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

Lisa Pearl & Sharon Goldwater

1. Statistical learning: Experimental evidence

Statistical learning has long been recognized as being part of the acquisition process (Chomsky 1955, Hayes & Clark 1970, Wolff 1977, Pinker 1984, Goodsitt, Morgan, & Kuhl 1993, among others), but traditionally it was viewed as playing a secondary role, rather than a primary one. Since young learners were generally perceived as poor learners, experience-independent innate knowledge was believed to be the driving force behind children's successful acquisition (Chomsky 1981, Fodor 1983, Bickerton 1984, Gleitman and Newport (1995), among others). Simply put, children were not believed to be capable of tracking statistical information in language input to the extent that they would need to for learning linguistic knowledge.

Saffran, Aslin, & Newport (1996) was a groundbreaking study in this respect, because it intended to demonstrate that young children have "powerful mechanisms for the computation of statistical properties of language input". Saffran et al. investigated the task of word segmentation, where the child must identify words in a stream of fluent speech. They 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. Aslin, Saffran, & Newport (1998) later affirmed that 8-month-old infants were tracking the syllable transitional probability to identify word boundaries. 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").[1]

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). However, Pelucchi, Hay, & Saffran (2009a) later found that 8-month-olds prefer Italian syllable sequences with a high transitional probability over syllable sequences with a low transitional probability, though it was unclear if infants regarded a low transitional probability sequence as a word. The precise way in which infants might use transitional probability information for realistic language data remains an open question.

Notably, however, the broader claim of Saffran et al. (1996) was not tied to transitional probability. Instead, they proposed that some aspects of acquisition may be

[1] This is more standardly known in statistics as the conditional probability of Y given X.

"best characterized as resulting from innately biased statistical learning mechanisms rather than innate knowledge". More specifically, it could be that humans are innately equipped with sophisticated statistical learning abilities which obviate the need for sophisticated prior knowledge about language in some cases. The ability to track transitional probabilities between temporally ordered sound sequences (here, syllables) is one innately biased statistical learning mechanism, but it need not be the only one. Further research has investigated a number of questions raised by these initial studies, most notably 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 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. However, 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, and that the statistics they use cannot always be characterized as something like 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, something traditionally considered a major challenge for early word learning (e.g., Markman1989, Waxman 1990) because these words overlap in the referents they apply to (a tabby is a cat, which is an animal). One sophisticated statistical inference that can help with this problem is related to what Xu & Tenenbaum call a *suspicious coinicidence*, 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 tabby cat. The learner at this point might (implicitly) have two hypotheses (*blick = cat, blick = tabby*), and expectations associated with these two hypotheses. Specifically, if *blick = cat*, other kinds of cats besides tabby cats should be labeled *blicks.* That is, the set of *blicks* is larger than just the set of tabby cats, and so other cats should also be labeled *blicks* sometimes. This, however, did not happen in the example situation above – three *blicks* were labeled, and all of them were tabby cats. This is, according to Xu & Tenenbaum, a suspicious coincidence if *blick* really means *cat*. Instead, it is more likely that *blick* is a subordinate label that is more specific, in this case *tabby*. Xu & Tenenbaum (2007) discovered that both adults and children between the ages of 3 and 5 are able to notice suspicious coincidences like this, and use them 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 and 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 are compatible with the idea that humans' statistical learning abilities are highly domain-general, showing robustness across modalities and presentation formats – particularly the results with the tactile modality, which is not used in natural languages.

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, 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 computational modeling. In the remainder of this chapter, we focus on the contributions of computational studies, discussing the kinds of questions they can answer and mentioning briefly some of the different computational approaches that have been used to answer these questions. We then focus in more detail on the Bayesian approach to computational modeling and provide some in-depth examples of work on language acquisition using this approach.

2. Computational models of statistical learning

There has been a great deal of work on computational modeling of language acquisition over the last three decades, and researchers have taken a number of different approaches. However, nearly all of these approaches have sought to answer one or more of the following questions:

1. What sources of information are available in the language input to children, and which of these might be useful in extracting linguistically meaningful generalizations?
2. What kinds of generalizations are learnable in principle (as opposed to being necessarily innate) and/or what innate knowledge is necessary?
3. What kind of mental process or neural architecture might be available and sufficient to extract these generalizations?

Different approaches have tended to focus more strongly on one or another of these questions, often due to particular theoretical views. For example, the connectionist approach is committed to the idea that language acquisition is entirely supported by domain-general learning abilities (Rumelhart and McClelland 1986, Elman et al. 1996). As a result, much of the connectionist literature has been devoted to showing that particular kinds of linguistic generalizations are learnable without the need for specific innate linguistic knowledge, and connectionist researchers do this by implementing models that learn those generalizations (e.g., Elman 1990, Elman 1993, Christiansen & Chater 1994, Oshma-Takane, Takane & Schultz 1999). Connectionist modeling research also focuses on mental representations and architecture, based on the belief that distributed representations and neural network architectures are critical to the success of the domain-general learner.

The connectionist approach to language acquisition is appealing to many researchers who do not believe in innate linguistic knowledge, but is typically rejected by those who do. Some of these researchers also reject the entire idea that linguistic generalizations may be acquired through statistical learning, but others have chosen to work with models that combine strong linguistic constraints with statistical learning. For example, a number of researchers have proposed models of acquisition based on variants of Optimality Theory. These models assume that constraint rankings are located along a numerical scale (Boersma 1997, Boersma & Hayes 2001), or that constraints themselves are weighted numerically (Goldwater & Johnson 2003, Hayes & Wilson 2008), and that evidence from the learner's input data is used to change the constraint rankings or weights. Although there are strong historical and formal connections between Optimality Theory and connectionism[2], they differ strongly in their view of innateness. Whereas

[2] Harmonic Grammar (Legendre et al. 1990, Smolensky et al. 1992), a precursor to Optimality Theory that has recently been regaining popularity among some linguists (Pater 2009, Potts et al. in press), was actually conceived of as a hybrid connectionist-symbolic model with a neural network architecture but strong linguistic constraints. This shows that in fact there is nothing inherent about connectionist models that forces a domain-general approach to learning.

most connectionist approaches assume language acquisition results from domain-general learning mechanisms, OT-based theories are rooted in the idea that the learner comes to the task with a set of innate universal linguistic constraints. Thus, models of learning in OT tend to focus on the interaction between the mental processes of learning and the universal constraints that are needed.

Other computational approaches not specifically tied to the two just mentioned have focused on identifying the useful statistical information available in the data. These include work by Redington, Chater, & Finch (1998) which examines the usefulness of the surrounding context of a word for grammatical categorization, studies by Mintz and collagues (Mintz 2003, Wang & Mintz 2008, Chemla et al. 2009) on the usefulness of frequent frames for grammatical categorization, and work by Albright & Hayes (2002) that investigates the morphological generalizations that can be posited based on comparing sets of word pairs.

Due to space constraints, it is impossible for us to provide a thorough review of computational work in all of these different areas. Rather than giving a cursory overview of several different modeling approaches, we focus here on a single one, Bayesian modeling. We do so for several reasons. First, Bayesian modeling is a relatively new approach to language acquisition, so there are few other resources available to those interested in learning more about it. Second, the Bayesian approach offers a concrete way to examine what knowledge is required for acquisition, and whether that required knowledge is domain-specific or domain-general, without committing to either view *a priori* . Finally, the Bayesian approach has led to the investigation of a new set of questions that previous approaches have not considered; specifically, whether human language learners can be viewed as being optimal statistical learners (i.e., making optimal use of the statistical information in the data), and in what situations. Whereas previous approaches (e.g., connectionist) have typically focused on *how* learners process their input to form generalizations, a Bayesian model can potentially address the question of *why* they make the generalizations they do, i.e., because these generalizations are statistically optimal given the available data and any learning biases, innate or otherwise. This view assumes that the learner is in some sense adapted to the task at hand – an assumption underlying the so-called *rational analysis* view of cognition (Anderson 1990, Chater & Oaksford 1999).

Some readers may not be comfortable with this idea, despite its success in modeling human behaviors in other areas of cognition such as numerical cognition (Tenenbaum 1996), causal induction (Griffiths & Tenenbaum 2005), and categorization (Kemp, Perfors, & Tenenbaum 2007). Nevertheless, recent work has begun to provide evidence that, in language acquisition as in other areas, humans do exhibit optimal behavior, at least in some circumstances (Feldman et al. 2009a, Xu & Tenenbaum 2007). It has also been argued that knowing what optimal behavior would be in a given situation is helpful even if humans do not exhibit this behavior, because we can then begin to investigate how and why humans might differ from it (Goldwater et al. 2009, Frank et al. in submission).

Put another way, the Bayesian approach to computational modeling investigates the problem of language acquisition at the Marr's (1984) *computational level*, seeking a declarative (rather than procedural) model of the learner. The learner's behavior is viewed as optimizing some set of goals, which are described mathematically using

Bayesian probability theory (see below). This contrasts with *algorithmic-level* approaches to understanding the information processing task facing the language learner – they hypothesize specific procedures that can be applied to the input to produce linguistically meaningful output: For example, learners might segment words by identifying syllable sequences with high frequency and mutual information (Swingley, 2005), define a grammatical category by the group of words clustered together by a frequent frame (Mintz 2003, Wang & Mintz 2008, Chemla et al. 2009), or use back-propagation to change the set of weights in a neural network (Elman 1990, Elman 1993).

Another feature of the Bayesian approach that sets it apart from most other computational modeling approaches is its focus on making the space of hypotheses considered by the language learner explicit, and encoding the learner's biases by assigning an explicit probability distribution over these hypotheses. This contrasts with neural network models, which have only implicit hypothesis spaces (those functions from inputs to outputs that are possible to learn given the structure of the network) and biases (functions that are easier or harder to learn), and no probability distribution over hypotheses. The hypothesis space in an OT learner is more explicit (all possible rankings of the constraints), but again there is no probability distribution assumed. Note that the Bayesian approach itself is agnostic as to whether the hypothesis space is governed by domain-general or domain-specific cognitive constraints, leaving this as an empirical question. This makes the approach appealing both to researchers who are interested in whether domain-specific constraints are necessary, and increasingly to those who are already committed to this position, but wish to investigate specific linguistic constraints. In addition, Bayesian models can operate over the kinds of highly structured representations that many linguists believe are correct (e.g., Regier & Gahl 2004, Perfors, Tenenbaum , & Regier 2006, Foraker et al. 2009, Pearl & Lidz 2009, Perfors et al. to appear).

To formalize the preceding discussion, 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 data. Note that the hypothesis space could be discrete (e.g., a finite or infinite set of 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 identify how probable each possible hypothesis h is, i.e. to estimate $P(h|d)$, the *posterior distribution* over hypotheses. Bayes' Rule states that the posterior can be reformulated as in (1):

(1) Bayes' Rule

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

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 normalizing factor that ensures that $P(h|d)$ is a proper probability distribution, summing to 1 over all values of h. In any particular situation, $P(d)$ is constant, so the denominator can often be safely ignored when comparing the relative probability of one hypothesis to another. Thus, 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 should the learner's input affect the learner's beliefs about which hypothesis is correct?

A simple example, adapted from Griffiths and 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 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 .0000531. If $\theta = 0.5$, then the same sequence has probability .000978. To determine which hypothesis is more plausible given that particular sequence, we can compute the *posterior odds ratio* as in (2):

(2) Posterior Odds Ratio

$$\frac{P(h_1 \mid d)}{P(h_0 \mid d)} = \frac{\frac{P(d \mid h_1)P(h_1)}{P(d)}}{\frac{P(d \mid h_0)P(h_0)}{P(d)}} = \frac{P(d \mid h_1)P(h_1)}{P(d \mid h_0)P(h_0)} = \frac{(.0000531)(1/3)}{(.000978)(2/3)} \approx \frac{1}{37}$$

That is, 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 and 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.[3] 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. $P(h)$ 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 the discussion of Perfors, Tenenbaum, & Regier

[3] 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

(2006) and Perfors et al. (to appear) in section 3 below for explicit examples of these ideas as applied to language acquisition.

The hypothesis space of a Bayesian model not only can be very complex and structured, but also may contain multiple levels of linguistic representation. 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. In a study we describe further in Section 3, Feldman et al. (2009) compared two Bayesian models of phonetic category acquisition to demonstrate that simultaneously learning phonetic categories and the lexical items containing those categories led to more successful categorization than learning phonetic categories alone. These types of joint learning models are 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.

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 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 (perhaps a very heuristic one) that is able to approximately identify the posterior distribution over hypotheses.

For example, most Bayesian models of language acquisition have used algorithms based on Markov chain Monte Carlo methods such as Gibbs sampling to obtain samples from the posterior distribution (Gilks et al., 1996; Geman and Geman, 1984; also see Resnik and Hardisty (2009) for an accessible tutorial and Knight (2009) for a humorous introduction). 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. However, it should be noted that a more recent line of work has begun to address the question of how learners might implement Bayesian predictions in a more cognitively plausible way. For example, Shi, Griffiths, Feldman, & Sanborn (to

appear) discuss how exemplar models may provide a possible mechanism for implementing Bayesian inference, since these models allow an approximation process called importance sampling. Another example is the work of Pearl, Goldwater, and Steyvers (2010) on word segmentation, which we will discuss further in Section 3.

One main contribution of Bayesian models is that they provide a way to formally evaluate claims about children's hypothesis space. For example, they can indicate if certain constraints or restrictions are required in order to learn some aspect of linguistic knowledge. Section 3 discusses several studies that investigate the prior knowledge children would need to learn the required linguistic information from the available child-directed speech data (e.g., Regier & Gahl 2004, Perfors, Tenenbaum, & Regier 2006, Foraker et al. 2009, Pearl & Lidz 2009, Perfors et al. to appear). In many cases, these models allow researchers to determine if the child's hypothesis space needs to be restricted in a specific way, or if it is possible to converge on the correct hypothesis within a larger, less restricted hypothesis space.

In this way, Bayesian models allow us to investigate if a particular hypothesis space is viable for language acquisition. More specifically, if a Bayesian learner looking for the optimal hypothesis given the data cannot converge on the correct hypothesis, this suggests that the current conception of the hypothesis space cannot be correct. Instead, some additional knowledge is required to successfully navigate the potential hypotheses and converge on the correct one. This additional knowledge may take the form of an additional constraint on the hypothesis space that gives preference to certain hypotheses over others, or eliminates some hypotheses entirely. On the other hand, if a Bayesian learner can converge on the correct hypothesis given the data, this suggests the hypothesis space is viable for children capable of approximating sophisticated statistical inference.

We should note that a Bayesian model is a tool that can be used to evaluate hypotheses in a predefined hypothesis space, but it is not a tool that creates a hypothesis space. If a hypothesis space is not already available, a Bayesian model cannot help. This property likely makes Bayesian modeling more appealing to linguists interested in learning specific abstract or structured representations of language, since the acquisition problem in these cases is often framed as choosing from a set of already existing hypotheses.

That being said, the predefined hypothesis space can be very broad. 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). As a simple example taken from Goodman (1955) and presented in Kemp, Perfors, & Tenenbaum, suppose people are learning the color distribution of marbles in a bag, where marbles can be either black or white. During training, people open bags and find out that the marbles in each bag are either all black or all white. When opening a new bag, they are allowed to draw one marble only before inferring the color distribution. A hypothesis about color distribution for this new bag might be that all the marbles are either all black or all white (which allows someone to make a strong inference after observing only a single marble from the new bag). An overhypothesis for color distribution is that all bags contain marbles that are uniform in their color distribution. Thus, during training, people learn to give this overhypothesis high probability as more and more bags are observed that are

uniform in color (as opposed to an overhypothesis that allows mixed colors in a bag). 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 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. Overhypotheses can be explicitly instantiated in hierarchical Bayesian models, such as the ones discussed in Kemp, Perfors, & Tenenbaum (2007) and Kemp & Tenenbaum (2008).

In the realm of syntactic acquisition, overhypotheses may correspond to what kinds of grammars are likely to be useful for analyzing the observable child-directed speech data (see the discussion of research by Perfors and colleagues in section 3 below). In addition, overhypotheses may also naturally correspond to linguistic parameters in generative linguistics, where a parameter is an abstract structural property that connects to many observable linguistic structures (Chomsky 1981) and thus constrains predictions on what structures should be observed in the language.

3. Specific example studies

We now turn to a survey of some specific representative studies in different areas of language acquisition, each one demonstrating how sophisticated probabilistic inference can be applied to a relevant problem within the chosen linguistic domain. For each study, we will describe the specific problem to be solved, discuss the hypothesis space of choices for each problem, and describe how Bayesian inference can operate over this hypothesis space to yield the same answers that humans seem to find. Of course, we cannot include all relevant studies here, but we hope to present illustrative samples of the application of Bayesian inference to problems in language acquisition.

Phonetics and perceptual learning

Feldman, Griffiths, & Morgan (2009b) 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 means of different vowel categories are different, there is significant overlap in the distributions, e.g. a particular token of /e/ may sound exactly like a token of /ɛ/, even if spoken by the same individual. See figure 1 below for an illustration of this variation in men's vowels sounds, taken from Feldman et al..

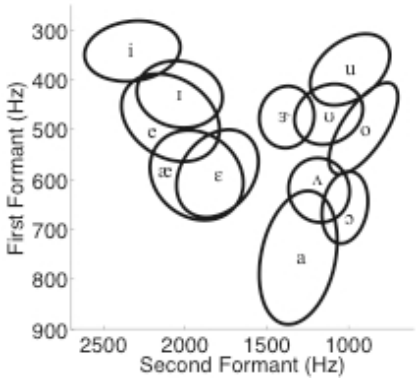

Figure 1. Example distribution of men's vowel sounds. Many vowel sounds have overlapping distributions, such as /e/ and / ɛ/.

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 also take into account the words those sounds occur in. Of course, young infants who are still learning the phonology of their language have very little knowledge of the lexicon. Feldman et al. present evidence from experimental studies suggesting that phonetic categorization and word segmentation and learning are acquired 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 (lexical items), or that words are acquired first and then used to disambiguate phonetic categories, Feldman et al. propose a joint model of learning in which phonetic categories and words 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 score 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 tokens rather than phonetic tokens (i.e., that the child is able to segment at least some words). The learner now has two

goals: to find phonetic categories (as in the distributional learner) but also to categorize word tokens into lexical items, 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 words are considered to be the same lexical item. On the other hand, if two word tokens are assigned to the same lexical item, then their phones should belong to the same categories. The hypothesis space for this model consists of pairs of categorizations (of phones into phonetic categories, and words into lexical items). Since the lexical learning task can also be viewed as categorization, it is modeled using another Dirichlet process, which again prefers lexicons containing fewer items when possible.

Using a toy data set, Feldman et al. show that the lexical-distributional model makes an interesting and counterintuitive prediction about minimal pairs. Specifically, if a pair of phones (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 phones. This is because the lexical-distributional learner can use phones 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 (phones and words) can actually make the learning problem easier than learning in sequence.

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!" 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).[4] 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 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 undersegments the input data -- the word boundaries it finds tend to be very accurate, but it often 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*, showing what kinds of solutions an idealized learner capable of extracting the necessary statistical cues would achieve given the available input and certain assumptions about the capabilities of the learner (i.e., whether the learner can track word-to-word dependencies or not). However, it does not tell us whether humans actually behave in ways consistent with the ideal learner, 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.

---

[4] 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.

In the work of Frank et al. (in submission), 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.[5] 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 words in each utterance (1-24 words), the total number of words/utterances heard in the training phase (48-1200 words), or the number of words in the vocabulary (3-9 words).

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. This can be explained by the fact that longer utterances have more possible segmentations, so there is a larger hypothesis space for the model to 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. This can be seen as a competition effect.

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.

In the experiment that manipulated the number of words in the vocabulary, subjects found languages with larger vocabularies more difficult to segment than those with smaller vocabularies. Although this finding was not surprising, all of the models tested predicted exactly the opposite result. This is because larger vocabularies require more memory to store, but they also make 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-

[5] 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.

word transitions. Frank et al. point out that the models under consideration have perfect memory, so the statistical properties of larger vocabularies make the task easier. 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 all of these are able to correctly model the data from all three experiments. Similar kinds of modifications were also explored by Pearl, Goldwater, & Steyvers (2010) in the context of word segmentation from naturalistic corpus data. The question of interest was 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. (2010) process 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).

The results of these simulations suggested that constrained learners could be nearly as successful at segmentation as the ideal learner in most cases, despite their processing and memory limitations. This suggests that children may not require an infeasible amount of processing power to identify words using this kind of sophisticated statistical inference. Moreover, Pearl et al. found that constrained learners did not always benefit from the bigram assumption which was helpful to the ideal learner. This may be because those constrained learners lacked sufficient processing resources to effectively exploit that information. So, useful information for an ideal learner may not necessarily be as useful for a constrained learner.

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 processing power (like the ideal learner has) is always better. However, these results are compatible with an idea for language acquisition called the "Less is More" hypothesis (Newport 1990), which suggests that less processing power may actually be beneficial for language acquisition. 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.

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). Frank et al. explore the utility of Bayesian inference for word-meaning mappings, incorporating the idea that speaker intention and the objects present in the world at the time of the linguistic utterance influence what words people choose to say. More specifically, the process of word-meaning mapping is part of a larger process that starts with a speaker's intention to refer to particular objects that are present, incorporates the speaker's knowledge of lexicon items for her language, and ends with the speaker choosing specific lexicon items that refer to specific objects. This process can be represented schematically as in Figure 2, where the words uttered by the speaker (W) in a given situation depend both on the lexicon items the speaker in general knows (L) and the referents presently available that the speaker intends to refer to (I). The intended referents then depend on the set of objects (O) presently available, with the intended referents presumably being some subset of the set of objects available.

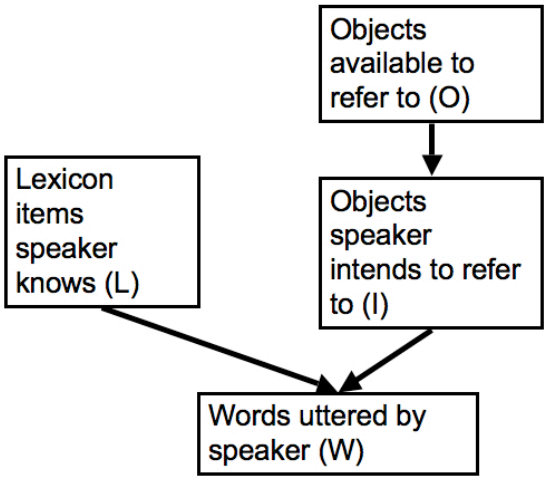

Figure 2. Generative process for producing words in a specific situation. The words uttered (W) depend on both the lexicon (L) and the intended objects (I). The intended objects (I) depend on what objects are current present (O).

Given realistic child-directed speech in situations with a number of objects present for a speaker to refer to, this Bayesian 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. Specifically, 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 items for the speaker (L) and which objects in the specific situation that speaker intended those lexicon items to refer to (I). It does this by pooling the data over many observable situations in which a speaker intended to refer to objects that were present.

Moreover, this Bayesian model is able to produce several known word-learning behaviors observed in humans. When tested with the experimental materials from Yu & Smith (2007), the model was able to learn from cross-situational information, as humans were. In addition, 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. That is, because word-meaning mapping was part of the larger process that incorporated speaker intentions, the mutual exclusivity bias that children show was a by-product of this model trying to find the most likely lexicon and the most likely speaker intentions.

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.

Another 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.

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. For instance, consider an example of anaphoric *one* in English:

(3) Example of English anaphoric *one*
"I have a black cat. He's wonderful – don't you want *one*, too?"

To interpret the second utterance, we must figure out what *one* refers to: Is it a black cat or just a cat in general that the speaker thinks we should want? The linguistic

antecedent of an anaphoric word can help. If we know *one* refers to *black cat,* we can interpret the last part of the second utterance as "don't you want a *black cat*, too?"; if we know *one* refers to *cat,* we can interpret it instead as "don't you want a *cat*, too?". How do we know which one to choose?

This is where the category of the anaphoric word can help. One common representation of the syntax of *a black cat* is shown below in Figure 3. If *one* is an N', it can substitute for either node 1a or node 1b; if *one* is N^0, it can substitute for node 2 only. These nodes are compatible with different linguistic strings: N' is compatible with both *cat* and *black cat*, while N^0 is compatible only with *cat*. Thus, one way we can decide the category of *one* is by observing the strings it can substitute for.

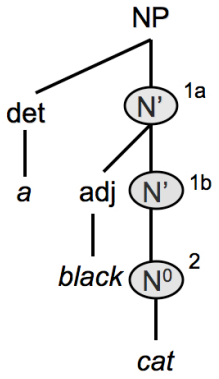

Figure 3. One representation of the syntax of *a black cat*. The numbered nodes (1a, 1b, 2) represent possible structures *one* might substitute for.

How then do we tell what strings it substitutes for? We can observe the intended referents. If *one* ever has *black cat* as its antecedent, we know *one* must be category N'. Here is one data point that would allow us to make this inference:

(4) Example of unambiguous data point for *one* = *black cat*
"I have a black cat, but you don't have one – you have a grey cat."

This utterance unambiguously indicates that *one* must have *black cat* as its antecedent. If *one* had *cat* as its antecedent, the interpretation would be that the listener does not have a cat of any kind. This would be a strange utterance since the listener clearly does have a cat (which is grey). So, we know that *one* can have strings like *black cat* as its antecedent, which means *one*'s category is N'. By understanding the intended referent (the black cat the listener does not have), we infer the linguistic antecedent (*black cat*) and so the syntactic category (N').

Interestingly, even in ambiguous examples like (X1), English adults often prefer *one* to take *black cat* as its antecedent – that is, they have a preference for *one* to substitute for the larger N' (node 1a in figure F1) rather than the smaller N' (node 1b in figure 3). Lidz, Waxman, & Freedman (2003) demonstrated that 18-month-olds appear to share these intuitions about anaphoric *one*'s interpretation, suggesting that children have acquired knowledge of *one*'s syntactic category by this age. This knowledge was traditionally considered unlearnable given the sparseness of unambiguous data like (4)

(Baker 1978, Hornstein & Lightfoot 1981, Crain 1991), which make up about 0.25% of children's anaphoric *one* input (Lidz, Waxman, & Freedman 2003). The traditional solution was then to assume that children possessed innate linguistic knowledge that referential words like *one* could not be category N^0.

Regier & Gahl (2004) discovered that a learner using Bayesian inference can learn from ambiguous examples like (3), in addition to unambiguous examples like (4). Specifically, for examples like (3), 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 Figure 3), the learner can then infer that *one* is of category N' as well. The only specific linguistic knowledge the learner required is (1) the definition of the hypothesis space (hypothesis 1: *one* = N' category, hypothesis 2: *one* = N^0 category), and (2) knowing to use both the unambiguous data and 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 in addition to the unambiguous data. 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 language data containing particular linguistic structures (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), and it is unclear if 18-month-old children would be able to do so.

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 in the corner is purring,* the yes/no question equivalent of this sentence is *Is the cat in the corner 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").

(5) Example of yes/no question formation
(i) Sentence:
The cat in the corner is purring.
(ii) Linear Rule: Move the first auxiliary verb
Is the cat in the corner t_{is} *purring*
(iii) Structure-Dependent Rule: Move the main clause auxiliary verb
Is [$_S$ *the cat in the corner* t_{is} *purring*]

It turns out that while both of these rules will account for simple yes/no questions like the one above, only the structure-dependent rule will account for behavior of more complex yes/no questions, such as in (6).

(6) 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*]

So, children must learn that structure-dependent rules are required to explain complex language word order properties like this one. Crain & Nakayama (1987) discovered that English children as young as three years old appear to know that structure-dependent rules are required for complex yes/no question formation in English. In addition, unambiguous examples like (6iii) that demonstrate this structure-dependent rule explicitly are quite 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 was therefore surprising that children seemed to 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 is structure-dependent, so they never consider the other kinds of analyses for their input, such as linear rules (e.g., Chomsky, 1971).

Perfors, Tenenbaum, & Regier (2006) 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. One main insight of their approach was 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.

This is determined by the posterior probability p(G | D), calculated as in (7). 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.

(7) Posterior probability of the grammar G, given the data D

$$p(G \mid D) \propto p(D \mid G)\, p(G)$$

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 are better than linear rules, and can apply this knowledge to 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 performing something 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 to appear) also used Bayesian learners to investigate how recursion might be instantiated in grammars. Recursion occurs when a rule has an expansion that eventually can call itself, as in (X4), where an S can be expanded into something containing an NP (8i) and an NP can be expanded into something containing an S (8ii).

(8) Recursive rule example

(rule i)	S → NP VP
(rule ii)	NP → N complementizer S

Recursion is of particular interest, as it has been argued to be a fundamental, 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. (to appear) evaluated grammars with and without recursive rules to decide which was optimal for child-directed speech data. Grammars with recursive rules allow infinite embedding (Depth 3+ in 9), while grammars without recursive rules allow embedding only up to a certain depth, e.g., 2 clauses deep (Depth 0, 1, and 2 in X5).

(9) 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 used had the same preferences as the one in Perfors, Tenenbaum, & Regier (2006): It attempted to identify the grammar that best balanced simplicity and the ability to account for the observed data. The issue with grammars containing recursive rules is that these grammars predict sentences that will rarely or never be observed, such as sentences with embedding of Depth 3+ in (9). So, recursive grammars may not be the best at predicting the observed data when compared to grammars that contain rules allowing only limited embedding. In addition, Perfors et al. investigated whether it was useful to have separate recursive rule types for subject-NPs (as in 9a) as opposed to object-NPs (as in 9b), since embedding is more often observed and more easily comprehended when it is object embedding (compare Depth2 in 9a to Depth 2 in 9b).

The Bayesian learner, when given child-directed speech data, inferred that the optimal grammar was one that had separate recursive rules for subject-NPs and object-NPs. In particular, the subject-NP rules included rules for limited embedding while the object-NP rules included only recursive rules. This is due entirely to the frequency of the recursion observed in object-NPs, as compared to the infrequency of recursion observed in subject-NPs.

From the perspective of language acquisition, the main result from Perfors et al. (to appear) is that a child able to approximate Bayesian inference well enough can discover when recursive rules are useful and when they aren't. More specifically, children do not need to innately know that recursion is required for representing object-NPs. Instead, if recursive rules are given as a potential option in their hypothesis space, children would be able to infer when recursive rules are most useful based on the data in their input.

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, word segmentation, word-meaning mapping, syntax-semantics mapping, and syntactic structure. Though computational 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. 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 attempted to provide a historical overview of statistical learning within the field of language acquisition, including experimental studies that demonstrate humans utilizing statistical information in sophisticated ways. We then discussed how computational modeling studies can contribute to our understanding of language acquisition and the role of statistical learning, focusing mainly on Bayesian inference. We also reviewed several computational studies that modeled acquisition of knowledge in different domains, specifically using Bayesian inference techniques. Sophisticated 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 may accomplish the feats of language acquisition that they do so quickly.

References:

Albright, A. & B. Hayes. 2002. Modeling Past Tense Intuitions with Minimal Generalization. In M. Maxwell (ed.), *Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology*. Philadelphia: ACL.

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.

Boersma, P. 1997. How we learn variation, optionality, and probability. In *Proceedings of the Institute of Phonetic Sciences of the Univ. of Amsterdam*, *21*, 43–58.

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

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.
Cairns, P., R. Shillcock, N. Chater, & J. Levy. 1997. Bootstrapping word boundaries: a bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, *33*, 111-153.
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.
Chater, N. and M. Oaksford. 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. 1981. *Rules and Representations.* New York: Columbia University Press.
Christiansen, M. & N. Chater. 1994. Generalization and connectionist language learning. *Ming and Language, 9,* 273-287.
Christiansen, M. H., J. Allen, & M. Seidenberg. 1998. Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13*, 221-268.
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. and M. Nakayama. 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.
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., 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
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., S. Goldwater, V. Mansinghka, T. Griffiths, & J. Tenenbaum. 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, and J. Tenenbaum. In submission. Modeling human performance in statistical word segmentation.
Frank, M.C., S. Goodman & J. Tenenbaum. 2009. Using Speakers' Referential Intentions to Model Early Cross-Situational Word Learning. *Psychological Science, 20(5),* 578-585.
Gambell, T. & C. Yang. 2006. Word Segmentation: Quick but not dirt. Ms. Yale University.
Geman, S. and D. Geman. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.
Gilks, W.R., S. Richardson, and D. J. Spiegelhalter, 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. & M. Johnson. 2003. Learning OT constraint rankings using a Maximum Entropy model. *Proceedings of the Workshop on Variation within Optimality Theory,* Stockholm University.
Goldwater, S., T. Griffiths, & M. Johnson. 2006. Interpolating between types and tokens by estimating power law generators. *Neural Information Processing Systems, 18*.
Goldwater, S., T. Griffiths, & M. Johnson. 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., J. L. Morgan, & P. K. Kuhl. 1993. Perceptual strategies in prelingual speech segmentation. *Journal of Child Language, 20*, 229-252.
Graf Estes, K., J. Evans, M. Alibali, & J. Saffran. 2007. Can Infants Map Meaning to Newly Segmented Words? *Psychological Science, 18(3)*, 254-260.
Griffiths, T. & J. Tenenbaum, 2005. Structure and strength in causal induction. *Cognitive Psychology, 51,* 334-384.

Griffiths, T. and A. Yuille. 2006. A primer on probabilistic inference. *Trends in Cognitive Sciences* 10(7). Supplement to special issue on Probabilistic Models of Cognition.
Hauser, M., N. Chomsky, & W. T. Fitch. 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., & H. Clark. 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.
Hayes, B. and C. Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry, 39*, 379-440.
Hornstein, N., & Lightfoot, D. 1981. *Explanation in linguistics: The logical problem of language acquisition*. London: Longmans.
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., A. Perfors, & J. Tenenbaum. 2007. Learning overhypotheses with hieracrchical Bayesian models. *Developmental Science, 10(3),* 307-321.
Kemp, C. & J. Tenenbaum. 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.
Knight, K. 2009. *Bayesian Inference With Tears.* Ms, University of Southern California. Available at http://www.isi.edu/natural-language/people/bayes-with-tears.pdf.
Legate, J. & C. Yang. 2002. Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review, 19,* 151-162.
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., H. Pasula, S. Russell, & Y. Peres. 2002. Decayed MCMC Filtering. In *Proceedings of 18th UAI*, 319-326.
Maye, J. & D. Weiss. 2003. Statistical cues facilitate infants' discrimination of difficult phonetic contrasts. *Proceedings of the 27th Annual Boston University Conference on Language Development.*
Maye, J., J. Werker, & L. Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* B101-B111.
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.
Oshma-Takane, Y., H. Takane, & T. Schultz. (1999). The Learning of First and Second Person Pronouns in English: Network Models and Analysis. *Journal of Child Language, 26(3),* 545-575.
Pater, J. 2009. Weighted Constraints in Generative Linguistics. *Cognitive Science, 33*, 999-1035.
Pelucchi, B., J. Hay, & J. Saffran. 2009a. Statistical Learning in Natural Language by 8-Month-Old Infants. *Child Development, 80(3)*, 674-685.
Pelucchi, B., J. Hay, & J. Saffran. 2009b. Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition,* 244-247.
Perfors, A., Tenenbaum, J., Gibson, T., and Regier, T. (forthcoming). How recursive is language? A Bayesian exploration. *Linguistic Review*.
Perfors, A., J. Tenenbaum, & T. Regier. 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.
Perruchet, P. & S. Desaulty. 2008. A role for backward transitional probabilities in word segmentation? *Memory & Cognition, 36(7),* 1299-1305.
Pinker, S. 1984. *Language learnability and language development*. Cambridge, MA: MIT Press.
Potts, C., J. Pater, K. Jesney, R. Bhatt and M. Becker. 2009. Harmonic Grammar with Linear Programming: From linear systems to linguistic typology. *Phonology*, *27(1),* 1-41.
Pullum, G. & Scholz, B. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review, 19,* 9-50.
Redington, M., C. Chater, & S. Finch. 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. and E. Hardisty. 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. and J. McClelland, editors. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, chapter 18. MIT Press.

Saffran, J., R. Aslin, & E. Newport. 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.

Shi, L., T. Griffiths, N. Feldman, & A. Sanborn. (to appear). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review.*

Smith, L., & C. Yu. 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.

Swingley, D. 2005. Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology, 50*, 86-132.

Tenenbaum, J. 1996. Learning the structure of similarity. In D. Touretzky, M. Mozer, & M. Hasslemo (eds.), *Neural Information Processing Systems, 8*. Cambridge: MIT Press.

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., J. McClelland, F. Pons, J. Werker, & S. Amano. 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.

Werker, J., F. Pons, C. Dietrich, S. Kajikawa, L. Fais, & S. Amano. 2007. Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition, 103(1)*, 147-162.

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

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., & J. Tenenbaum. 2007. Word learning as Bayesian inference. *Psychological Review, 114*, 245-272.
Yang, C. 2004. Universal Grammar, statistics, or both? *Trends in Cognitive Sciences, 8(10),* 451-456.
Yu, C. & L. Smith. 2007. Rapid Word Learning Under Uncertainty via Cross-Situational Statistics. *Psychological Science, 18(5)*, 414-420.