Grammar Learning as Model Building

LouAnn Gerken and Colin R. Dawson

University of Arizona

This research was supported by NICHD #R01 HD042170 and NSF 0950601 to LAG.

This chapter focuses on recent causal model-building accounts of statistical learning by infants and young children, of which Bayesian inference models are the most popular and well known. We begin by placing model-building in contrast to both traditional nativist and learning accounts of language acquisition. Like traditional learning accounts, model-building accounts assume that the input to the child is very rich and that learners are able to take advantage of this richness and find statistically stable patterns in the input. However, like nativist accounts, model-building accounts assume that the outcome of language acquisition is a grammar (a mental representation of the system that gave rise to the observed data). Next, the chapter discusses five hallmarks of a model-building learner and reviews some of the developmental data that are consistent with these hallmarks. Because much of the work that sparked the human developmental research began in machine learning models, the chapter outlines the relation of these models and theories of human language learning. Finally, the chapter ends with a summary and speculation about the next challenge for the view that infants and children construct causal models of their linguistic input.

Prediction, Explanation, and Grammar

A central tenet of language development is that children generalize beyond their input, producing word and morpheme combinations that they have never encountered. The classic example is the child who produces the past tense of *go* as *goed*: It is unlikely that the child ever heard the word *goed* before, because it is in fact not part of the adult linguistic system. Most scientific and popular writing about the generalization mechanism and the acquisition of human language more broadly has focused on the "nature vs. nurture" question (e.g., Elman, et al., 1996; Pinker, 1994). Nevertheless, it is possible to re-construe nearly all discussion of language development in terms of linguistic creativity: Is this central characteristic of human language merely a form of *predicting* new linguistic encounters from old data, or does it require that the language user have tacit access to a possible *explanation* of prior linguistic experiences? In the domain of language, an explanation of the input is a *grammar*.

Over the past 25 years or so, the main theoretical alternatives in the grammar vs. no-grammar frameworks indeed mapped neatly onto the nature-nurture debate. Within both the Principles and Parameters and Optimality Theory frameworks (e.g., Chomsky & Lasnik, 1993; Prince & Smolensky, 1997), a set of possible grammars was innately given to the infant, and language learners needed only minimal input to rule out by *deductive inference* all but the grammar(s) consistent with the input of their language community. The main no-grammar alternatives are all forms of *inductive inference*, which include various versions of Associationism or Connectionism. Here, the language learner encodes a set of associations between linguistic experiences (e.g., present and past tense forms, words and their referents), and the form of encoding coupled with

compression of stored associations leads to generalizations like *goed* (e.g., Rumelhart & McClelland, 1987).

However, the mapping between innate vs. learned on the one hand, and explanation vs. prediction, on the other, was not always so neat. While it is true that proponents of the no-grammar view (e.g., Skinner, 1957) have largely held firm to the notion that language is learned through experience, mid-20th-century discussions of the grammar view seemed to allow for more learning from experience than do more recent discussions. For example, Chomsky (1959) in his response to Skinner (1957) notes that

> The child who learns a language has in some sense constructed the grammar for himself on the basis of his observation of sentences and non-sentences (corrections by the verbal community). Study of the actual observed ability of a speaker to distinguish sentences from nonsentences, detect ambiguities, etc., apparently forces us to the conclusion that this grammar is of an extremely complex and abstract character, and that the young child has succeeded in carrying out what from the formal point of view, at least, seems to be a remarkable type of theory construction. Furthermore, this task is accomplished in an astonishingly short time, to a large extent independently of intelligence and in a comparable way by all children. Any theory of learning must cope with these facts.

This quote contains within it a snapshot of the modern study of language development. For our immediate purposes, two phrases stand out: “on the basis of his

observation of sentences and nonsentences" and "remarkable type of theory construction." Thus, it appears that, at this time, Chomsky thought that children constructed an explanation of their observations of language – a fairly clear de-coupling of grammar from innateness. However, the notion that children had any opportunity to observe nonsentences by being corrected for producing ungrammatical forms fell by the wayside in the next decade, partly due to studies by Roger Brown and his colleagues showing that parents correct misstatements of fact, but not ungrammatical sentences (Brown, Cazden, & Bellugi, 1969). That is, it has been argued that children receive only positive evidence in the form of the sentences produced around them. The view that children are deprived of negative evidence, coupled with formal proofs of the necessity of such evidence for learning (Gold, 1967), led Chomsky and others to abandon the notion that language learners *construct* grammars as explanations for their linguistic experiences (but see Valian, Winzemer, & Erreich, 1981). Rather, in order to maintain the notion that language development entails the acquisition of a grammar, many in the field of language development adopted the view that the set of grammars must be innate.

In the last decade, however, the possibility that children construct grammars as explanations for their linguistic input has resurfaced as Bayesian learning models have seen a growth in popularity. An essential component of such models is that the experiences we have in the world are treated as samples from a distribution of possible experiences. If we stop there and say that the learner's goal is to make inferences about the characteristics of the population from which the sample is drawn, we could say that a Bayesian learner is an inductive learner, albeit one of a different sort than an

associative learner. However, Bayesian inference can go beyond merely *generalizing* from a sample, by making inferences about genuine *causes* of the data. Inference about causes or explanations of data that have been encountered is *abductive inference*, of which Bayesian inference can be thought of as a probabilistic version. If, as suggested earlier, we think of a linguistic grammar as an explanation for linguistic data, then a Bayesian account of grammar-acquisition is a form of probabilistic abduction. Thus, the study of language development encompasses all three types of inference detailed by Aristotle in his *Prior Analytics:* deduction, induction, and abduction.

The above quote from Chomsky (1959) contains at least one additional phrase of note: He describes the child language learner as operating "to a large extent independently of intelligence." Thus, if we are to revisit the notion of the child as inferring a grammar from the input sample that she has encountered, we must ask what kind of intelligence might be involved, and determine whether the child or infant in fact has the capacity to make such an inference. Luckily, in the past two decades, the field has amassed ample evidence that human infants are capable of encoding and performing a range of calculations over their linguistic input. In particular, we now have evidence that young learners can compute over linguistic input a variety of properties, including descriptive statistics (e.g., Maye, Werker, & Gerken, 2002), transitional probabilities between adjacent strings (e.g., Saffran, Aslin, & Newport, 1996), transitional probabilities between non-adjacent strings (Gómez, 2002; Santelmann & Jusczyk, 1998), relations between identical elements (Dawson & Gerken, 2009; Gómez & Gerken, 1999; Marcus, Vijayan, Rao, & Vishton, 1999), and morphological paradigms (Gerken, Wilson, & Lewis, 2005) (for a review, see Gerken, 2005). Recent evidence

suggests that young learners are able to perform the sorts of computations required of a Bayesian learner in particular, and an abductive learner more generally. In the following two sections of this chapter, we will first describe what we take to be the hallmarks of an abductive learner, and then present evidence that infants and children have the cognitive wherewithal to meet these requirements. The behavioral evidence will come from language as well as other cognitive domains. Finally, we evaluate more generally what we see as the advantages and challenges for the putative language learner who seeks causal explanations of their input, and point out some parallel advantages and challenges for a similar approach to artificial intelligence.

What Every Young Abductive Learner Needs To Know

Before we turn to the evidence that infants and children generate and test causal models of their world, we need to consider what characteristics are hallmarks of such a learner. We will describe five characteristics – the first three are central to formal accounts of Bayesian inference, while the fourth and fifth are more general properties of abductive learners.

Samples and populations

As noted above, Bayesian inference is, in essence, a way of making guesses about the population distribution that gave rise to a data sample that you have encountered. This description of Bayesian inference suggests that a Bayesian learner needs to know something of the relationship between samples and populations from which the samples are drawn. In particular, when given evidence about a population, learners should be able to predict a likely sample. For example, if shown a box

(population) containing mostly red balls with only a few white balls, a set of balls randomly[1] selected from this box (sample) should also contain mostly red balls. Conversely, when shown a sample, learners should be able to select among a set of possible populations the one that is most likely to have generated the sample. Using the box of balls example, a sample of 4 red balls and 1 white ball is more likely to have come from a box with mostly red balls than one with mostly white balls. Below, we will examine the evidence that infants under the age of 1 year behave in accordance with this requirement of Bayesian inference.

Random vs. strong sampling

A Bayesian learner needs to know that the relation that holds between samples and populations only does so under random sampling. If some other more selective type of sampling is in play, different relations should hold. Sampling only those items from a population that meet certain criteria (e.g., things I like) is called "strong sampling." For example, if the person selecting balls from a box containing mostly red balls indicates a preference for white balls, then a learner should not be surprised if the sample does not reflect the distribution of the population. We will examine evidence that infants make different inferences from the same input if they are given evidence that the input reflects a random sample vs. a sample that was selected by human volition.

The Size Principle

A third aspect of Bayesian inference concerns how learning proceeds when the same sample might have come from two different populations that are in a subset-

[1] Here and in the subsequent discussion, by "random" sampling, we mean a process in which each member of a population has an equal chance of being selected. This is in contrast with other forms of sampling discussed just below.

superset relation. To handle these cases, Tenenbaum and Griffiths (2001) noted a property that falls out of Bayes' Rule under a random sampling assumption: the Size Principle. This principle states that, when input supports two possible models in a subset-superset relation, and each new data point is consistent with both models, the learner should increasingly favor the smaller model as more data are observed. For example, if a learner encounters the set of numbers 30, 400, 90, 60, at least two possible populations are consistent with this input: all numbers divisible by 10 and all numbers divisible by 5. If the actual population is really all multiples of 5, it is a "suspicious coincidence"[2] that sample so far encountered contains no numbers that are divisible by 5 but not divisible by 10. Therefore, it is more likely that the sample comes from the smaller population of numbers that are divisible by 10. Importantly, the Size Principle has a bigger and bigger (exponential) effect the more input you encounter, as the "suspiciousness" of the coincidence grows. Therefore, if you have only a single input example (e.g., 30), you do not yet have a very strong reason for choosing divisible by 10 over divisible by 5 as the more likely underlying population, because a single sample from the latter would be divisible by 10 half the time anyway. Is there evidence that infants and young children behave in accordance with the Size Principle, increasingly favoring the smaller population as they encounter more data?

Hypothesis updating

Perhaps it goes without saying that a learner who is attempting to build models of the world to account for her experiences must update those models as new data come in. However, it is useful to consider three differences between abductive learners, who

[2] More precisely, the probability of obtaining a sample with those characteristics purely by chance (i.e, random sampling) is quite low.

build and update explanatory models, and inductive learners of an associative variety. The latter generalize based on an encoding of the distribution of input forms, not on explanatory models. If this distribution changes with the addition of more data, the generalization will change. One difference between these two types of learners concerns the usefulness to learning of input examples that are different from previous examples (new input types) vs. input examples that are copies of previous examples (new input tokens that are not new input types). New types promote learning and abstraction for an abductive learner, whereas exposure to additional tokens of familiar types need only impact "within-type" inferences. In contrast, many associative models do not explicitly differentiate between types and tokens, both of which contribute to the encoded distribution of the input data (e.g., Xu & Tenenbaum, 2007a).

A second difference between an abductive learner and an associative learner concerns the amount of data required for changes in generalization. Because an abductive learner has the ability to represent the differences between competing models, a single piece of data that rules out (or is at least very unlikely under) Model A, but is consistent with Model B, can be sufficient to make Model B a considerably better explanation of the data than Model A. However, because changes in the generalizations made by an associative learner must be driven by the overall input distribution, it is more difficult for a single counterexample to shift the distribution. That is, while both types of learner can accommodate noisy input, only the model-based learner can differentiate environments that *should* be noisy (under the current model) from environments where deviations from predictions signal new structure. This "novelty-detection" ability of a model-based learner is discussed in the context of

generative machine-learning models below.

Finally, if Model B is a more general model of the world than Model A, it can "explain away" the data that Model A was originally created to account for, potentially making those data less valuable in future contexts. This explaining away phenomenon is illustrated by the following intuitive example, modified slightly from Pearl (1988). Consider a lawn, which is equipped with a sprinkler system. Discovering the lawn is wet (W) is evidence for the hypothesis that the sprinkler system ran recently (S). Finding out that the lawn next door is wet (N) is independent evidence for rain (R), which would also explain W. Although N by itself would have no impact one way or another on the sprinkler hypothesis (assuming a sprinkler system which is insensitive to the weather), it does undermine – that is, *explain away* – the original *evidence* for S via its support for the more general "wetness generator" R. As we will discuss further below, generative (abductive) models of machine learning produce this inference easily; however it is quite difficult, and at the very least unnatural, for discriminative (inductive) maching learning models to produce it. Do infants and children show evidence of changing the generalizations that they make in keeping with these characteristics?

Information-seeking

A learner who can represent multiple models of the world can also potentially represent the degree of uncertainty inherent in the dimensions of each model. Such a learner, like a scientist, can selectively seek new data that are informative about the aspects of the environment about which she is the most uncertain. Indeed, Bayesian experimental design is a field in statistics that deals formally with exactly this problem in scientific inquiry. Although this sort of information-seeking is not a formal requirement of

abductive learners, it would be difficult to create an information-seeking learner who could not represent competing explanatory models. Therefore, if infants and children have the capacity to seek information in a directed way, we would have some additional support for the notion that they, like scientists, are model-builders.

Developmental Data

In this section, we consider available behavioral data that address each of the five hallmarks of an abductive learner outlined above. Whereas the studies dealing with the relation between samples and populations and with random vs. strong sampling do not focus on language, the other three headings include a number of studies that apply abductive inference to word and linguistic structure learning.

Samples and populations

Xu and Garcia (2008) performed a series of experiments with 8-month-olds to ask how much these infants implicitly understood about the relation between samples and populations. Each experiment began by allowing infants to play with three red and three white ping-pong balls in a small container. In one experiment, each infant was shown four familiarization trials with a large box that contained either mostly red ping-pong balls and a few white ones (2 familiarization trials) or mostly white ping-pong balls and a few red ones (2 familiarization trials). On each test trial, the same box was presented, but its contents could not be seen by the infant. The experimenter closed her eyes and pulled out a series of five balls, placing them in a holder for infants to see. On half of the test trials, the holder contained four red and one white ball, while on the other half, it contained four white and one red. After the five balls were shown, the

experimenter opened the front panel of the box so that the infant could see whether it contained mostly red or mostly white balls. Infants looked longer on those trials in which the ratio of red to white balls in the sample of five did not match that of the population of balls in the box than when the sample was representative. Another experiment helped to rule out the possibility that infants were responding to a mismatch in the proportion of balls of each color in the holder and box, but not treating the balls in the holder as a sample from the box. In this experiment, the experimenter drew the balls from her pocket instead of from the box. Here, infants showed no preference based on the relation between the balls in the holder and those in the box.

A final experiment asked if infants could not only reason from samples to predicted populations, but also from populations to predicted samples (also see Denison & Xu, 2009; Teglas, Girotto, Gonzalez, & Bonatti, 2007). This experiment began like the first with four familiarization trials in which the contents of the box (population) were shown. In contrast with the earlier experiments, infants could see the contents of the box on each test trial before the experimenter began to extract balls. However, the box was closed before the sample of five balls was drawn. Again, the experimenter drew four red balls and one white one or four white and one red. This time, however, the infant never saw the large box again, and only their looking times to the sample of five balls were measured. As before, infants looked longer when the sample failed to match the proportion of red and white balls in the population, even though the population was only available to the infant in her memory.

These studies suggest that, at least with discrete physical objects or representations thereof (Teglas et al, 2007), infants as young as 8 months are able to

make inferences from samples to populations and from populations to samples.

Random vs. strong sampling

Did the infants in the studies just described implicitly understand that the proportion of red to white balls in the sample and population should match only under random sampling? To ask this question, Xu and Denison (2009) performed an experiment with 11-month-olds similar to the ones described above, but beginning with a "preference" phase, in which any preference for a particular ball color the experimenter might have was demonstrated. For infants in the random sampling condition, the experimenter appeared to have no preference. In this condition, infants saw a holder with three red and three white balls. The experimenter picked up the red balls and smiled and then picked up the white balls and smiled (the order of the colors was counterbalanced). Other infants participated in an experimenter preference condition in which the experimenter picked up three balls of one color and smiled, and then picked up the same three balls and smiled. After the preference phase, the experimenter selected five red balls or five white balls from the box, placing that sample in a holder. For infants in the random sampling condition, the experimenter closed her eyes before selecting the sample from the population. For half of the infants who saw the experimenter show a preference, the experimenter drew the sample while looking into the box, while for the other half, the experimenter was blindfolded during sample selection. For all conditions, after the sample of five balls was selected, the experimenter revealed the contents of the box (population). For infants in the random sampling and blindfolded conditions, looking times were longer when the sample did not match the majority ball color of the population (e.g., red sample, mostly white box and

vice versa). However, for infants in the condition in which the experimenter both showed a preference during the preference phase and had the ability to see the contents of the box when the sample was drawn, infants appeared to expect the sample to match the previously seen preference, regardless of the match of the sample and population. Therefore, it appears that infants have some understanding of when samples should reflect populations (random sampling) and when they should not (strong sampling).

Another study suggests that children as young as 20 months can use the appearance of strong sampling (mismatch between sample and population) to infer the intent of a person who selected the sample (Kushnir, Xu, & Wellman, 2010). Children saw a person select five toys of one type (either ducks or frogs) out of a box. The boxes contained either mostly the selected toy or mostly the unselected toy in an 18% vs. 82% split. The box of toys was then put away, and the child was presented with two bowls of toys – ducks in one bowl and frogs in the other. The experimenter held her palm up between the two bowls and said to the child "Oh goody! Just what I wanted! Can you give me one?" More children touched and offered the target toy (14) than the alternate toy (10) when the selection of the target violated random sampling (came from a box with only 18% of that toy). In contrast, more children touched and offered the alternate toy (13) than the target toy (5) in the random sampling condition. The authors interpret these findings to suggest that that infants and children can not only differentiate random from strong sampling, but that they can also use the appearance of strong sampling to make important inferences about human motivation.

The Size Principle

As noted above, the Size Principle applies to inferences about the relative likelihood of two populations that might have produced a sample, with one population being a subset of the other. Although the smaller population is treated as more likely when even a single input datum has been encountered, the relative likelihood of the subset population increases exponentially with each datum that is consistent with the subset. At least two published studies, as well as an unpublished study from our lab, provide support that infants and children behave in accordance with the Size Principle in language learning. Xu and Tenenbaum (2007b) showed 3- to 4-year-olds either a single Dalmatian or three different Dalmatians and labeled each example *fep*. They then asked children to give them another fep from a set of toys that included Dalmatians, non-Dalmatian dogs, and other animals. Children always treated a Dalmatian as the most likely extension of *fep*. However, when the label was applied to three different Dalmatians, children (and adults) were less likely to select a dog that was not a Dalmatian than when the label was applied to a single Dalmatian. Importantly, Xu and Tenenbaum (2007b) compared a Bayesian model with the Size Principle to both an associative (Hebbian) learning model and a model that included the Subset Principle (e.g., Berwick, 1986), in which the most narrow hypothesis is always preferred. The Bayesian model better matched the behavioral data.

A study with 9-month-olds suggests that something like either the Size Principle or Subset Principle is at work in the first year of life. Gerken (2006) presented 9-month-olds with four three-syllable strings that obeyed either an AAB (1st and 2nd syllables are the same) or ABA pattern (1st and 3rd syllables are the same). Half of the infants were familiarized to strings in which the B syllable was always the syllable *di,* whereas the

other half heard strings in which the B syllable varied among four different syllables. Infants in the latter condition generalized to new AAB or ABA strings, whereas infants in the former *di* condition only generalized to new AA*di* or A*diA* strings. This study suggests that when learners encounter data that are consistent with two different grammars (AAB vs. AA*di*), they treat the more narrow grammar as more likely, at least when they have encountered four different input examples. Recent work in our lab suggests that, consistent with the Size Principle, and inconsistent with the Subset Principle, 9-month-olds who are presented just a single example of an input string (e.g., *leledi*) generalize to new AAB and new AA*di* strings (Gerken, Dawson, Chatila, & Tenenbaum, in preparation). This result, coupled with the one in Gerken (2006), suggests that infants become less likely to make the broader generalization as the number of examples that are consistent with the narrower grammar increases from one to four.

Hypothesis updating

As noted above, a learner that uses input data to construct explanatory models changes their bases of generalization differently than an associative learner. One difference between a model-building learner and an associative learner concerns the role of types and tokens in generalization. We have already described the study by Xu and Tenenbaum (2007b), in which they labeled as *fep* a single Dalmatian vs. three Dalmatians. Another manipulation in that study was to present children with three identical Dalmatians (3 tokens of a single type) vs. three different Dalmatians (3 different types). Children were more likely to treat *fep* as referring to the category Dalmatian when they were shown three different types than when they were shown one

token of one type or three tokens of one type. Xu and Tenenbaum demonstrate that an associative model does not perform differently when given three types than when given three tokens of one type, whereas a Bayesian model better reflects the children's performance.

Several studies with infants also demonstrate that they are more likely to generalize when they are presented with three different input types from a category than when presented with three tokens of a single type. For example, Needham and colleagues (Needham, Dueker, & Lockhead, 2005) exposed 4-month-olds to between one and three exemplars of a visual category and then tested them on new items that were either consistent or inconsistent with the category. They found that infants did not generalize based on one or two exemplars of the category, but they did generalize from three exemplars (also see Quinn & Bhatt, 2005). Gerken and Bollt (2008) found that infants exposed to three- and five-syllable words that exhibited patterns of stressed and unstressed syllables based on one of two artificial languages were able to generalize the principle 'stress syllables ending in a consonant' after hearing three different syllables ending in a consonant (types) in the input, but not after hearing multiple tokens of just one syllable ending in a consonant (1 type).

Learners who construct explanatory models are also able to reject models very quickly when faced with conflicting data. For example, Kushnir and Gopnik (2007) determined that 100% of 3- and 4-year-olds were able to learn that placing an object on a detector activated the detector, whereas considerably fewer children were able to learn that placing on object over the detector caused activation. This finding is consistent with the view that children have a strong *a priori* belief that physical contact

causes physical changes. However, when children were shown once that putting the block on the detector did not activate it, while holding a block over the detector did activate it, they abandoned their prior physical contact hypothesis as an explanation about how the detector worked.

Gerken (2010) showed that 9-month-olds also rapidly reject a prior hypothesis with a handful of counterexamples. Recall that infants tested by Gerken (2006) made only the narrower AA*di* or A*di*A generalization when presented for 2 min. with four three-syllable words exhibiting this pattern *leledi, wiwidi, jijidi*, *dededi* (AA*di*) or *ledile, widiwi, jidiji, dedide* (A*di*A). Gerken (2010) presented the same stimuli, but added three counterexamples that did not contain *di* to toward the end of the list, such that the last five stimuli were *wiwije, jijidi, dedewe, jijili, wiwidi* (AAB) or *wijewe, jidiji, dewede, jlijii, wididwi* (ABA). As a control, another group of infants heard 2 min. of music followed by the same five stimuli. Infants who heard 2 min. of linguistic stimuli now generalized to new AAB or ABA patterns that did not contain *di*, whereas infants who heard the music plus five linguistic stimuli did not generalize. This study, coupled with the failure of infants studied by Gerken (2006) to generalize to strings that did not contain *di* when not presented with the three counterexamples, suggests that the three counterexamples in the context of the 2 min. of AA*di* or A*di*A stimuli caused infants to change their favored hypothesis from a narrower one to a broader one.

In some circumstances, it appears that infants' rejection of one hypothesis in favor of another happens over considerably longer stretches of developmental time. In these instances, older infants appear to form a general model of a domain that subsumes, or "explains away", previously compelling data. In one such apparent

example, Gerken and Bollt (2008) found that 7-month-olds, but not 9-month-olds, were able to learn a stress-assignment rule in which syllables beginning with /t/ are stressed. One possible reason behind the developmental change might be the types of statistics that younger and older infants were able to perform on English words due to differences in vocabulary. The raw frequency of stressed syllables beginning in /t/ is relatively high, and if a learner knows primarily monosyllabic words, it would be possible to entertain the hypothesis that different syllable onsets differentially assign stress. However, the conditional probability of stressed and unstressed syllables beginning in /t/ or any other consonant suggests that syllable onsets are not implicated in stress assignment. But in order to calculate meaningful conditional probabilities on stressed vs. unstressed syllables, learners would need to know a large enough number of polysyllabic words. Thus, the developmental change between 7 and 9 months may be due to changes in receptive vocabulary, which in turn allows a more complete assessment of the factors that affect stress assignment, ultimately ruling out onsets as a candidate.

Another example of developmental change being driven by a more complete or general model of a domain can be seen in music. Marcus and colleagues (2007) demonstrated that, while 7-month-olds were able to generalize to new AAB vs. ABA patterns in language (also see Gerken, 2006, 2010), they were unable to do so for musical tones. Dawson and Gerken (2009) replicated the null result in music with 7-month-olds but found that 4-month-olds could generalize to new AAB vs. ABA musical patterns. One possible reason for the developmental change is that adjacent or near adjacent repetition is indeed very common in Western tonal music (e.g., Dawson,

2007).[3] However, the frequency of such repetition can be subsumed or explained away by a more general locality constraint: notes nearby in the scale to the immediately preceding note are more likely to occur than notes that are farther away in the scale (Dawson, 2007; Dowling, 1967; Ortmann, 1926; Temperley, 2008). On this view, 4-month-olds, who do not yet have a sufficiently complete model of Western tonal music, think that repetition in AAB or ABA musical patterns is something that needs explanation (perhaps just as 7-month-olds think that a correlation between syllable stress and starting with /t/ is something that needs explanation). However, if 7-month-olds have begun to develop a model of Western tonal music, they may no longer treat repetition in AAB or ABA strings as requiring a separate model or rule.

In order to explore this explaining away notion experimentally, Dawson presented adults with a musical context phase in which note repetition could either be subsumed under a more general locality constraint or not (Dawson, 2010; Dawson & Gerken, 2011). Adults who heard the prior explaining away context were less likely to generalize on the basis of note repetition in a subsequent AAB-style learning task than adults for whom repetition required a separate explanation in the context phase.

The studies in this section suggest that learners change their generalizations about their input in a manner not predicted easily by associative learning models. Areas of difference between the two accounts include the usefulness for generalization of types vs. tokens, the amount of input needed to change generalizations, and the way in which the winning generalization changes how subsequent data are perceived or encoded.

[3] Adjacent or near adjacent repetition is *not* very common in language, which may be why such occurrences continue to require an explanation by learners of all ages in language learning tasks.

Information seeking

Scientists are the prototypical abductive learners, because they seek causal explanations of their data and can often say what kind of new data is needed to differentiate between competing explanations. Indeed, the ability to seek certain types of data is central to the model-building that transpires in science. Do young children have a similar ability to seek out new data that would allow them to decide among competing hypotheses?

One of the clearest demonstrations that young children seek out data under just those conditions where new data would be more helpful comes from Schulz and Bonawitz (2007). In this study, an experimenter presented almost-5-year-olds with a toy with two levers, each of which caused a different figure to pop up when pushed. Children were given either ambiguous or unambiguous evidence about how the toy worked. For example, showing the child that both figures pop up when the two levers are pressed simultaneously is ambiguous evidence for the inference that each figure pops up independently, depending on which lever is pressed. When children were given the chance to play with the old toy or a new toy, children who received ambiguous evidence about the cause of the figures popping up were more likely to choose the old toy than children who received unambiguous evidence.

Many studies of infants assume an ability to determine whether visual or auditory information presented to them in the laboratory is already known to them or is something new to be learned (Hunter & Ames, 1988). In these studies, infants are allowed to experience a stimulus for as long as they show interest. When interest flags (when the infant has habituated), it is assumed that the infant has learned as much as

subjectively possible. For example, infants who were interrupted while exploring a set of toys were more likely to continue exploring the original set of toys when given the choice of this set or a new set, than infants who were allowed to play with the toys until they had begun to turn their attention elsewhere. The latter group was more likely to choose a new set of toys when given a choice (Hunter, Ames, & Koopman, 1983).

A recent study asked whether toddlers' level of interest in linguistic stimuli reflected a form of information seeking. Gerken, Balcomb, and Minton (2011) allowed 17-month-olds to listen to words from a Russian morphological gender paradigm for as long as they showed interest. Half of the infants listened to stimuli that could be independently classified as learnable, while the other half listened to stimuli that could be independently classified as unlearnable. Across two experiments, infants exposed to the learnable stimuli took more trials and more overall time to habituate. These infants also showed more reversals in listening times, with later trials having longer listening times than earlier trials. This study suggests that even very young learners might monitor their state of uncertainty and attend longer to input that appears to hold new information.

Although the studies of information seeking by young learners do not paint as clear a picture of model-building as the studies described under the other four sections, they suggest an important direction for researchers interested in whether learners are model-builders.

Summary of the Developmental Data

We suggest that the studies in this section are generally compatible with the view of human infants and children as model builders. Nevertheless, we acknowledge that for

many or even most of the studies presented, other types of learners, including associative learners, might demonstrate similar abilities. Although some of the studies we presented directly compared model-building learners and associative learners (e.g., Xu & Tenenbaum, 2007b), most have not. Therefore, even though the five abilities that we have identified seem to us central to a model-building learner, future comparisons between different computationally instantiated learners are needed. Because of the central role of computational instantiation in deciding among theories of human learning, the next section of this chapter provides what we hope is a helpful outline of machine learning models and their parallels to human learning.

Parallels to Machine Learning

The recent resurgence of interest in the notion that language learning is a type of model-building has been sparked by Bayesian approaches to machine learning. This section will survey various approaches to learning in artificial intelligence and draw parallels to the types of developmental questions we discussed above.

The field of artificial intelligence (AI) has incorporated all three modes of inference – deduction, induction, and abduction, at various times. In the earliest stages of AI, purely symbolic, Boolean logic algorithms were developed to carry out automated deductive inference. These systems gave rise to automated theorem-proving computers, and discrete algorithms for playing highly constrained games such as chess. However, such methods are only useful in cases where (nearly) complete information about a problem domain is available. For problems where it is desirable to generalize beyond the available data, inductive and abductive methods are needed.

The extension of artificial intelligence methods beyond rule-based deductive inference systems led to the creation of a new field of *machine learning*, which borrowed much more from fields outside computer science, such as statistics and engineering, than did traditional AI. Today, machine learning methods can be roughly divided into *discriminative* and *generative* methods, each with a different approach to learning about the structure of the environment. Loosely speaking, discriminative methods can be said to carry out inductive inference, generalizing from the data without modeling its causes explicitly. In other words, the system learns to discriminate among patterns of data based on what generalizations (often categories) they are associated with. Generative methods, on the other hand, carry out abductive inference: equipped with a model of how observable data arises from a set of causes, they are able to make inferences concerning which causes are responsible for a given data point. Because they represent the causal structure of the environment, they are capable of generating new, hypothetical data, a capacity outside the purview of discriminative methods.

Both generative and discriminative approaches to learning can be further categorized based on their learning goals and the nature of the input provided (see Table 1). Supervised learning occurs when the system is provided with training data for which a "correct output" (often a category label) is known. For these learners, the goal is to determine how to generalize beyond the training examples associated with each output to choose outputs for unlabeled data. The goal of unsupervised learners, on the other hand, is simply to learn about relationships among data points. Most often, this takes the form of clustering data into groups, without attaching specific meaning to the classes. Intermediate approaches between these extremes exist as well. In this chapter

we are more concerned with the generative-discriminative distinction than the supervised-unsupervised distinction.

	Discriminative	**Generative**
Supervised	Neural networks Support Vector Machines Radial Basis Functions Decision Trees	Naïve Bayes Classifier Hidden Markov Models General Bayes Nets Relevance Vector Machines Markov Random Fields
Unsupervised	Clustering: k-Means k-Nearest Neighbors Self-Organizing Maps Feature-Extraction: Principal Components Analysis (PCA) Multidimensional Scaling (MDS)	Clustering: Expectation-Maximization for Mixture Models Dirichlet Process Modeling Feature Extraction: Bayesian PCA Independent Components Analysis (ICA)

Table 1: Some popular Machine Learning techniques

Discriminative methods in machine learning

Discriminative approaches to machine learning have both deterministic and probabilistic versions. One of the earliest deterministic systems capable of generalizing beyond training data was the perceptron learning algorithm (Rosenblatt, 1958), based on the McCulloch-Pitts "logic gate" model of a neuron (McCulloch, 1943). This simple algorithm assigns binary classes to arbitrarily high-dimensional data by learning a set of linear weights governing the relationship between each input and output feature. Rumelhart and McClelland (1986) developed a workable method for training networks with multiple layers of McCulloch-Pitts-style neurons, augmented with continuous, rather than binary threshold, activation functions. With continuous activations, gradient

assignment of classes or output features can be made, which may or may not be interpretable as probabilities.

Rumelhart and McClelland (1987) presented one of the first major alternatives to the Chomskian, rule-based framework of language acquisition when they showed that a simple neural network could learn the past-tense forms of English verbs, generalizing to novel verbs, and even learning sub-patterns among "irregular" forms like *sing/sang – ring/rang*, and *read/read – feed/fed*.

The most common variety of unsupervised learning involves clustering of data points into categories. These algorithms rely on the assumption that observations that belong to the same category should be more similar to each other than they are to members of another category. Provided a sensible measure of similarity is defined, "clusters" of observations can be inferred from data, and new data points assigned (either probabilistically or in a "winner-take-all" manner) to a cluster. Algorithms such as k-means clustering (MacQueen,1967) accomplish this in a discriminative manner by choosing cluster centers and cluster assignments such that the aggregate distance from observations to their respective centers is minimized.

Clustering the input into categories is clearly an important aspect of language acquisition at every level, from low-level phonological categories to abstract semantic categories. In addition, the sequential nature of language presents a somewhat different form of clustering in the need to segment continuous input into constituent units (from phoneme to sentence, and beyond). Saffran, Aslin and Newport (1996) proposed a simple probabilistic discriminative method for finding boundaries between words: namely, by tracking transitional probabilities between successive syllables and placing a

boundary in the observed stream where these probabilities are low. Elman (2001) suggests that a Simple Recurrent Network, which is a form of neural network with internal feedback, could accomplish something similar by tracking its own prediction uncertainty at each syllable, and placing a boundary where uncertainty is high.

Probabilistic generative methods in machine learning

The defining characteristic of generative learning models is that they represent distributions of *combinations* of variables (both observed and unobserved). The conditional distribution over possible values of an unobserved variable (such as a category label), given the observed data, can then be calculated from this joint distribution using the rules of probability. As the number of variables grows, the full joint distribution becomes difficult to model directly; however, by imposing a causal structure on the set of variables, the set of dependencies among variables can be constrained, and this "curse of dimensionality" can be mitigated.

Several types of generative models are in wide use in machine learning. Many of these can be described in terms of a graph, where each variable is represented as a node, and each direct dependence is represented as an edge connecting two nodes. One particular variety of graphical model, known as a Bayes Net (Pearl, 1988), is especially useful in formalizing abductive inference, as the directional relationships between variables lend themselves to a causal interpretation. In a Bayes Net, pairs of variables are connected via directional edges (in other words, arrows). In causal terms, a connection from A to B can be thought of as a statement that A has a direct (unmediated) influence on B. We outline two frequently used Bayes Nets: the Naïve Bayes Classifier and Hidden Markov models.

The Naïve Bayes Classifier

One of the simplest forms of Bayes Net is the Naïve Bayes Classifier, which, as the "classifier" part of its name suggests, is often used to assign observations to categories. However, as a generative model, it can also be used to synthesize a new data set. The simplifying assumption behind this model is that the observable features of an object are conditionally independent given a category label: that is, once the category is known, learning about one feature is not informative about the others.

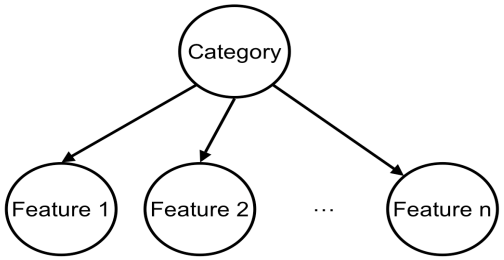

Figure 1: A Naïve Bayes Classifier. Features are assumed to be drawn from a distribution that depends only on the category and not on other features.

Hidden Markov Models

Another version of a Bayes Net is the Hidden Markov Model (HMM), used for data with a temporal or other sequential component (such as language). Here, a hidden process is assumed to evolve over time such that the state at time *t* depends only on the state at time *t-1*. Moreover, for an observable sequence produced by the hidden process, the state at time *t* is assumed to directly depend on the state of the hidden process only at the concurrent time step. Thus, given a set of transitional probabilities of evolving from one hidden state to another, along with a set of "emission probabilities" of each observable state given a particular hidden state, it is possible to estimate the

probability of any given hidden sequence, given an observed sequence. There are standard algorithms that can efficiently learn transitional and emission probabilities from labeled or unlabeled training data (the latter is an unsupervised learning problem, the standard solution to which makes use of a version of the Expectation-Maximization algorithm for probabilistic clustering). Because of their sequential nature, Hidden Markov Models are natural choices to model a variety of language phenomena. For example, a generative alternative to the discriminative "boundary-finding" approach to word segmentation discussed in the previous section is proposed by Goldwater, Griffiths and Johnson (2009). In this approach, the goal of inference is abductive: what vocabulary best explains the observed sequence of syllables?

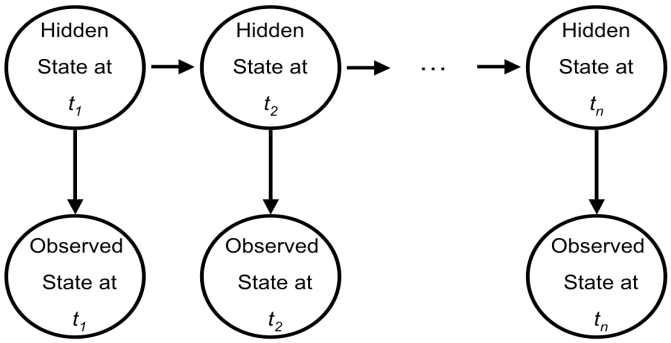

Figure 2: A Standard Hidden Markov Model. The observed state in a sequence is assumed to depend on a synchronous hidden state, which in turn depends on the previous hidden state.

Advantages of Generative Models

The defining difference between discriminative and generative models is the ability of the latter to simulate new data. In some cases this may be an end in itself. However, even if the generation of a realistic distribution of novel data is not a goal, the

ability to do so confers a number of other advantages.

By developing a “theory” about the processes that created the data, one can more naturally generalize beyond what has already been observed. Although all machine learning models are designed with some generalization in mind, for many discriminative models this generalization is limited to interpolation. In contrast, a learner who is guided by a theory, especially one that is hierarchical in nature, can make useful predictions in an entirely novel context. By drawing on information at a higher level of abstraction, one can connect the new context to one about which specific details are known. This ability to share information across contexts provides a way for a generative learner to rapidly make initial “sensible” (if rough) generalizations in a new environment from very little data, reducing its sensitivity to the specific input it encounters. In contrast, the generalizations made by discriminative models are less robust: one particular sample may result in overly specific inferences that “overfit” the data.

Another practical advantage of a generative model is its ability to detect novel situations. By representing joint distributions over variables of interest, it is possible to compute the probability of a particular set of variables taking on some observed values, even without making a commitment to particular values of the unobserved variables. In cases where this probability is low, it may be desirable to add a new context to one's theory. Even probabilistic discriminative models, on the other hand, do not represent probabilities of the *observed* data at all, and simply produce a distribution over the unobserved variables given the data, however unlikely it was. Hence, in any application where the formation of categories is involved (as is the case in almost any linguistic domain), discriminative models will, in general, need to specify the number of categories

in advance, whereas a generative learner may dynamically increase its number of clusters as the distribution of the data demands.

One of the most distinctive advantages of the generative approach is the natural emergence of explaining away behavior. In the formalism of a Bayes Net, whenever an observed node has more than one path leading to it (i.e., it has multiple "influences"), learning about the state of one of its influences affects inferences about the other. This "explaining away" behavior is quite a rational and desirable consequence in causal systems.

Disadvantages of Generative Models

The use of generative models for machine learning applications is not without costs. Although generative models are more robust to variability across data sets, and can more easily learn from small samples, for well-defined problems with well-defined solutions, discriminative methods often produce superior performance with large data sets. Since discriminative models typically place fewer restrictions on the generalizations that can be learned, in the limit of infinite data (provided they represent the input in a way that is conducive to the problem), they can often do a better job of finding the "right answer" without bias. This interplay between flexibility and robustness is known as the "bias-variance" tradeoff.

In addition to their potential bias, generative models typically require more complex computations than discriminative models. In some cases exact inferences are infeasible, with practical implementations instead relying on approximation algorithms. In speed-sensitive engineering applications, where computations are performed on serial digital computers, complexity considerations often lead to the use of

discriminative algorithms. In modeling human cognition, however, where computations are performed in massively parallel and inherently probabilistic brains, it is not clear that the same measures of computational complexity apply; moreover, the notion of an "exact solution" seems implausible for any model.

The greatest challenge associated with generative models is less a feature of the models themselves than of the model-construction process. The process of abductive inference is a form of hypothesis-testing. Although probabilistic generative models need not be restricted to a finite number of hypotheses, they do need to begin with a hypothesis space of some form. By taking advantage of the hierarchical capacity and abstraction ability of generative modeling, this space may be extremely generic and domain-general, but it must nonetheless be defined. In machine-learning applications, suitable hypothesis spaces can often be constructed from expert knowledge, but if abductive inference is to characterize human cognition, we as scientists need to say something about how hypotheses are generated by minds without invoking an infinite regress.

Summary and Future Directions

This chapter explored the possibility that human infants and children build models of their linguistic input (grammars) and converge on an appropriate model via a process of hypothesis generation and testing. Much of the current framework for thinking of learners as model builders comes from research on Bayesian inference, both as an account of human learner and as an approach to machine learning. Prompted in part by Bayesian learning models, we outlined what we take to be five characteristics

that are consistent with a model-building learner. Such a learner should be able to: (1) predict populations from samples and *vice versa*, (2) differentiate random sampling from strong sampling, (3) increasingly favor a smaller model as more data consistent with a smaller model come in, (4) change their generalizations based on a handful of input types (but not tokens) and use the currently most probable model to discount or explain away new data, and (5) seek information that differentiates between models. A growing body of developmental research in language and other areas suggest that human learners have these abilities. As we have already acknowledged, however, it seems likely that other non-model-building learners might show similar behavior in these studies. Therefore, interactions between behavioral researchers and machine learning researchers will be crucial for determining what sort of learners human infants and children really are.

Because the distinction between generative and discriminative learners is relatively well understood in computer science, appealing to machine learning to help adjudicate among different conceptions of human learners seems a reasonable strategy. As we have noted, generative learners, particularly those incorporating Bayesian probabilistic inference, have a number of advantages over discriminative learners. These include speed and robustness in making appropriate generalizations, the ability to generalize in entirely novel contexts, the ability to detect new contexts, and the ability to account for explaining away behavior. Many of the developmental studies presented in this chapter ask, in essence, whether human learners show similar advantages.

Generative models are not without their challenges, however. They are

ultimately less flexible in their generalizations than discriminative models, they require more complex computations and therefore may require us to impute greater abilities to human learners, and crucially, they require some sort of hypothesis space to be defined in advance. Indeed, our inability to say how we generate new hypotheses to explain our input, given that the hypothesis space is potentially infinite, is sufficient reason for some thinkers to say that all hypotheses must be innate (e.g., Fodor, 1981).

Let us end by providing a reason to be hopeful that we might surmount the origin of hypotheses problem. Along with the recent rise of popularity of Bayesian inference comes renewed interest in the work of philosopher Charles Sanders Peirce (1935), who describes abduction as follows:

1. The surprising fact, C, is observed.

2. But if A [some hypothesis] were true, C would be a matter of course.

3. Hence, there is reason to suspect that A is true.

In our view, the key insight embodied in Peirce's conception of model-building is that it is triggered by surprise. On this view, English-learning infants exposed to an AAB or ABA grammar instantiated in syllables are surprised by the repeated syllables, because syllable repetition is rare in English. Their surprise drives them to seek an explanation. In contrast, infants who have formed a model of Western tonal music do not find repetition to be surprising, because it is "a matter of course" in their model (Dawson, 2010; Dawson & Gerken, 2011). Younger infants who do not have such a music model *are* surprised by repetition and *do* seek an explanation (Dawson & Gerken, 2009).[4]

[4] It may be that repetition is *a priori* surprising unless it is explained away by a model in a particular domain (e.g., Gervain et al., 2008).

Other research in our lab extends the putative role of surprise beyond repetition to a domain in which what is surprising depends on the specifics of the learner's input statistics (Gerken et al., in prep). The phoneme /zh/ is very rare in English and thus potentially surprising. Infants who are exposed to AAB or ABA strings containing /zh/ appear to include the presence of /zh/ in their explanation of the input and do not generalize to new test items unless these contain /zh/. Note that this finding is in contrast to findings in which syllables starting with the more predictable /d/ phoneme do not cause infants to posit a separate explanation. These preliminary findings suggest that surprise not only tells the learner that there may be something in the input that requires explanation, but it may also drive them to differentiate their hypothesis space in ways that can accommodate the surprising aspect of the input.

Can focusing on what's surprising limit the hypothesis space enough for learning to be computationally tractable? How can we determine what is surprising? Both of these questions will need to be answered if we are to proceed with this approach. Nevertheless, the research presented in this chapter, coupled with Peirce's insight as to where hypotheses might come from, suggest that viewing human children as model builders will continue to yield fruitful insight about language development.

References

Berwick, R. C. (1986). Learning from positive-only examples: The subset principle and three case studies. In J. G. Carbonell, R. S. Michalski & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 2, pp. 625-645). Los Altos, CA: Morgan Kauffman.

Brown, R., Cazden, & Bellugi, U. (1969). The child's grammar from I to III. In J. P. Hill (Ed.), *Minnesota symposium on child psychology*. Minneapolis, MN: University of Minnesota Press.

Chomsky, N. (1959). A review of Skinner's "Verbal Behavior". *Language, 35*, 26-58.

Chomsky, N., & Lasnik, H. (1993). *Principles and parameters theory, in syntax*. Berlin: de Gruyter.

Dawson, C. (2007). *Infants learn to attend to different relations when forming generalizations in different domains*. University of Arizona, Tucson

Dawson, C. (2010). *"Explaining-away" effects in rule-learning: evidence for generative probabilistic inference in infants and adults.* University of Arizona, Tucson.

Dawson, C., & Gerken, L. A. (2009). Learning to learn differently: The emergence of domain-sensitive generalization in the second six months of life. *Cognition, 111*, 378-382.

Dawson, C., & Gerken, L. A. (2011). When global structure "explains away" evidence for local grammar: A Bayesian account of rule induction in tone sequences. *Cognition*.

Denison, S., & Xu, F. (2009). Twelve- to 14-month-old infants can predict single-event probability with large set sizes. *Developmental Science, 13*, 798-803.

Dowling, W. J. (1967). *Rhythmic fission and the perceptual organization of tone sequences.* Harvard University, Cambridge, MA.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: The MIT Press.

Elman, J. L. (2001). Connectionism and language acquisition. In M. Tomasello & E. Bates (Eds.), *Language development: The essential readings* (pp. 295–306). Oxford, U. K.: Blackwell.

Fodor, J. (1981). Representations. Cambridge, MA: MIT Press.

Gerken, L. A. (2005). What develops in language development? In R. Kail (Ed.), *Advances in Child Development and Behavior* (Vol. 33, pp. 153-192). San Diego, CA: Elsevier.

Gerken, L. A. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition, 98*, B67-B74.

Gerken, L. A. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition, 115*(2), 362-366.

Gerken, L. A., Balcomb, F. K., & Minton, J. (in press). Infants avoid “laboring in vain” by attending more to learnable than unlearnable linguistic patterns. *Developmental Science*.

Gerken, L. A., & Bollt, A. (2008). Three exemplars allow at least some linguistic generalizations: Implications for generalization mechanisms and constraints. *Language Learning and Development, 4*(3), 228-248.

Gerken, L. A., Dawson, C., Chatila, R., & Tenenbaum, J. (in preparation). What

grammars do infants consider based on a single input example?

Gerken, L. A., Wilson, R., & Lewis, W. (2005). 17-month-olds can use distributional cues to form syntactic categories. *Journal of Child Language, 32*, 249-268.

Gervain, J., Macagno, F., Cogoi, S., Peña, M., & Mehler, J. (2008). The neonate brain detects speech structure. *PNAS, 105*(37), 14222–14227.

Gold, E. M. (1967). Language identification in the limit. *Information and Control, 10*, 447–474.

Goldwater, S. and Griffiths, T. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*(5), 431-436.

Gómez, R. L., & Gerken, L. A. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70*(2), 109-135.

Hunter, M., & Ames, E. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research, 5*, 69-95.

Hunter, M., Ames, E., & Koopman, R. (1983). Effects of stimulus complexity and familiarization time on infant preferences for novel and familiar stimuli. *Developmental Psychology, 19*(3), 338-352.

Kushnir, T., & Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental Psychology, 43*(1), 186-196.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to

infer the preferences of other people. *Psychological Science, 21*(8), 1134-1140.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281-297). Berkeley, CA: Univeristy of California Press.

Marcus, G. F., Fernandes, K., & Johnson, S. P. (2007). Infant rule learning facilitated by speech. *Psychological Science, 18*(5), 387-391.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science, 283*, 77-80.

Maye, J., Werker, J. F., & Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101-B111.

McCulloch, W., & Pitts. W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 7*, 115-133.

Needham, A., Dueker, G., & Lockhead, G. (2005). Infants' formation and use of categories to segregate objects. *Cognition, 94*, 215-240.

Ortmann, O. (1926). On the melodic relativity of tones. *Psychological Monographs, 35*. 1-35.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan-Kaufmann.

Peirce, C. S. (1935). Pragmatism and Abduction. In C. Hartshorne (Ed.), *Collected Papers of Charles Sanders Peirce*, pp. 112-135. Cambridge, MA: Harvard University Press.

Pinker, S. (1994). *The language instinct*. New York, NY: William Morrow & Co.

Prince, A., & Smolensky, P. (1997). Optimality: From Neural Networks to Universal Grammar. *Science, 275*(5306), 1604-1610.

Quinn, P. C., & Bhatt, R. S. (2005). Learning perceptual organization in infancy. *Psychological Science, 16*(7), 511-515.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information-storage and organization in the brain. *Psychological Review*, *65*(6), 386-408.

Rumelhart, D., & McClelland, J. (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge, MA: MIT Press.

Rumelhart, D., & McClelland, J. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing? In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 195-248). Mahwah, NJ: Lawrence Erlbaum Associates.

Saffran, J. R., Aslin, R. N., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928.

Santelmann, L. M., & Jusczyk, P. W. (1998). Sensitivity to discontinuous dependencies in language learners: Evidence for limitations in processing space. *Cognition, 69*(2), 105-134.

Schulz, L., & Bonawitz, E. B. (2007). Serious fun: Preschoolers play more when evidence is confounded. *Developmental Psychology, 43*(4), 1045-1050.

Skinner, B. F. (1957). *Verbal Behavior*. New York: Appleton-Century-Crofts.

Teglas, E., Girotto, V., Gonzalez, M., & Bonatti, L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences 104*, 19156-19159.

Temperley, D. (2008). A probabilistic model of melody perception. *Cognitive Science: A Multidisciplinary Journal, 32*(2), 418-444.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24*, 629-640.

Valian, V., Winzemer, J., & Erreich, A. (1981). A little-linguist model of syntax learning. In S. Tavakolian (Ed.), *Language acquisition and linguistic theory* (pp 188-209). Cambridge, MA: M.I.T. Press.

Van Gael, J., Vlachos, A., & Gharamani, Z. (2009). The infinite HMM for unsupervised PoS tagging. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, 678-687.

Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition, 112*, 97-104.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences of the United States of America, 105*, 5012-5015.

Xu, F., & Tenenbaum, J. B. (2007a). Sensitivity to sampling in Bayesian word learning. *Developmental Science, 10*, 288-297.

Xu, F., & Tenenbaum, J. B. (2007b). Word learning as Bayesian inference. *Psychological Review, 114*, 245-272.