modifications to Goldwater et al.'s (2009) Bayesian model that incorporate such limitations through algorithmic means, and find that these are able to correctly model the data from all three experiments. Similar kinds of modifications were also explored by Pearl, Goldwater, & Steyvers (2011) in the context of word segmentation from naturalistic corpus data. Like Frank et

al., Pearl et al. wanted to examine cognitively plausible algorithms that could be used to implement an approximate version of Goldwater et al.'s Bayesian model.

To simulate limited cognitive resources, all the algorithms explored in Pearl et al. (2011) processed utterances one at a time, rather than in a batch as the ideal learner of Goldwater et al. (2009) did. Two algorithms used variants of a method called dynamic programming, which allows a learner to efficiently calculate the probability of all possible segmentations for a given utterance. A third algorithm attempted to additionally simulate the human memory decay process, and so focus processing resources on data encountered more recently. This algorithm was a modified form of the Gibbs sampling procedure used for ideal learners, and is called decayed Markov Chain Monte Carlo (DMCMC) (Marthi et al. 2002). Notably, the DMCMC algorithm can be modified so it does significantly less processing than the ideal learner's Gibbs sampling procedure (for the simulations in Pearl et al, the DMCMC algorithm did 89% less processing than the ideal learner's algorithm).

Simulations using these algorithms showed that in most cases, constrained learners were nearly as successful at segmentation as the ideal learner, despite their processing and memory limitations. These results suggest that children may not require an infeasible amount of processing power to identify words using an approximation of Bayesian learning. On the other hand, Pearl et al. found that constrained learners did not always benefit from the bigram assumption which was helpful to the ideal learner, perhaps because those constrained learners lacked sufficient processing resources to effectively exploit that information.

Interestingly, Pearl et al. also found that some of their constrained learners actually *outperformed* the ideal learner when the learners used a unigram assumption. This is a somewhat counterintuitive finding, since we might naturally assume that having more memory and more processing power (like the ideal learner has) is always better. However, these results are compatible with the "Less is More" hypothesis (Newport 1990), which suggests that fewer cognitive resources may actually be beneficial for language acquisition. This turned out to be true in the unigram learners Pearl et al. examined. In particular, a property of all unigram learners is that they will undersegment frequent short words that often appear in sequence (like *it's* and *a*), preferring to make them a single word (*itsa*) in order to explain why they appear together so frequently. Because the ideal learner has a perfect memory, it can see all the data at once and realize that this sequence of phonemes often occurs. In addition, because the ideal learner also has more processing resources, it has more opportunity to "fix" a mistaken hypothesis that these are two separate lexical items instead of one. In contrast, the constrained learner lacks both the ability to see all the data at once and the processing resources to easily fix a "mistake" it made earlier on in learning (where the "mistake" is viewing the phoneme sequence *it's a* as two lexical items). As such, it does not make the undersegmentation errors the ideal learner does. Though the Bayesian modeling studies discussed here are preliminary and the robustness of the results should be verified on other languages, they provide a tantalizing example of this idea that is used to explain children's excellent language acquisition abilities.

3.3. Word-meaning mapping

There have been two notable recent studies involving Bayesian models for learning word-meaning mappings. In Section 1 we briefly mentioned some experimental results from one of these, Xu & Tenenbaum (2007), and refer the reader to that paper for a description of the computational aspects of the study. Here we discuss instead the work of Frank, Goodman, & Tenenbaum (2009), who developed a Bayesian model that incorporates both non-linguistic context and speaker intentions in learning noun-object mappings. Their model assumes that the words uttered by a speaker (and therefore observed by the learner) are determined by the process represented schematically in Figure 2. Given the set of objects *O* that are currently present, the speaker chooses some subset of those objects *I* as intended referents. The speaker also has a lexicon *L* containing one or more labels for each object. The utterance *W* contains one referring word for each intended referent, where that word is chosen at random from the labels available in *L* for the referent. *W* can also include non-referring words (verbs, determiners, etc.), but the model is set up in such a way that it prefers words to be referential if possible.

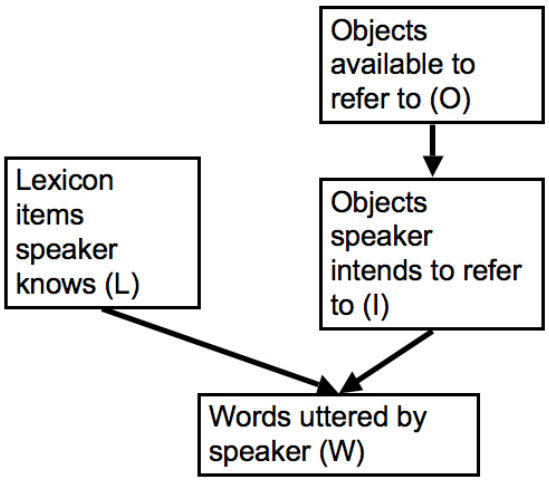

Figure 2. Generative process for producing words in a specific situation. The words uttered (W) depend on both the lexicon (L) and the intended objects (I). The intended objects (I) depend on what objects are current present (O).

The model is tested using a corpus derived from videos of parent-child interactions, where each utterance was transcribed and annotated with the small number of objects that were visible during that utterance. Given the words uttered by a speaker (*W*) in the presence of a set of objects (*O*), the model simultaneously infers the most probable lexicon for the speaker (*L*) and which objects in *O* the speaker intended to refer to (*I*). Although each (*W,O*) pair can be highly ambiguous, pooling the data across many observable pairs allows the model to disambiguate the word-meaning mappings, just as humans were able to do in the cross-situational word learning experiments of Yu & Smith (2007) and Smith & Yu (2008). The model far out-performed other statistical learning methods such as conditional probability and mutual information, identifying the most accurate set of lexicon items and speaker-intended objects.

In addition to its overall high accuracy, the Bayesian model reproduced several known word-learning behaviors observed in humans. For example, the model exhibited a

mutual exclusivity preference (Markman & Wachtel 1988, Markman 1989, Markman, Wasow, & Hansen 2003) because having a one-to-one mapping between a lexicon item and an object referent maximized the probability of a speaker using that lexicon item to refer to that object. The Bayesian model can also reproduce a behavior that children show called *one-trial learning* (Carey 1978, Markson & Bloom 1997), where it only takes one exposure to a word to learn its meaning. This occurs when the learner's prior knowledge and the current available referents in the situation make one word-meaning mapping much more likely than others. For example, suppose there are two objects in the current situation, a bird and an unknown object. Suppose the word *dax* is used. If the child has prior knowledge of the word *bird* and what it tends to refer to, then the model will view the lexicon item *dax* as most likely referring to the unknown object after only this one usage.

A third child behavior this model can capture is the use of words for individuating objects (Xu 2002). Xu (2002) found that when infants hear two different labels, they expect two different objects and are surprised if only one object is present; when only one label is used, they expect only one object to be present. That is, infants have an expectation that words are used referentially. This behavior falls out naturally in the Bayesian model because the model has a role for speaker intentions. Specifically, the models used its assumptions about how words work (they are often used referentially) to make inferences about the states of the world that caused a speaker to produce particular utterances (i.e., one label indicates one object, and two labels indicate two objects). In this way, the model replicated the infant behavior results from Xu (2002).

In a similar fashion, this model can directly incorporate speaker intention to explain behavioral results such as those of Baldwin (1993). Baldwin found that children could learn the appropriate label for an object even if a large amount of time elapsed between the label and the presentation of the object as long as the speaker's intention to refer to the object with that label was clear. In the Bayesian model, this information can be directly incorporated at the level of speaker intentions.

3.4. Syntax-semantics mapping

The meaning of a word is not always directly connected to a referent in the world, however. Some words are *anaphoric* – that is, they refer to something previously mentioned. Here is an example of *one* being used anaphorically in English:

(1) *"Look! A black cat. Oh, look - there's another one!"*

In this situation, most adults would expect to see a second black cat. That is, the default interpretation of *one* is a cat that has the property black and this utterance would sound somewhat strange (without additional context) if the speaker actually was referring to a grey cat. This interpretation occurs because most adults assume the *linguistic antecedent* of *one* is the phrase *black cat* (i.e., *one* could be replaced by *black cat* without the meaning of the utterance changing: *Look! A black cat. Oh, look – there's another black cat.*) Lidz, Waxman, & Freedman (2003) ran a series of experiments with 18-month-old children to test their interpretations of *one* in utterances like (1), and found that they too shared this intuition. So, Lidz, Waxman, & Freedman (2003) concluded that this knowledge about how to interpret *one* must be known by 18 months.

The interpretation of *one* depends on what antecedents *one* can have. According to common linguistic theory, an anaphor and its antecedent must have the same syntactic category. But what category is that? A common representation of the syntactic structure for *black cat* is in (2), where N^0 refers to a basic noun like *cat* and N′ is a category that includes both basic nouns like *cat* and nouns containing modifiers like *black cat*.

(2) $\left[_{N'}\ \text{black}\ \left[_{N'}\ \left[_{N^0}\ \text{cat}\right]\right]\right]$

Since *one* can have the string *black cat* as its antecedent, and *black cat* is category N′, then *one* should also be category N′. If *one* were instead category N^0, it could never have *black cat* as its antecedent – it could only have *cat* as its antecedent, and we could not get the interpretation most adults do for (1). Note that the bracketing notation in (2) indicates that *cat* can be labeled as both syntactic category N′ ($\left[_{N'}\ \left[_{N^0}\ \text{cat}\right]\right]$) and syntactic category N^0 ($\left[_{N^0}\ \text{cat}\right]$). This is what allows us to have *cat* as *one*'s antecedent sometimes, as in *I don't want a black cat – I want a grey one.* In this utterance, *one* must refer to *cat*, rather than *black cat*, and this is possible because *cat* can also be category N′, just as *one* is.

Under this view, adults (and 18-month-olds) have apparently learned that *one* should be category N′, since they allow *black cat* to be its antecedent. This includes both syntactic and semantic interpretation knowledge. On the syntactic side, *one* is category N′ in utterances like (1). On the semantic side, when the potential antecedent of *one* contains a modifier (such as *black*), that modifier is relevant for determining the referent of *one*. This relates to syntax-semantics mapping: because the modifier is relevant for determining the referent, the larger of the two N′ options should be chosen as the intended antecedent (*black cat* instead of just *cat*).

How this knowledge is acquired by 18 months has long been debated. The problem is the following. Suppose the child encounters utterance (1) from above in the context where two black cats are present. Suppose this child is not sure which syntactic category *one* is in this case – N′ or N^0. If the child thinks the category is N^0, then the only possible antecedent string is *cat*, and the child should look for the referent to be a cat. Even though this is the wrong syntactic category for *one*, the observable referent will in fact be a cat. How will the child realize that this is the wrong category, since the observable referent (a cat that also happens to be black) is compatible with this hypothesis? The same problem arises even if there's no modifier in the antecedent: *Look! A cat. Oh, look – there's another one.* The hypothesis that *one* is category N^0 is compatible with the antecedent *cat*, which is compatible with the observable referent being a cat, as indeed it is.

These ambiguous data dominate children's input for anaphoric *one*, and only rarely do unambiguous data appear (approximately 0.25% according to a corpus analysis by Lidz, Waxman, & Freedman (2003) and never in the corpus analysis conducted by Pearl & Mis (2011)). This is unsurprising once we realize that unambiguous data require a specific coincidence of utterance and situation. For example, suppose there are two cats present, one black and one gray. An unambiguous utterance would be *Look! A black cat. Hmmm…there's not another one around, though.* In this utterance, *one* cannot refer to *cat*, since there is clearly another cat around. Instead, *one* must refer to *black cat*, which would allow the utterance to make sense – there's not another black cat around (the other

cat is gray). Because *black cat* includes both a modifier and a noun , it must be N′, so this data point is unambiguous for *one*'s syntactic category and semantic interpretation.

Because unambiguous data are so rare in children's input, knowledge of anaphoric *one* was traditionally considered unlearnable without innate, domain-specific biases on the hypothesis spaces of children. In particular, many nativists such as Baker (1978), Hornstein & Lightfoot (1981), and Crain (1991) proposed that children already knew that *one* was category N′, so the choice between N^0 and N′ would never occur, eliminating the syntactic component of the learning problem.

Regier & Gahl (2004) discovered that a learner using Bayesian inference can leverage useful information from ambiguous examples that include a modifier, like (1). Specifically, for examples like (1), the learner observes how often the referent of *one* is a cat that is black. If the referents keep being black cats, this is a suspicious coincidence if *one* referred to *cat*, and not to *black cat*. The learner capitalizes on this suspicious coincidence and soon determines that *one* takes *black cat* as its antecedent in these cases. Since the string *black cat* can only be an N′ string (see (2)), the learner can then infer that *one* is of category N′ as well. The only specific linguistic knowledge the learner requires is (1) the definition of the hypothesis space (hypothesis 1: *one* = N′ category, hypothesis 2: *one* = N^0 category), and (2) knowing to use these specific informative ambiguous data.

Pearl & Lidz (2009) later explored the consequences of a Bayesian learner that did not know this second piece of information, and instead attempted to learn from all potentially informative ambiguous data involving anaphoric *one* (such as *Look, a cat! Oh, look – another one*). Pearl & Lidz found that this "equal opportunity" learner made the wrong choice, inferring that *one* was category N^0 due to the suspicious syntactic coincidences available in the additional ambiguous data. Thus, the second piece of information is vital for success, and Pearl & Lidz speculated that it is linguistic-specific knowledge since it requires the child to ignore a specific kind of language data (note, however, that it could be derived using a domain-general strategy - see Pearl & Lidz (2009) for more detailed discussion of this point).

Foraker, Regier, Khetarpal, Perfors, & Tenenbaum (2009) investigated another strategy for learning the syntactic category of *one*, this time drawing only on syntactic information and ignoring information about what the intended referent was. In particular, a learner could notice that *one* is restricted to the same syntactic arguments (called *modifiers*) that words of category N′ are restricted to, rather than being able to have both modifiers and another syntactic argument (*complements*) that words of category N^0 can have. That is, *one*, like N′ words, can take only modifiers as arguments, while N^0 words can take both modifiers and complements as arguments. This restriction is a suspicious coincidence if *one* is really category N^0. So, a Bayesian learner can infer that *one* is category N′. Notably, however, the ability to distinguish between modifiers and complements requires the child to make a complex conceptual distinction (see Foraker et al. (2009) for more discussion on this point) as well as link that conceptual distinction to the syntactic distinction of complements and modifiers, and it is unclear if 18-month-old children would be able to do this.

Pearl & Mis (2011, submitted) considered expanding the learner's view of the relevant data to include what they call *indirect positive evidence*. Specifically, they note that *one* is not the only anaphoric element in English. Other pronouns also have this referential property, such as *it, her, him,* etc. Moreover, other pronouns share

distributional properties with *one*: *Look, at the black cat! I want it/her/him/one*. This might cause children to view data involving these other pronouns as informative for learning about *one*. Notably, the antecedents for these pronouns always include the modifiers – in the above utterance, the antecedent is *the black cat*, and so the referent will be the black cat in question. If a Bayesian learner is tracking whether the mentioned property (e.g., "black", as indicated by the modifier *black*) is important for picking out the intended referent, these additional pronoun data will cause that learner to assume that mentioned properties are indeed important for interpreting anaphoric elements. So, when the child encounters an ambiguous example with anaphoric *one* like *Look, a black cat! Oh, look – another one,* the child will assume the mentioned property *black* is important, and pick *one*'s antecedent to be *black cat* rather than *cat*. This then leads to the correct interpretation. Moreover, because *black cat* can only be category N′, this also leads to the correct syntactic category for *one* in this context. Interestingly, while the learner gets the correct interpretation in this context, and so matches the 18-month-old behavioral data from Lidz et al. (2003), the learner actually has the wrong hypothesis about *one*'s category (*one*=N^0) when no modifier is present (*Look, a cat! Oh, look – another one*.) However, this wrong hypothesis does not lead to the wrong interpretation in this utterance, and so could easily go undetected. (See Pearl & Mis (submitted) for discussion of examples where the wrong hypothesis about *one* has observable consequences.) This suggests that the 18-month-olds from Lidz et al. (2003) may not have the full range of adult intuitions either.

3.5. Syntactic structure

Children must also discover the rules that determine what order words appear in. For example, consider the formation of yes/no questions in English. If we start with a sentence like *The cat is purring,* the yes/no question equivalent of this sentence is *Is the cat purring?* But how does a child learn to form this yes/no question? One rule that would capture this behavior would be "Move the first auxiliary verb to the front", which would take the auxiliary verb *is* and move it to the front of the sentence. This rule is a *linear* rule, since it only refers to the linear order of words ("first auxiliary"). Another rule that would capture this behavior is "Move the main clause auxiliary verb to the front". This is a *structure-dependent* rule, since it refers to the structure of the sentence ("main clause").

(3) Example of yes/no question formation

(i) Sentence:

The cat is purring.

(ii) Linear Rule: Move the first auxiliary verb

Is the cat t_{is} *purring*

(iii) Structure-Dependent Rule: Move the main clause auxiliary verb

Is [$_S$ *the cat* t_{is} *purring*]

Both of these rules account for simple yes/no questions like the one above, but only the structure-dependent rule accounts for behavior of more complex yes/no questions, as in (4).

(4) Example of complex yes/no question formation

(i) Sentence:

The cat who is in the corner is purring.

(ii) Linear Rule: Move the first auxiliary verb

**Is the cat who* t_{is} *in the corner is purring*

(iii) Structure-Dependent Rule: Move the main clause auxiliary verb

Is [$_S$ *the cat* [$_S$ *who is in the corner*] t_{is} *purring*]

Children as young as three years old appear to know that structure-dependent rules are required for complex yes/no question formation in English (Crain & Nakayama, 1987), yet unambiguous examples like (4iii) that explicitly demonstrate this structure-dependence are rare in child-directed speech (Pullum & Scholz 2002, Legate & Yang 2002). Since the yes/no question data children usually see are compatible with both linear and structure-dependent rules, it seems surprising that children know the structure-dependent rule for complex yes/no questions at such an early age. A standard explanation is that children innately know that language rules are structure-dependent, so they never consider other kinds of analyses for their input, such as linear rules (e.g., Chomsky, 1971).

Perfors, Tenenbaum, & Regier (2006, 2011) investigated whether a Bayesian learner that considered both linear and structure-dependent analyses could correctly infer that structure-dependent analyses were preferable, given child-directed speech data. Children must have a structure-dependent analysis of the linguistic data before they can hypothesize structure-dependent rules, so inferring a structure-dependent representation is a foundation for later inferring structure-dependent rules. Perfors et al. proposed that while complex yes/no questions implicating structure-dependent analyses might be rare, other data in the input, taken together, might collectively implicate structure-dependent analyses for the language as a whole. This could indirectly implicate the correct complex yes/no question structure without the need to observe complex yes/no questions in the input.

The hypothesis space of the Bayesian learner included both a linear set of rules (a linear grammar) and a structure-dependent set of rules (a hierarchical grammar) to explain the observable child-directed speech data. That is, given data (*D*), the learner inferred which grammar (*G*) satisfied two criteria:

(1) the grammar best able to account for the observable data
(2) the simplest grammar, where a grammar with fewer and/or shorter rules can be thought of as simpler

The posterior probability $P(G|D)$, which Bayes' Theorem tells us is proportional to $P(D|G)*P(G)$, incorporates both criteria. The likelihood $P(D|G)$ rewards grammars that are best able to account for the observable data, while also rewarding simpler derivations using the available grammar rules. The prior $P(G)$ rewards simpler grammars.

For data, Perfors et al. used the child-directed sentences from the Adam corpus (Brown 1973) of the CHILDES database (MacWhinney 2000), and divided the sentences into six groups based on frequency. The most frequent sentences also tended to be simpler. Perfors et al. found that a hierarchical grammar was optimal for all the data sets

that included more complex sentence forms, i.e. those that included at least some sentences that occurred less frequently than 100 times. Thus, if the Bayesian learner is exposed to enough complex sentences, it can infer that structure-dependent rules for generating the observed data are better than linear rules, and can apply this knowledge to analyzing and proposing rules for complex yes/no questions, even if no complex yes/no questions have been encountered before. Interestingly, even the earliest data in the Adam corpus shows a diversity of linguistic forms, suggesting that young children's data may be varied enough for them to prefer structure-dependent analyses if they are approximating the Bayesian inference procedures used by Perfors, Tenenbaum, & Regier. An open question is whether children have the memory and processing capabilities to make these approximations.

Perfors and colleagues (Perfors, Tenenbaum, Gibson, & Regier 2010) also used Bayesian learners to investigate how recursion might be instantiated in grammars. Recursion occurs when a phrasal category can be expanded using rules that eventually include another instance of that category, as in (5), where an S can be expanded using an NP (5i) and an NP can be expanded using an S (5ii).

(5) Recursive rule example
(rule i) S → NP VP
(rule ii) NP → N complementizer S

Recursion has been argued to be a fundamental and possibly innate part of the language faculty (Chomsky 1957), as well as the one of the only parts of the language faculty specific to humans (Hauser, Chomsky, & Fitch 2002). Perfors et al. (2010) evaluated grammars with and without recursive rules to decide which was optimal for parsing child-directed speech data. Grammars with recursive rules allow infinite embedding (Depth 3+ in (6)), while grammars without recursive rules allow embedding only up to a certain depth, e.g., 2 clauses deep (Depth 0, 1, and 2 in (6)).

(6) Embedding
(a) Subject-embedding
[Depth 0] [$_{Subj}$ *The cat*] *is purring.*
[Depth 1] [$_{Subj}$ *The cat that* [$_{Subj}$ *the girl*] *petted*] *is purring.*
[Depth 2] [$_{Subj}$ *The cat that* [$_{Subj}$ *the girl that* [$_{Subj}$ *the boy*] *kissed*] *petted*] *is purring.*
[Depth 3+] [$_{Subj}$ *The cat that* [$_{Subj}$ *the girl that* [$_{Subj}$ *the boy that* [$_{Subj}$…] *kissed*] *petted*] *is purring.*

(b) Object-embedding
[Depth 0] *The cat chased* [$_{Obj}$ *the mouse*].
[Depth 1] *The cat chased* [$_{Obj}$ *the mouse that scared* [$_{Obj}$ *the dog*]].
[Depth 2] *The cat chased* [$_{Obj}$ *the mouse that scared* [$_{Obj}$ *the dog that barked at* [$_{Obj}$ *the mailman*]]].
[Depth 3+] *The cat chased* [$_{Obj}$ *the mouse that scared* [$_{Obj}$ *the dog that barked at* [$_{Obj}$ *the mailman that* [$_{Obj}$…]]]].

The Bayesian learner had the same preferences as the one in Perfors, Tenenbaum, & Regier (2006, 2011): it attempted to identify the grammar that best balanced simplicity and the ability to account for the observed data. Note that grammars with recursive rules predict sentences that will rarely or never occur, such as the sentences with embedding of Depth 3+ in (6), so these grammars will not fit the data as well as grammars with limited embedding. However, a recursive grammar is often simpler than one that needs to encode exactly a specific depth of embedding, so whether recursion is learned will depend on the trade-off between these two factors with respect to the observed data.

Perfors et al. could have assumed that the learner considered only two kinds of hypotheses: grammars where rules are recursive whenever possible, or grammars where no rules are recursive. Instead, they also allowed the learner to have separate recursive rule types for subject-NPs (as in 6a) as opposed to object-NPs (as in 6b), since embedding is more often observed and more easily comprehended when it is object embedding (compare Depth 2 in 6a to Depth 2 in 6b).

The Bayesian learner, when given child-directed speech data, inferred that the optimal grammar was one where the subject-NP rules allowed both recursive rules and depth-limited embedding while the object-NP rules were only recursive. This result is due to the fact that multiple embeddings are much more frequently observed in object-NPs than in subject-NPs. More broadly, it also suggests that a statistical learner may be able to discover when recursive rules are useful and when they aren't. A child would not necessarily need to innately know that recursion is required for representing object-NPs. Instead, if recursive rules are available in their hypothesis space, children would be able to infer from their input that recursion is required for some parts of the grammar.

3.6. General summary of studies

We have tried to review several studies that highlight the contribution of Bayesian inference to language acquisition, including studies in the domains of phonetics and phonology, word segmentation, word-meaning mapping, syntax-semantics mapping, and syntactic structure. Though Bayesian modeling is only one approach to understanding language acquisition, it provides a way to investigate questions about the utility of statistical information in the data and which acquisition problems statistical learning can deal with effectively. In addition, it can often provide a coherent account of observed human behavior by demonstrating what a learner using Bayesian inference would do with the available data.

4. Conclusion

In this chapter, we have discussed ways in which Bayesian modeling can be used to explore questions of interest to the language acquisition community. As Bayesian models assume humans can use statistical information in sophisticated ways, we also provided a historical overview of statistical learning within the field of language acquisition, including experimental studies that demonstrate human statistical learning ability. We then discussed the Bayesian modeling framework, including some of its benefits that may be particularly interesting to both developmental and theoretical linguists. Finally, we reviewed several computational studies that modeled acquisition of knowledge in different domains using Bayesian inference techniques. Statistical learning techniques such as Bayesian inference, when coupled with well-defined problems and

hypothesis spaces, can help us understand both the nature of the data available to children and how they are able to acquire complex linguistic generalizations so rapidly.

References:

Anderson, J. R. 1990. *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Aslin, R., J. Saffran, & E. Newport. 1998. Computation of Conditional Probability Statistics by 8-Month-Old Infants. *Psychological Science, 9(4)*, 321-324.

Baker, C. L .1978. *Introduction to generative-transformational syntax*. Englewood Cliffs, NJ: Prentice-Hall.

Baldwin, D. 1993. Early referential understanding: Infants' ability to recognize acts for what they are. *Developmental Psychology, 29*, 832–843.

Bickerton, D. 1984. The Language Bioprogram Hypothesis. *Behavioral and Brain Sciences, 7(2)*, 173-222.

Bishop, C. 2006. *Pattern Recognition and Machine Learning*. Springer:

Boersma, P. & B. Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, *32(1),* 45–86.

Bonatti, L., Peña, M., Nespor, M., & Mehler, J. 2005. Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech process. *Psychological Science, 16,* 451-459.

Bonawitz, E., Denison, S., Chen, A., Gopnik, A., & Griffiths, T.L. 2011. A Simple Sequential Algorithm for Approximating Bayesian Inference. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, Austin, TX: Cognitive Science Society.

Brent, M. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*, 71–105.

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

Carey, S. 1978. The child as word learner. In J. Bresnan, G. Miller, & M. Halle (Eds*.), Linguistic theory and psychological reality*. Cambridge, MA: MIT Press, 264-293.

Cascells, W., Schoenberger, A., & Grayboys T. 1978. Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 299(18),* 999-1001.

Chater, N. & Manning, C. 2006. Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences, 10,* 287-291.

Chater, N. & Oaksford, M. 1999. Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences, 3(2),* 57-65.

Chemla, E., T. Mintz, S. Bernal, and A. Christophe. 2009. Categorizing words using 'frequent frames': what cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science, 12(3)*, 396-406.

Chomsky, N. 1955. *The logical structure of linguistic theory*. MIT Humanities Library. Microfilm. Published in 1977 by Plenum.

Chomsky, N. 1957. *Syntactic structures*. The Hague: Mouton.

Chomsky, N. 1971. *Problems of Knowledge and Freedom*. Fontana, London.

Chomsky, N. 1973. Conditions on transformations. In S. Anderson & P. Kiparsky (Eds.), *A Festschrift for Morris Halle*, (pp. 237-286). New York: Holt, Rinehart and Winston.

Chomsky, N. 1981. *Rules and Representations.* New York: Columbia University Press.
Conway, C. M., & Christiansen, M. H. 2005. Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning Memory and Cognition, 31 (1)*, 24-3916.
Crain, S. 1991. Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, *14*, 597–612.
Crain, S. & Nakayama, M. 1987. Structure dependence in grammar formation. *Language, 24*,139–186.
Dietrich, C., D. Swingley, & J. Werker. 2007. Native language governs interpretation of salient speech sound differences at 18 months. *Proceedings of the National Academy of Sciences, 104(41)*, 16027-16031.
Dillon, B., Dunbar, E., & Idsardi, B. 2011. A single stage approach to learning phonological categories: Insights from Inuktitut. Manuscript, University of Massachusetts, Amherst and University of Maryland, College Park.
Dresher, E. & Kaye, J. 1990. A Computational Learning Model for Metrical Phonology. *Cognition, 34*, 137-195.
Duda, R., Hart, P., & Stork, D. 2000. *Pattern Classification*.Wiley-Interscience.
Elman, J. 1990. Finding structure in time. *Cognitive Science, 14,* 179-211.
Elman, J. 1993. Learning and development in neural networks: the importance of starting small. *Cognition, 48,* 71-99.
Elman, J., E. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett. 1996. *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press/Bradford Books.
Feldman, N. 2011. *Interactions between word and speech sound categorization in language acquisition.* Unpublished doctoral dissertation, Brown University.
Feldman, N., T. Griffiths, & J. Morgan. 2009a. The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review, 116,* 752-782.
Feldman, N., T. Griffiths, T., & J. Morgan. 2009b. Learning phonetic categories by learning a lexicon. *Proceedings of the 31st Annual Conference of the Cognitive Science Society.* http://cocosci.berkeley.edu/tom/papers/lexicon.pdf.
Feldman, N., Griffiths, T., Goldwater, S., & Morgan, J. Submitted. A role for the developing lexicon in phonetic category acquisition. Manuscript, University of Maryland, College Park, University of California, Berkeley, and University of Edinburgh.
Ferguson, T. 1973. A Bayesian analysis of some nonparametric problems. *Annals of Statistics, 1(2),* 209-230.
Fiser, J., & Aslin, R. N. 2002. Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, 99 (24)*, 15822-15826.
Fodor, J. 1983. *Modularity of Mind.* Cambridge, MA: MIT Press.
Foraker, S., Regier, T., Khetarpal, A., Perfors, A., & Tenenbaum, J. 2009. Indirect evidence and the poverty of the stimulus: The case of anaphoric *one*. *Cognitive Science*, *33*, 287–300.
Frank, M. C., Goldwater, S., Mansinghka, V., Griffiths, T., & Tenenbaum, J. 2007. Modeling human performance on statistical word segmentation tasks. *Proceedings of the 29th annual meeting of the Cognitive Science Society* (pp. 281–286), Austin, TX:

Cognitive Science Society.
Frank, M. C., S. Goldwater, T. Griffiths, & Tenenbaum, J. 2010. Modeling human performance in statistical word segmentation. *Cognition, 117, 107-125.*
Frank, M. C., S. Goodman & Tenenbaum, J. 2009. Using Speakers' Referential Intentions to Model Early Cross-Situational Word Learning. *Psychological Science, 20(5),* 578-585.
Gambell, T. & Yang, C. 2006. Word Segmentation: Quick but not dirt. Ms. Yale University.
Gelman, A., Carlin, J., Stern, H., & Rubin, D. 2003. *Bayesian Data Analysis.* Chapman & Hall.
Geman, S. & Geman, D. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.
Gibson, E. & Wexler, K. 1994. Triggers. *Linguistic Inquiry, 25,* 407-454.
Gilks, W.R., Richardson, S., & Spiegelhalter D.J., editors. 1996. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, Suffolk.
Gleitman, L. & E. Newport. 1995. Language: An Invitation to Cognitive Science. In L. Gleitman & M. Liberman (Eds.), *An Invitation to Cognitive Science: Vol 1: Language.* Cambridge, MA: MIT Press, 1-24.
Goldwater, S. 2006. Nonparametric Bayesian Models of Lexical Acquisition. Ph.D. Dissertation, Brown University.
Goldwater, S., Griffiths, T., & Johnson, M. 2006. Interpolating between types and tokens by estimating power law generators. *Neural Information Processing Systems, 18*.
Goldwater, S., Griffiths, T., & Johnson, M. 2009. A Bayesian Framework for Word Segmentation: Exploring the Effects of Context. *Cognition, 112(1)*, 21-54.
Gómez, R. 2002. Variability and detection of invariant structure. *Psychological Science*, *13,* 431-436.
Gómez, R. & Gerken, L. 1999. Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70,* 109-135.
Goodman, N. 1955. *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.
Goodsitt, J. V., Morgan, J.L., & Kuhl, P. K. 1993. Perceptual strategies in prelingual speech segmentation. *Journal of Child Language, 20*, 229-252.
Graf Estes, K., Evans, J., Alibali, M., & Saffran, J. 2007. Can Infants Map Meaning to Newly Segmented Words? *Psychological Science, 18(3)*, 254-260.
Griffiths, T., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. 2010. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences, 14,* 357-364.
Griffiths, T. & Tenenbaum, J. 2005. Structure and strength in causal induction. *Cognitive Psychology, 51,* 334-384.
Griffiths, T. & Yuille, A. 2006. A primer on probabilistic inference. *Trends in Cognitive Sciences* 10(7). Supplement to special issue on Probabilistic Models of Cognition.
Harris, Z. 1954. Distributional Structure. *Word, 10,* 146-162.
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,* 1569-1579.

Hauser, M. D., Newport, E. L., & Aslin, R. N. 2001. Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition, 78*, B53-B64.
Hayes, J., & Clark, H. 1970. Experiments in the segmentation of an artificial speech analog. In J. R. Hayes (Ed.), *Cognition and the development of language*. New York: Wiley. 221-234.
Hornstein, N., & Lightfoot, D. 1981. *Explanation in linguistics: The logical problem of language acquisition*. London: Longmans.
Huang, C.-T.J. 1982. Logical relations in Chinese and the theory of grammar. Doctoral dissertation. MIT, Cambridge, MA.
Johnson, M. 2008. Using adapter grammars to identify synergies in the unsupervised learning of linguistic structure. In *Proceedings of Association for Computational Linguistics 2008.*
Kemp, C., Perfors, A., & Tenenbaum, J. 2007. Learning overhypotheses with hieracrchical Bayesian models. *Developmental Science, 10(3),* 307-321.
Kemp, C. & Tenenbaum, J. 2008. The discovery of structural form. *Proceedings of the National Academy of Sciences, 105(31),* 10687-10692.
Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. 2002. Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition, 83*, B35-B42.
Lasnik, H. & Saito, M. 1984. On the nature of proper government. *Linguistic Inquiry, 15,* 235- 289.
Legate, J. & Yang, C. 2002. Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review, 19,* 151-162.
Legate & Yang. 2007. Morphosyntactic learning and the development of tense. *Language Acquisition, 14(3),* 315-344.
Legendre, G., Y. Miyata, and P. Smolensky. 1990. Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. *Technical Report 90-5*, Institute of Cognitive Science, Univ. of Colorado.
Lidz, J., Waxman, S., & Freedman, J. 2003. What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition*, *89*, B65–B73.
MacWhinney, B. 2000. T*he CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates, third edition.
Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. 1999. Rule learning by seven-month-old infants. *Science*, *283,* 77-80.
Markman, E.M. 1989. *Categorization and naming in children: Problems of induction*. Cambridge, MA: MIT Press.
Markman, E.M., & Wachtel, G.F. 1988. Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology, 20*, 121–157.
Markman, E.M., Wasow, J.L., & Hansen, M.B. 2003. Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology, 47*, 241–275.
Markson, L., & Bloom, P. 1997. Evidence against a dedicated system for word learning in children. *Nature*, *385*, 813–815.
Marr, D. 1982. *Vision*. San Francisco: W.H. Freeman.
Marthi, B., Pasula, H., Russell, S., & Peres, Y. 2002. Decayed MCMC Filtering. In *Proceedings of 18th UAI*, 319-326.

Maye, J. & Weiss, D. 2003. Statistical cues facilitate infants' discrimination of difficult phonetic contrasts. *Proceedings of the 27th Annual Boston University Conference on Language Development.*

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

Mehler, J., Peña, M., Nespor, M., & Bonatti, L. 2006 . The soul of language does not use statistics: reflections on vowels and consonants. *Cortex*, *42(6),* 846-854.

Medina, T., Snedecker, J., Trueswell, J., & Gleitman, L. 2011. How words can and cannot be learned by observation, *Proceedings of the National Academy of Sciences, 108,* 9014-9019.

McClelland, J. 1998. Connectionist models and Bayesian inference. In M. Oaksford & N. Chater (eds.), *Rational Models of Cognition.* Oxford: Oxford University Press. 21-53.

McClelland, J., Botvinick, M., Noelle, D., Plaut, D., Rogers, T., Seidenberg, M., & Smith, L. 2010. Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences, 14,* 348-356.

Mintz, T. 2002. Category induction from distributional cues in an artificial language. *Memory & Cognition*, *30*, 678-686.

Mintz, T. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90, 91-117.

Mintz, T. 2006. Finding the verbs: distributional cues to categories available to young learners. In K. Hirsh-Pasek & R. M. Golinkoff (Eds.), *Action Meets Word: How Children Learn Verbs* (pp. 31-63). New York: Oxford University Press.

Newport, E. (1990). Maturational constraints on language learning. *Cognitive Science, 14,* 11-28.

Newport, E., & Aslin, R. 2004. Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48 (2),* 127-162.

Niyogi, P. & Berwick, R. 1996. A language learning model for finite parameter spaces. *Cognition, 61,* 161-193.

Pearl, L. 2011. When unbiased probabilistic learning is not enough: Acquiring a parametric system of metrical phonology. *Language Acquisition, 18(2),* 87-120.

Pearl, L., Goldwater, S., & Steyvers, M. 2011. Online Learning Mechanisms for Bayesian Models of Word Segmentation, *Research on Language and Computation, 8(2)*, 107-132.

Pearl, L. & Lidz, J. 2009. When domain general learning fails and when it succeeds: Identifying the contribution of domain specificity, *Language Learning and Development, 5(4),* 235-265.

Pearl, L. & Mis, B. 2011. How Far Can Indirect Evidence Take Us? Anaphoric One Revisited. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.* Boston, MA: Cognitive Science Society.

Pearl, L. & Mis, B. Submitted. What Indirect Evidence Can Tell Us About Universal Grammar: Anaphoric One Revisited. University of California, Irvine. http://www.socsci.uci.edu/~lpearl/papers/PearlMis2011Manu_AnaOne.pdf

Pelucchi, B., Hay, J., & Saffran, J. 2009a. Statistical Learning in Natural Language by 8-Month-Old Infants. *Child Development, 80(3)*, 674-685.

Pelucchi, B., Hay, J., & Saffran, J. 2009b. Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition,* 244-247.

Perfors, A., Tenenbaum, J.B., Gibson, E., Regier, T. 2010. How recursive is language? A Bayesian exploration. In H. van der Hulst (ed.) *Recursion and human language.* Mouton: DeGruyter: 159-175.

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

Perfors, A., Tenenbaum, J., Regier, T. 2011. The learnability of abstract syntactic principles. *Cognition 118*(3): 306-338.

Perruchet, P. & Desaulty, S. 2008. A role for backward transitional probabilities in word segmentation? *Memory & Cognition, 36(7),* 1299-1305

Peters, A. 1983. *The Units of Language Acquisition, Monographs in Applied Psycholinguistics*. New York: Cambridge University Press.

Pinker, S. 1984. *Language learnability and language development*. Cambridge, MA: MIT Press.

Prince, A. & Smolensky, P. 2004. *Optimality Theory: Constraint interaction in generative grammar.* Blackwell.

Prince, A. & Tesar, B. 2004. Learning phonotactic distributions. *Constraints in phonological acquisition,* 245-291.

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

Redington, M., Chater, C., & Finch, S. 1998. Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science, 22(4),* 425-469.

Regier, T., & Gahl, S. 2004. Learning the unlearnable: The role of missing evidence. *Cognition*, *93*, 147–155.

Resnik, P. & Hardisty, E. 2009. Gibbs sampling for the uninitiated. Unpublished manuscript, Version 0.3, October 2009. Available from http://www.umiacs.umd.edu/~resnik/pubs/gibbs.pdf.

Rumelhart, D. & McClelland, J. editors. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, chapter 18. MIT Press.

Saffran, J., Aslin, R., & Newport, E. 1996. Statistical Learning by 8-Month-Old Infants. *Science, 274,* 1926-1928.

Saffran, J. R., Hauser, M., Seibel, R. L., Kapfhamer, J., Tsao, F., & Cushman, F. 2008. Grammatical pattern learning by infants and cotton-top tamarin monkeys. *Cognition, 107*, 479-500.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. 1999. Statistical learning of tone sequences by human infants and adults. *Cognition, 70(1)*, 27-52.

Sakas, W. This volume. Parameter Setting. In J. Lidz, W. Snyder, & C. Pater (eds), *The Oxford Handbook of Developmental Linguistics.*

Sampson, G. 2005. The 'Language Instinct' Debate (Revised Edition). Continuum: London.

Shi, L., T. Griffiths, N. Feldman, & A. Sanborn. 2010. Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review, 17(4),* 443-464.

Smith, L., & Yu, C. 2008. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*, 1558–1568.

Smolensky, P., G. Legendre, & Y. Miyata. 1992. Principles for an integrated connectionist/symbolic theory of higher cognition*, Report No. CU-CS-600-92*. Computer Science Department, University of Colorado at Boulder.
Smolensky, P. & Legendre, G. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar.* MIT Press: Cambridge, MA.
Swingley, D. 2005. Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology, 50*, 86-132.
Swingley, D. 2009. Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B, 364,* 3617-3632.
Tenenbaum, J. 1996. Learning the structure of similarity. In D. Touretzky, M. Mozer, & M. Hasslemo (eds.), *Neural Information Processing Systems, 8*. Cambridge: MIT Press.
Tenenbaum, J. & Griffiths. T. 2001. Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629-641.
Tesar, B. & Smolensky, P. 2000. *Learnability in Optimality Theory.* MIT Press: Cambridge, MA.
Thiessen, E. 2007. The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language, 56*, 16-34.
Thiessen, E., & J. Saffran. 2003. When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39*, 706-716.
Thompson, S. & Newport, E. 2007. Statistical Learning of Syntax: The Role of Transitional Probability. *Language Learning and Development, 3,* 1-42.
Toro, J. M., & Trobalon, J. B. 2005. Statistical computations over a speech stream in a rodent. *Perception and Psychophysics, 67 (5),* 867-875.
Vallabha, G., McClelland, J., Pons, F., Werker, J., & Amano, S. 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the U.S., 104(33)*, 13273-13278.
Wang, H. & Mintz, T. 2008. A Dynamic Learning Model for Categorizing Words Using Frames. *BUCLD 32 Proceedings*, Chan, H., Jacob, H., & Kapia, E. (eds.), Somerville, MA: Cascadilla Press, 525-536.
Waxman, S. R. 1990. Linguistic biases and the establishment of conceptual hierarchies: Evidence from preschool children. *Cognitive Development, 5,* 123–150.
Wexler, K. & Culicover, P. 1980. *Formal Principles of Language Acquisition.* Cambridge, MA: MIT Press.
Wolff, J. G. 1977. The discovery of segments in natural language. *British Journal of Psychology, 68*, 97-106.
Xu, F. 2002. The role of language in acquiring object concepts in infancy. *Cognition, 85*, 223–250.
Xu, F., & Tenenbaum, J. 2007. Word learning as Bayesian inference. *Psychological Review, 114*, 245-272.
Yang, C. 2002. *Knowledge and Learning in Natural Language.* Oxford: Oxford University Press.
Yang, C. 2004. Universal Grammar, statistics, or both? *Trends in Cognitive Sciences, 8(10),* 451-456.
Yu, C. & Smith, L. 2007. Rapid Word Learning Under Uncertainty via Cross-Situational Statistics. *Psychological Science, 18(5)*, 414-420.