What Bayesian modelling can tell us about statistical learning: what it requires and why it works

Amy Perfors
School of Psychology
University of Adelaide

Daniel J. Navarro
School of Psychology
University of Adelaide

Abstract

This paper explores the *why* and *what* of statistical learning from a computational modelling perspective. We suggest that Bayesian techniques can be useful for understanding what kinds of learners and assumptions are necessary for successful statistical learning. The inferences that can be made by a learner are driven by both the units that such learning operates over and the levels of abstraction it includes. Other assumptions made by the learner have non-trivial affects as well, including assumptions about the process in the world generating the data, as well as whether it is more reasonable to make inferences on the basis of types, tokens, or a mixture of the two. Finally, of course, any learner must incorporate – whether explicitly or implicitly – certain assumptions in the form of their prior biases and the nature of the hypotheses they can represent and consider. We discuss the ways in which these assumptions might drive what is learned, and how Bayesian modelling can be a useful way of exploring these issues.

Introduction

The purpose of this chapter is to address some issues regarding the *why* and *what* of statistical learning, with a particular focus on Bayesian computational modelling and language acquisition. We begin with a brief introduction to Bayesian modelling, contrasting it with the other primary computational approach to statistical learning (connectionist modelling), and demonstrating how it clarifies some common confusions about what statistical learning is and is not. The chapter is structured around a series of questions: What is statistical learning? What data does statistical learning operate on? What knowledge does learner acquire from the data? What assumptions do learners make about the data? What prior knowledge does the learner possess? Finally, why does statistical learning work? Each of these is a big topic in itself, so we aim only to provide a general introduction to them, covering some but not all of the issues involved.

What is statistical learning?

Statistical learning encompasses a wide variety of learning situations in which the knowledge acquired by the learner is highly dependent on the statistical structure of the data that they are given. From an empirical perspective, researchers are interested in finding out whether and to what extent people are sensitive to statistical structure (e.g., the frequencies of different events) when learning from data. From a formal perspective, we aim to describe the abstract principles and processes that are necessary to explain how the learner might acquire knowledge based on statistical input. Statistical learning can be distinguished from learning that relies solely on deterministic rules, such as the subset principle (Berwick, 1986) or learning that requires a certain type of input before acting, like "trigger" learning (Gibson & Wexler, 1994).

Strictly speaking, *any* model that learns primarily by exploiting the statistical structure of the data is a statistical learner, but in practice computational modelling in cognitive science has focused on two particular types of statistical learners: connectionist models and Bayesian models. Connectionist modelling grew in popularity in cognitive science beginning in the 1980s, and has led to advances in our understanding of multiple areas in cognitive science, including categorisation, verb learning, and semantic representation. At its core connectionism is a learning theory are inspired by the architecture of the brain. Connectionist networks consist of a collection of nodes connected by weighted links, which propagate activation between the nodes. Computation is performed by the pattern of activation that is passed, and learning is achieved by adjusting the link weights. Bayesian modeling is a somewhat more recent approach in cognitive science, having emerged over the last decade, but is inspired by the statistical theory of probabilistic inference that dates back to the 18th century. It differs from connectionist modelling largely in terms of its assumptions about representation and learning. Since this chapter approaches statistical learning from a Bayesian perspective, we will now give a brief introduction to the Bayesian approach. Later on, we discuss the relationship between Bayesian and connectionist theories.

Bayesian theories of language and cognition draw their inspiration from probability theory, which provides a normative theory for describing how to learn from noisy data. The framework revolves around Bayes' rule, which describes how an ideal learner should update his or her beliefs in light of data, denoted d. Suppose that there are a set of hypotheses $\mathcal{H} = \{h_1, h_2, h_3, \ldots\}$ that the learner could potentially believe to be the correct theory as to the origin of the data d. This set of hypotheses is referred to as the *hypothesis space*. Before the data have been observed, the degree of belief that the learner assigns to the ith hypothesis is $P(h_i)$, the *prior probability* that h_i is the correct one. Because each of these hypotheses yields precise predictions about what data would be expected if it were true, it is possible for the learner to assess the *likelihood* $P(d \mid h_i)$, the probability of observing data d if the true hypothesis were h_i. Bayes' rule then provides a method for belief updating, in which the *posterior probability* $P(h_i \mid d)$ that the learner assigns to this hypothesis is given by:

$$P(h_i \mid d) = \frac{P(d \mid h_i)P(h_i)}{\sum_{h_j \in \mathcal{H}} P(d \mid h_j)P(h_j)} \tag{1}$$

In this expression, the numerator multiplies the prior by the likelihood: hypotheses that are more consistent with the data receive a larger multiplier than those that are not. The

denominator is just a normalizing term, to make sure that the posterior probabilities all sum to 1. Qualitatively, the key idea is that the prior beliefs $P(h_i)$ are modified by the data through the likelihood $P(d \mid h_i)$. As more data are observed, the likelihood term becomes more and more important, and so the learner will come to assign the highest posterior probability to those hypotheses that are most consistent with the data, regardless of what the prior beliefs were. A Bayesian model, therefore, is built from three distinct parts. Firstly, we need to specify the hypothesis space $\mathcal{H}$, the set of things that might be true. Secondly, we need to specify the learner's prior beliefs $P(h)$. Finally, we need to specify the likelihood function $P(d \mid h)$ that relates the hypothesis to data.

A natural question to ask is how to interpret the model. In general, Bayesian models do not try to describe any particular cognitive processes. People are not literally assumed to be computing posterior probabilities by mechanically applying Equation 1. Indeed, as is often pointed out in both the statistics literature and the cognitive science literature, these calculations can be extremely time consuming. Even given the impressive computational power that the brain provides, it is highly unlikely that Bayesian updating describes a literal mechanism for human learning. Instead, what it gives us is an abstract, ideal solution to the learning problem at hand. As such, it provides answers about what is and is not possible for the learner to learn, and provides a standard against which real learners can be assessed. As a consequence, Bayesian models are not focused on cognitive processes: they are focused on trying to understand the abstract goals of the cognitive system. What problem does it solve? How do the constraints under which it solves that problem affect what is learned? Why does the cognitive system have these goals? What would a good solution look like, and why would it be good?

Bayesian models, like any other family of models, vary in many particulars. Depending on the nature of the problem, they may incorporate different hypotheses and hypothesis spaces, different assumptions about how the data is sampled from the environment, and different assumptions about which hypotheses have the highest prior probability. One advantage of Bayesian modelling is that the assumptions of the model are made explicit. Representational constraints are specified clearly by describing the hypothesis space, preexisting beliefs are specified through the prior, and the learners interpretation of data is specified through the likelihood. As a consequence, it is comparatively easy to manipulate these these assumptions and evaluate precisely how they affect what can be learned given certain data. Indeed, much of this chapter will discuss this question.

The flip side of this advantage, however, leads to one of the main perceived disadvantages of Bayesian models: they can appear to have much more "built in" than other types of statistical learning models. This is particularly problematic for those who are interested in questions of innateness, for whom the goal is to build models that assume as little knowledge as possible *a priori*. While there are specific cases where Bayesian models rely on a lot of assumed knowledge, as a general critique of Bayesian models this problem is overstated, for two basic reasons. Firstly, drawing on work Bayesian statistics, it is quite possible to construct Bayesian models that make very few assumptions. In nonparametric Bayesian models, for instance, the hypothesis spaces are constructed so as to have "support" across the space of all probability distributions. Without going into technical details, what this means is that these models do not place strong prior constraints on what knowledge the learner can acquire. Secondly, it is important to recognise that all models must

incorporate some assumptions about what hypotheses can be represented, what hypotheses are considered more likely *a priori* and so on. As illustrated by learnability problems such as Goodman's (1955) problem of induction, Quine's (1960) indeterminacy of translation, and Gold's (1967) theorem, when faced with a complex learning problem that can have an infinite number of possible solutions, some prior biases are necessary in order to learn effectively. In this respect, the main difference between the various modeling frameworks is *not* whether they make prior assumptions: it is the extent to which these assumptions are made explicit.

Having described Bayesian models of statistical learning, we now explore how they can be used to elucidate and explore some of the issues involved in such learning, in particular the *what* and the *why*. We begin with one of the important "what" questions, in particular, how *what units* the statistical learning operates over affects the sort of inferences that can be made.

What data does statistical learning operate on?

At its most basic, statistical learning describes a process in which people acquire knowledge from probabilistic data. Within the Bayesian framework, this learning is governed by the likelihood function, $P(d \mid h)$. A natural question to ask, therefore, is what actually constitutes the data d from which people learn. This fundamental question attaches to a range of problems in the study of language. Two cases of particular interest are: (1) At what level of language (phoneme, syllable, word, etc) should we describe the input? (2) What should count as a single observation: a type or a token? We choose these because they are interesting questions in their own right, but also because each of these will serve as the motivation for one of the later sections in the chapter.

At what level should we describe the learner's data?

In order to motivate our discussion, we consider one of the most empirically robust demonstrations of human statistical learning: the fact that people are extremely sensitive to the transition probabilities that characterize the short-range sequential dependencies between linguistic units (Saffran, Aslin, & Newport, 1996; Aslin, Saffran, & Newport, 1998; Saffran & Thiessen, 2003). The natural assumption is that sensitivity to transition probabilities can help the learner solve problems like word segmentation, since the transition probability for pairs of linguistic units that cross word boundaries tends to be lower than transition probabilities for units that are contained within the same word (Harris, 1955). However, the story is more subtle than this: transition probability based statistical learning is dependent on the units over which transition probabilities are calculated and used.

In empirical work it has been typical to define transition probabilities at the level of the syllable, and to examine how these transition probabilities help people solve the word segmentation problem (e.g., Saffran et al., 1996; Aslin et al., 1998), although there are studies that look at word level transition probabilities to learn syntax (e.g., Thompson & Newport, 2007). A typical experiment involves exposing learners to input from an artificial language in which the syllable transition probability tends to be low when two syllables lie on either side of the word boundary, and high when the two syllables are wholly contained within a single word. The fact that people can learn the correct word segmentations within

the artificial language suggests that syllable-level transition probabilities are a powerful source of evidence. Natural language, unfortunately, is full of instances in which syllables cross word boundaries: Brent (1999b) gives the example of *teak rail*, whose syllabic boundaries are /ti/ and /krel/, but whose word boundaries are /tik/ and /rel/. Even if the syllable transition probability from /ti/ to /krel/ were low enough to indicate the presence of a word boundary, it does not provide the learner with the information required to infer that the location of the boundary is between /tik/ and /rel/.

As a consequence of this ambiguity, most computational models of word segmentation use input in which phonemes are the basic unit, not syllables. However, computational models allow a finer grained investigation of the issue: by comparing the performance of models that rely on different assumptions, we can obtain more insights into the level of abstraction at which transition probabilities are calculated. For instance, it might be possible for the learner to calculate transition probabilities at the syllable level only, and try to memorize the ambiguous cases. Alternatively, a model might operate solely at the phonetic level, requiring no syllable-level calculations. Finally, the learner might try something in between, in which syllable-level transition probabilities do the vast majority of the work, but the learner relies on phoneme-level transition probabilities to handle ambiguous cases. In short, comparing the performance of different computational models can yield further insight about what units transitional probabilities must be calculated over in order for the successful segmentation of words from fluent speech.

Learning from types or from tokens?

The second sense in which the units of analysis matters is whether we learn on the basis of types or tokens. Although there is extensive psycholinguistic and developmental evidence that people are sensitive to token-level frequency variation in a variety of contexts, there are also reasons to think that for at least some kinds of inference, it may be more sensible to rely on type frequencies rather than token frequencies (or an interpolation between the two). For instance, one of the most robust empirical regularities in language is the power-law distribution in the frequency of word tokens: a few words are extraordinarily common, and many are extremely infrequent (Zipf, 1932). These power-law distributions appear to apply to many other elements of language, not just words (see Briscoe, 2006, for an overview). However, many standard statistical models fail to capture this distribution; for instance, context-free grammars capture an exponential rather than power-law distribution. Recent work addresses this lack using a two-stage Bayesian model for language learning called an *adaptor grammar*, which separates the question of how forms are generated from the question of how frequent those forms are (Goldwater, Griffiths, & Johnson, 2006b; Johnson, Griffiths, & Goldwater, 2007). The framework corresponds to assuming that language users can generate tokens either by drawing on a memory store of familiar types, or by generating a type anew based on deeper linguistic principles. The model can simultaneously infer which underlying linguistic generalisation is correct, as well as whether it is more sensible to perform inferences on the basis of types, tokens, or a mixture of the two.

This model has been used for unsupervised acquisition of the morphology of English (e.g., learning that the word *helped* can be parsed into the stem *help* and the separate past-tense suffix *ed*). This sort of knowledge is useful to children for making productive generalisations of novel verbs. Adaptor grammars have also been applied in the problem of

grammar induction. For instance, work by Perfors, Tenenbaum, and Regier (2011) suggests that the nature of the abstract grammatical inferences a learner is justified in making can depend on whether they assume that learning should be done over sentence types or tokens. A learner who assumes that grammar induction should be done on the basis of types will infer that grammars with hierarchical phrase structure are the best fit to child-directed speech; a learner who assumes that it should be done on the basis of tokens will prefer grammars without hierarchical phrase structure. Which of these learners is most sensible? We can address this question by evaluating what a learner capable of interpolating between types and tokens – determining which interpretation of the data will lead to the highest overall probability – would conclude, as in (Johnson et al., 2007). Results indicate that such a learner would probably[1] infer that a more type-based analysis is more appropriate, and that, overall, the grammars with the highest probability are those with hierarchical phrase structure. The point here is that the inferences possible from types can be different from the inferences possible from tokens, and that realising which is most appropriate for a given problem is an important task facing the learner. Research into this question is still in its infancy, so many open questions remain.

What knowledge does the learner acquire from the data?

Uncovering the form and content of the mental representations that the learner relies upon is a central question in any discussion of language acquisition and cognitive science more generally. Are the rules of syntax best described in terms of regular grammars, context free grammars, context sensitive grammars, or something else? Are phonetic categories represented in the same way as other perceptual categories? If so, are these categories represented in terms of prototypes, exemplars, distribution estimates, decision boundaries, or something else? These questions and many others all form part of the broader issue of describing the knowledge acquired by the learner. The scope of this issue is thus too wide to cover in detail. With this in mind, we restrict ourselves to two topics. Firstly, we follow on from our previous discussion of sensitivity to transition probabilities in word segmentation, and try to show the range of mental representations that can be explored by computational models. We then pick a single issue that these models open up (learning on multiple levels), and follow it through in some detail.

Modelling the word segmentation process

In our earlier discussion of transition probabilities, the focus was primarily on the data: when addressing the word segmentation problem, is it sensible to assume that the learner computes transition probabilities between phonemes, between syllables, or something more complex. However, while transition probabilities are presumably useful to the learner when trying to solve the word segmentation problem, it is clearly the case that the actual knowledge acquired by the learner is much richer. To see this, suppose that the learner has successfully calculated all transitional probabilities (be they phoneme transition probabilities or syllable transition probabilities). How should he or she segment the speech

[1]Because the search problem in this sort of grammatical learning is so difficult, the results are based on approximations, rather than an exhaustive search of the entire space of grammars and type-to-token-based interpolations.

into words on the basis of this knowledge? Is it sufficient for the learner to infer word boundaries whenever the transition probability is under some threshold? If so, how is that threshold set? Alternatively, do the statistics of the rest of the language play some role in determining the word boundaries? If so, how does this work?

These are the kinds of questions that computational models can help address. In doing so, they open up new and interesting questions. For example, few models directly apply a simple threshold to learn word segmentations. Instead, transition probability is assumed to be one cue available to the learner, who seeks to extract some explicit theory about the lexicon. In some cases, the model aims to learn the word boundaries directly (e.g., Elman, 1990; Cairns, Shillcock, & Chater, 1997; Christiansen, Allen, & Seidenberg, 1998), while other models focus on learning the lexicon itself, identifying the boundaries as a side effect (de Marcken, 1995; Brent & Cartwright, 1996; Perruchet & Vinter, 1998; Brent, 1999a; Venkataraman, 2001; Swingley, 2005; Goldwater, Griffiths, & Johnson, 2006a, 2009). The motivation for the second group of models is that word segmentation is not the primary goal of the learner: rather, the main goal is to identify the words themselves. As such, it is argued the better approach is to build word segmentation models that are focused on learning the lexicon itself. Thus we have three different ideas about the nature of the knowledge that the learner acquires during the word segmentation process: raw transition probabilities, word segmentation data, lexical knowledge. These three ideas differ in terms of the extent to which the learner is assumed to abstract away from the raw data. Calculating transition probabilities involves very little abstraction, assigning word boundaries involve only the minimal amount of abstraction required to solve the segmentation problem, and lexical knowledge involves moving beyond the basic problem and focusing on the broader problems facing the language learner. Evaluating all three of these possibilities is important for understanding what kind of knowledge the learner acquires, and what kind of data is required to acquire it.

Recent Bayesian models of word segmentation (Brent, 1999a; Goldwater et al., 2009), which tend to outperform other models on naturalistic corpora, have tended to be of the third type (focusing on lexical knowledge). They also more accurately capture human performance on artificial languages in the lab (Frank, Goldwater, Griffiths, & Tenenbaum, 2007). The interesting thing about these models is the extent to which they highlight how complex the learning problem is. Within these models there is a tension between the desire to have a simple lexicon, and the desire to assign high probability to the observed corpus. One way to maximise the probability of the data is to assume that the lexicon contains only a single "word" with that word being precisely identical to the entire corpus. At the other extreme, the simplest possible "lexicon" consists of a small number of words, one per phoneme, with word boundaries placed between all pairs of phonemes. In between these two extremes exists an optimal compromise solution, with a moderately large lexicon consisting of fairly short words, which assigns fairly high probability to the data. The most recent of these models (Goldwater et al., 2009) goes further than this, and shows that better word segmentation can be achieved if the learner calculates the transitional probabilities not just between the observed phonemes, but *also* between the words in the (learned) lexicon. As it turns out, if the learner ignores the transitions probabilities between words (e.g., by assuming words are independent) then the result is systematic under-segmentation; learning based on within-word as well as between-word dependencies greatly improves the

segmentation of naturalistic child-directed speech. In other words, by constructing detailed computational models that can solve the statistical learning problem, we learn that the most successful representations are likely to be complex, and involve learning an explicit lexicon and tracking transition probabilities at *multiple* levels of linguistic analysis.

Learning on multiple levels

In the previous section, we illustrated how computational models help "flesh out" the ideas associated with the statistical learning of language, using the word segmentation problem as an example. In this section we now pick one of the issues raised in that discussion, and follow it through in some detail, across multiple problems. That issue is the question of learning on multiple levels. In the previous example, we saw that word segmentation is improved when the learner has the capacity to use knowledge from two different levels (in this case, phonemes and words) to assist them. However, the issue is quite general, and turns up in a range of problems in language and cognition. One especially interesting case is when the two "levels" refer to levels of abstraction: learning both information about specific items and information about general principles. For instance, acquiring categories involves learning information about how specific categories are organised (e.g., that balls tend to be round) as well as information about how categories *in general* tend to be organised (e.g., that count nouns are organised by shape). Empirical work demonstrates that children learn both of these elements (Landau, Smith, & Jones, 1988; Imai & Gentner, 1997; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002), and that this learning is probably driven by the statistical nature of the input (Samuelson & Smith, 1999).

An analogous problem arises in a very different domain – acquiring verb argument constructions. In every language, different verbs take arguments in distinct constructions; for instance, the verb *load* can occur with two distinct locative constructions ("He loaded apples into the cart" and "He loaded the cart with apples"). Not all verbs can occur in all constructions: one can pour apples into a cart but not pour a cart with apples, and one can fill a cart with apples but not fill apples into a cart. Knowing which verbs can occur with which constructions is verb-specific knowledge, but children learning language also acquire verb-general knowledge about the sorts of constructions that verbs of different types can appear with (see, e.g., Baker, 1979; Pinker, 1989). This allows them to generalise sensibly about novel verbs, spontaneously producing sentences such as "He is mooping the cloth with marbles" when introduced to the novel verb 'mooping' in the context of an experimenter placing marbles into a cloth (Gropen, Pinker, Hollander, & Goldberg, 1991). Simultaneous verb-general and verb-specific learning on the basis of the statistical structure of the input has been demonstrated in the lab as well, in artificial language learning studies (Wonnacott, Newport, & Tanenhaus, 2008).

The fact that human learners are able to learn knowledge organized at multiple levels of abstraction raises a natural question: how is it possible to describe this learning, and how do we describe the structure of the knowledge that they acquire? This kind of learning has been shown to be possible for both connectionist (e.g., Kruschke, 1992; Colunga & Smith, 2005) and Bayesian (Navarro, 2006; Kemp, Perfors, & Tenenbaum, 2007; Griffiths, Sanborn, Canini, & Navarro, 2008; Heller, Sanborn, & Chater, 2009) learners, and some of these models have been applied to verb construction learning (Hsu & Griffiths, 2009; Perfors, Tenenbaum, & Wonnacott, 2010). The Bayesian framework in particular is quite revealing

about the general computational principles that allow this learning to occur and the kind of knowledge that is acquired. The one thing that all of the Bayesian models discussed above have in common is that they are *hierarchical* models. In a standard Bayesian model the learner is assumed to postulate hypotheses that explain the observed data. Each individual learning problem (e.g., learning a single category) maps onto a single hypothesis. In a hierarchical model, however, the learner goes one step further and postulates more general hypotheses that can explain each of the individual ones. Following Goodman (1955), these "hypotheses about hypotheses" are called overhypotheses. Mathematically, what this means is that instead of having a single set of prior beliefs $P(h_i)$ (as per Equation 1), the learner's belief about a specific hypothesis h_i are constrained by an overhypothesis. Thus, if the learner has the overhypothesis o_k, his or her beliefs are described by the more structured distribution, $P(h_i|o_k)P(o_k)$. What this means is that as the learner refines their beliefs about specific hypotheses, they also acquire more general knowledge in the form of the overhypothesis. In principle, the level at which learning occurs could be extended upward even more, until the knowledge at the highest level is weak or general enough that it can be plausibly assumed to be innate.

As in the problem of word segmentation, Bayesian models demonstrate what can be learned by a learner capable of statistical inference on multiple levels at once. Do people *actually* learn in this way? This is an empirical question that requires much more work to flesh out fully, but some indications are promising. For instance, it is empirically observed that children are capable of acquiring higher-order word learning generalisations after learning relatively few words (Smith et al., 2002); this sort of rapid inference is one of the trademarks of Bayesian models, but is rarely observed in connectionist models. A recent model captures this learning (Kemp et al., 2007), and the same model captures human verb learning in artificial language learning tasks (Perfors et al., 2010). A related version of the model predicts that both higher-level and lower-level generalisations in category learning should be acquired at the same time (i.e., on the basis of the same amount of data). This prediction which was empirically supported in an experiment with adults (Perfors & Tenenbaum, 2009). These models also demonstrate how it may be possible to acquire two different higher-order generalisations about two different kinds of things at the same time (for instance, learning that solids are organised by shape but non-solids are organised by texture). Although there is evidence that young children are capable of this sort of learning over the course of their first years, it is an open question whether it is possible for adults in a laboratory setting.

What assumptions do learners make about the data?

In the previous section we discussed the kind of knowledge that a statistical learner is able to acquire. Closely related to this issue is the question about what the learner assumes about the data from which such knowledge is acquired. This issue was hinted at in our earlier discussion of the type-token distinction. As noted previously, inferences drawn from type-level data can be different from those drawn from token-level data. As such, if the learner assumes that the "power law" statistics observed at the token level are caused by extraneous processes (i.e., the adaptor in adaptor-grammars) then the inferences will be different than if the learner treats these as an important characteristic of the data (i.e., if the token-level data are sampled independently). In other words, in order to successfully

extract linguistic knowledge from data, the learner must also make assumptions about the manner in which those data were generated. These are referred to as *sampling assumptions*. In this section we discuss how sampling assumptions can play a role in influencing what a statistical learner learns, and the rate at which this knowledge is acquired.

Much of the literature on sampling assumptions focuses on two extreme but illustrative cases, known as "strong sampling" and "weak sampling". Strong sampling corresponds to assuming that data is generated by the underlying process that the learner seeks to acquire: for instance, sentences in a language are probably strongly sampled from some sort of underlying grammar, since the grammar is used to generate the sentences. Conversely, assuming weak sampling corresponds to assuming that the data can be generated independently of the process being learned about, and that process serves simply to label the data. For instance, the typical paradigm in a category-learning experiment corresponds to a weak sampling process: items are presented to the learner, and labelled as being either examples of the concept or *not* examples of the concept. One could assume weak sampling even if one only sees positive examples, of course, if one assumes that items are generated independently but are labelled by the underlying process, and that for some reason the negative items are never seen.

Because Bayesian models force an explicit specification of the learner's assumptions about the nature of the generative process, these models can clarify how sampling assumptions change the nature of the inferences that can be made. If a learner assumes strong sampling, then it is possible to learn a surprising amount from just a few data points. This is because the data points tell you something about the "size" of the concept. Broadly speaking, if there are h items in the concept – for instance, h animals in the immediate world that correspond to the label "dog", or h sentences that a grammar can generate – then the probability of generating any *specific* item is proportional to $\frac{1}{h}$.[2] As a result, if there are n items generated independently, the probability of all n of them will be proportional to $\frac{1}{h^n}$. In other words, after very few data points, the highest-probability hypotheses will be those that are most conservative with respect to those data points. This principle is known as the "size principle" (Tenenbaum & Griffiths, 2001; Navarro & Perfors, 2010).

As a result of the size principle, it is possible to make strong inferences on the basis of relatively little data. Work by Xu and Tenenbaum (2007) demonstrates this type of inference in a word learning context. Intuitively, if we were shown one object – say, a dalmatian – and told that it were an example of a "fep", we would not necessarily infer that "fep" means dalmation; it could mean dog or animal or pet. Yet, if we were shown *three* examples of a "fep" and they were all dalmatians, we would think it much more likely that "fep" meant dalmation. This is because dalmation is a "smaller" concept than dog: dogs include all of the same items as dalmations do, plus many others. If the underlying concept actually *were* dog, it would be somewhat surprising that only dalmations happened to be generated, but this would not be a puzzle if the underlying concept were dalmation.

[2]It is only *precisely* $\frac{1}{h}$ in the case that items are sampled with equal probability from the underlying process, and completely independently from each other. If that changes – as, for instance, might occur in the case of grammars, because some sentences might be more probable than others by virtue of being shorter or using more frequent words or constructions – then the precise probability of any one item might diverge slightly from $\frac{1}{h}$. Still, the probability will scale proportionally to $\frac{1}{h}$, which is all that is necessary to drive the effect being discussed.

Xu and Tenenbaum (2007) found that both adults and children reason according to this intuition in a word-learning task like this. Even infants appear to make inferences that are consistent with a strong sampling assumption: after observing an experimenter blindly draw four red balls and one white ball out of a box, infants tend to assume that the box is predominately red and are surprised if it turns out to be predominantly white (Xu & Garcia, 2008). Intriguingly, this effect vanishes when the infant is presented with evidence that the balls are being sampled in a different way. If they see the experimenter look into the box and selectively remove four red balls after evincing a preference for red balls, they are no longer surprised to find that the box is predominantly full of white ones: the sample wasn't drawn independently and at random from the underlying concept (Xu & Denison, 2009).

Strong sampling assumptions may also enable a learner to overcome the problem of learning in the absence of negative evidence. If items are strongly sampled from the concept, there is never any negative evidence given, but the learner can nevertheless eventually constrain their generalisation in a sensible way. This is because the size principle captures the notion of a suspicious coincidence: as the number of examples increases, hypotheses that make specific predictions – those with more explanatory power – tend to be favored over those that are more vague. As the size of the data set approaches infinity, a Bayesian learner rejects larger or more overgeneral hypotheses in favor of more precise ones. With limited amounts of data, the Bayesian approach can make more subtle predictions, as the graded size-based likelihood trades off against the preference for simplicity in the prior. This is essentially the same idea as implicit negative evidence, which others have suggested (e.g., Braine & Brooks, 1995); the mathematics of Bayesian probability theory provides a principled quantitative justification for *why* a sensible learner should take it into account, and precisely how much to do so.

Do learners always assume that data in the world is a result of strong sampling? This has not yet been studied extensively, but early work suggests both that there are individual differences, and that assumptions appear to depend somewhat on the nature of the task and domain. In some recent work, people were given different cover stories explaining how different sorts of categorical data were created: for instance, in one, participants were told that they had observed bacteria at certain temperatures and then were asked what range of temperatures such bacteria could survive (Navarro, Lee, Dry, & Schultz, 2008; Navarro, Dry, & Lee, submitted). Although there were individual differences in how people performed, many inferences suggested that people were assuming some mixture between strong and weak sampling: their generalisations sharpened somewhat with increasing data, but not as much as the size principle would predict. Interestingly, as long as a learner assumes *anything* other than purely weak sampling – that is, as long as they sharpen their generalisations with increasing data at all – it is still possible to constrain generalisations even without negative evidence. As before, the mathematics of probability theory can explain precisely how much generalisations should sharpen as a function of the degree to which the learner assumes strong sampling.

What prior knowledge does a statistical learner possess?

Assumptions about the sampling process is tied to another set of assumptions that can drive the shape and nature of a learner's generalisations, captured in Bayesian models

via the prior. In the model of Navarro et al. (submitted) above, people's assumptions about sampling were included as a part of the model parameters. This isn't meant to imply that such parameters can be completely arbitrary: for instance, consider the priors. Although most Bayesian models have enough flexibility to accommodate extremely strange priors if the modeller wants to, in practice a preference for simpler or more parsimonious hypotheses will emerge naturally without having to be deliberately engineered. This preference derives from the generative assumptions underlying the Bayesian framework, in which hypotheses are themselves generated by a process that produces a space of candidate hypotheses and the prior probability $P(h)$ reflects the probability of generating h under that process.

As an example, consider the grammar-learning work of Perfors et al. (2011) referred to earlier. It assumes that grammars are generated by an underlying grammar-generating process in which individual grammar rules are created by generating non-terminal and terminal nodes from an underlying vocabulary according to certain specifications (e.g., for a context-free grammar rule, the left-hand-side must always be a non-terminal, and so forth). For each choice that is made, there is some probability of choosing differently: for instance, one might have chosen *NP* to go in a particular "slot" in some rule, but it could equally have been a *VP* or *PP*. As a result, grammars that are longer and more complicated – that have more rules, more non-terminals, and more ways of creating legitimate rules – will be disfavoured in the prior; the more choices a hypothesis requires, the more likely it is that those choices could have been made in a different way, resulting in an entirely different hypothesis. More formally, because the prior probability of a hypothesis is the product of the probabilities for all choices needed to generate it, and the probability of making any of these choices in a particular way must be less than one, a hypothesis specified by strictly more choices will in general receive strictly lower prior probability.

Although we have illustrated this by the grammar example, this is a general property of Bayesian modelling, and most priors (unless deliberately engineered otherwise) will naturally favour more parsimonious hypotheses with fewer parameters. The word segmentation model of Goldwater et al. (2009), for instance, favours segmentations that reflect a smaller underlying vocabulary of words, and category-learning models (e.g., Perfors & Tenenbaum, 2009) generally favour fewer categories. None of these models blindly prefers the simplest of all possible explanations, of course, because Bayesian probability theory trades off the preference for simplicity (in the prior) with goodness-of-fit to the data (in the likelihood). Since likelihood is weighted more and more heavily as the amount of data increases, there can be interesting differences in what is learned as the amount of data increases. With little data, preferred hypotheses tend to be simpler, but as it increases, the complexity of the preferred hypothesis can also increase.

Bayesian modelling makes explicit how different learning assumptions lead to a different kind of learning in another way, too. All Bayesian models explicitly specify the nature of the hypotheses that the model entertains. It is rare for all hypotheses to be explicitly enumerated and compared, since most hypothesis spaces are simply too large, if not infinite in size; in practice, the best hypothesis (the one with maximum posterior probability) is identified through intelligent search of the space. Models with different hypothesis spaces can therefore learn different things. For instance, a learner could not learn that context-free grammars are the best fit to child-directed speech if they were not capable of representing and evaluating context-free grammars in the first place; a learner could not learn about

an underlying category if it was incapable of representing the features to that category. This explicit representation of the hypothesis space can appear as a disadvantage to many, especially by comparison to connectionist models, which appear to build in less. However, as discussed earlier, *all* models must build in *something*, and connectionist models implicitly build in hypothesis spaces through their choice of architecture and other parameters; learning in connectionist models corresponds to searching over some of the hypotheses in the space. The main difference between the models is not, therefore, that Bayesian models necessarily build in a lot more innate machinery in terms of their hypothesis spaces than connectionist models do: it is that the *nature* of these hypothesis spaces is made explicit, rather than implicit, and it is that for connectionist models the focus is on the process of the search, whereas in Bayesian models the focus is on the nature of the solution.

This leads to our final section, which addresses the issue of *why?* Why does statistical learning work in the first place, what does it mean to "work", and why is it beneficial to study statistical learning from a modelling perspective?

When and why does statistical learning work?

In a very real sense, Bayesian theories are genuinely rational theories of inference. What we mean by this is that Bayesian probability theory is not simply a set of *ad hoc* rules useful for manipulating and evaluating statistical information: it represents a set of unique, consistent rules for conducting plausible inference (Jaynes, 2003). In essence, it is a extension of deductive logic to the case where propositions have degrees of truth or falsity (and indeed, deductive inference is a special case of Bayesian inference). Just as formal logic describes a deductively correct way of thinking, Bayesian probability theory describes an inductively correct way of thinking. As Laplace (1816) said, "probability theory is nothing but common sense reduced to calculation." To illustrate what this means, suppose we were to try to come up with a set of desiderata that a system of "proper reasoning" should meet. This might include things like consistency, and qualitative correspondence with common sense. For instance, if you see some data supporting a new proposition A, you should conclude that A is more plausible rather than less; the more you think A is true, the less you should think it is false; if a conclusion can be reasoned multiple ways, its probability should be the same regardless of how you got there; etc. Probability theory and Bayesian inference can be seen to follow from a mathematical formalization of these basic desiderata (Cox, 1946, 1961). In that sense, Bayesian inference follows from basic common sense, and supplies a sensible normative standard for inductive inference. Moreover, there are a number of results that show that Bayesian models have optimal or near-optimal predictive capabilities (see de Finetti, 1937; Dawid, 1984; Grünwald, 2007): a non-Bayesian reasoner attempting to predict the future will typically make worse predictions than a Bayesian one.

On the surface, then, it would appear to be the case that Bayesian reasoning solves all of the problems of inductive inference for us, since it is a normative model for induction that follows from qualitative common sense reasoning. However, the story is somewhat more subtle than this. Earlier in this chapter we briefly referred to work by Goodman (1955), Quine (1960) and Gold (1967), all of which implied that when the learning problem is "complex" in some sense, it is impossible for *any* variety of statistical inference to work effectively. In the statistics literature, for instance, it is is well-established that if the

world lacks any structure or the learner is simply insensitive to such structure, then no learning is possible (Schaffer, 1994; Rao, Gordon, & Spears, 1995). We mentioned this at the time in order to make the point that both connectionist and Bayesian models require some assumptions in order to work. In many ways, this point is unremarkable. There simply has to be some minimal set of *a priori* assumptions required for statistical learning to work, and both Bayesian and connectionist models must rely on such assumptions. From the connectionist perspective, what this means is that any suggestion that connectionist models do not "build things in" must be viewed with skepticism. Whether it be through the architecture, the learning rule, or the activation mechanisms, a connectionist model must supply constraints and assumptions: it cannot be otherwise. From a Bayesian perspective, one must be cautious about overly strong claims about "optimality". For instance, if the learner lives in a world in which "tomorrow is like today", a Bayesian model that assumes that that "tomorrow will be the opposite of today" will perform extremely poorly. Moreover, there are even cases where it is possible to make Bayesian models behave suboptimally (Grünwald & Langford, 2007) or pathologically (Diaconis & Freedman, 1986); there are also clever ways of speeding up Bayesian learning beyond what the standard analysis might predict (Van Erven, Grünwald, & de Rooij, 2009). In other words, Bayesian modelling is not a panacea, nor an excuse to avoid thinking carefully about assumptions. To the extent that a Bayesian model makes poor assumptions, it can perform just as poorly as any other type of model. In large part, this is the reason why we compare different Bayesian models against each other: to find out which ones make more sensible assumptions about the world.

Returning to the cognitive science questions, it is an open question of if (and to what extent) human learners actually do behave in accordance with the optimal predictions of Bayesian theory. In some domains, it may certainly appear that they do not: for instance, it has long been noted that human decision making is rife with biases that appear to diverge markedly from what Bayesian reasoning would predict (e.g., Tversky & Kahneman, 1974). Nevertheless, in other areas, Bayesian models appear to make surprisingly strong fits to human behaviour, some of which we have seen already, some of which applies in areas other than language acquisition (e.g., causal reasoning (Griffiths & Tenenbaum, 2009), sensorimotor control (Kording & Wolpert, 2006), and vision (Yuille & Kersten, 2006), among others). Even when humans are non-optimal, it is impossible to know this without being able to precisely specify what optimal thinking would amount to. Put another way, understanding how humans *do* think is often made easier if one can identify the ways in which people depart from the ideal: this is approximately the methodology by which Kahneman and Tversky derived many of their famous heuristics and biases.

Moreover, specifying an optimal (or at least near-optimal) solution to the problem often makes it clear how it might be the case that people could *still* be reasoning sensibly, but be operating under additional constraints emerging from the nature of the underlying representation, or from having limited memory or processing available (e.g., Chase, Hertwig, & Gigerenzer, 1998). It is possible to capture these sort of constraints within the Bayesian framework using a relatively recent approach called rational process modelling (Sanborn, Griffiths, & Navarro, 2010), which focuses more on the process by which the hypothesis space is searched, and provides for a way to evaluate who a learner with limited capacity might approximate the optimal solution. Although rational process models have not yet been applied much in language (though see Pearl, Goldwater, & Steyvers, 2010, for an

example where it has), they have been applied with some success in areas like decision making (Vul, Goodman, Griffiths, & Tenenbaum, 2009) and category learning (Sanborn et al., 2010), and are a promising future direction for those interested in exploring the extent to which humans approximate optimal inference.

There is one final way in which it is interesting and useful for those interested in language acquisition to be able to precisely specify and understand what optimal reasoning would look like: it is useful for performing ideal learnability analysis. What must be "built into" the newborn mind in order to explain how infants eventually grow to be adult reasoners, with adult knowledge? One way to address this question is to establish the bounds of the possible: if some knowledge couldn't possibly be learned by an optimal learner presented with the type of data children receive, it is probably safe to conclude either that actual children couldn't learn it, either, or that some of the assumptions underlying the model are inaccurate. The tools of Bayesian inference are well-matched to this sort of problem, both because they force modelers to make all of these assumptions explicit, and also because of their representational flexibility and ability to calculate optimal inference.

Conclusion

We have explored here a little bit about the *why* and *what* of statistical learning, using the Bayesian framework to illuminate all of these questions. We demonstrated that Bayesian techniques can be useful for understanding what kinds of learners and assumptions are necessary for successful statistical learning: the units that such learning operates over and the levels of abstraction it includes both drive the nature of the inferences that can be made. Learners, too, must make assumptions about whether it is more reasonable to make inferences on the basis of types or tokens (or a mixture of the two) and whether the data was strongly or weakly sampled (or a mixture of the two). And, of course, any learner must incorporate – whether explicitly or implicitly – certain assumptions in the form of their prior biases and the nature of the hypotheses they can represent and consider. We discussed how these assumptions might drive what is learned, and how the Bayesian paradigm, since it forces them to be made explicit, can make it relatively straightforward to manipulate them and evaluate how that changes what can be learned.

Of course, Bayesian modelling is only one tool in the toolbox available to researchers studying language acquisition. Although it can be very useful to be able to specify and understand the nature of the problem facing language learners as well as what an optimal solution would entail, this is only part of the scientific problem. In our view, maximal scientific progress can be made with a combination of computational modelling (of all sorts) in conjunction with rich and precise empirical work exploring what people actually *do* learn, and clarifying and testing the assumptions and predictions made by the models. Only by working in tandem can we most efficiently arrive at a full understanding of how and why people learn from the statistical regularities in their environment.

References

Aslin, R., Saffran, J., & Newport, E. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.

Baker, C. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, *10*(4), 533–581.

Berwick, R. (1986). Learning from positive-only examples: The subset principle and three case studies. *Machine Learning*, *2*, 625–645.

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

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

Brent, M. (1999b). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences*, *3*(8), 294–301.

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*, 93–125.

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

Cairns, P., Shillcock, R., & Chater, N. (1997). Bootstrapping word boundaries: A bottom-up approach to speech segmentation. *Cognitive Psychology*, *33*, 111–153.

Chase, V., Hertwig, R., & Gigerenzer, G. (1998). Visions of rationality. *Trends in Cognitive Sciences*, *2*(6).

Christiansen, M., Allen, J., & Seidenberg, M. (1998). Learning to segment speech using multiple cues. *Language and Cognitive Processes*, *13*, 221–268.

Colunga, E., & Smith, L. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review*, *112*(2), 347–382.

Cox, R. (1946). Probability, frequency, and reasonable expectation. *American Journal of Physics*, *14*, 1–13.

Cox, R. (1961). *The algebra of productive inference.* Baltimore, MD: Johns Hopkins University Press.

Dawid, A. P. (1984). Present position and potential developments: Some personal views, statistical theory, the prequential approach. *Journal of the Royal Statistical Society, Series A*, *147*, 278-292.

de Finetti, B. (1937). Prevision, its logical laws, its subjective sources. In H. Kyburg & H. Smokler (Eds.), *In studies in subjective probability* (2nd ed.). New York: J. Wiley and Sons.

de Marcken, C. (1995). *The unsupervised acquisition of a lexicon from continuous speech* (A.I. Memo No. 1558). Cambridge, MA: Massachusetts Institute of Technology.

Diaconis, P., & Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics*, *14*, 1-26.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2007). Modeling human performance in statistical word segmentation. In D. McNamara & J. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (p. 281-286). Austin, TX: Cognitive Science Society.

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, *25*(3), 407–454.

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

Goldwater, S., Griffiths, T. L., & Johnson, M. (2006a). Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics* (p. 673-680). Sydney, Australia.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2006b). Interpolating between types and tokens by

estimating power law generators. In Y. Weiss, B. Scholkopf, & J. Platt (Eds.), *Advances in Neural Information Processing Systems* (Vol. 18, p. 459-466). Cambridge, MA: MIT Press.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21–54.

Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Griffiths, T. L., Sanborn, A., Canini, K., & Navarro, D. (2008). Categorization as non-parametric Bayesian density estimation. In M. Oaksford & N. Chater (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (p. 303-328). Oxford: Oxford University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*, 661–716.

Gropen, J., Pinker, S., Hollander, M., & Goldberg, A. (1991). Affectedness and direct objects: The role of lexical semantics in the acquisition of verb argument structure. *Cognition*, *41*, 153–195.

Grünwald, P. (2007). *The minimum description length principle*. Cambridge, MA: MIT Press.

Grünwald, P., & Langford, J. (2007). Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, *66*, 119-149.

Harris, Z. (1955). From phoneme to morpheme. *Language*, *31*, 190–222.

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

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

Imai, M., & Gentner, D. (1997). A cross-linguistic study of early word meaning: Universal ontology and linguistic influence. *Cognition*, *62*(169–200).

Jaynes, E. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

Johnson, M., Griffiths, T. L., & Goldwater, S. (2007). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in Neural Information Processing Systems* (Vol. 19). Cambridge, MA: MIT Press.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Kording, K., & Wolpert, D. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*(7), 319–326.

Kruschke, J. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

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

Laplace, P. S. (1816). *A philosophical essay on probabilities*. Dover Publications.

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

Navarro, D., Dry, M., & Lee, M. (submitted). Sampling assumptions in inductive generalization.

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

Navarro, D., & Perfors, A. (2010). Similarity, feature discovery, and the size principle. *Acta Psychologica*, *133*, 256–268.

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

Perfors, A., & Tenenbaum, J. B. (2009). Learning to learn categories. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 136–141). Austin, TX: Cognitive Science Society.

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

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

Perruchet, P., & Vinter, A. (1998). Parser: A model for word segmentation. *Journal of Memory and Language*, *39*, 246–263.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure.* Cambridge, MA: MIT Press.

Quine, W. v. O. (1960). *Word and object.* Cambridge, MA: MIT Press.

Rao, R. B., Gordon, D., & Spears, W. (1995). For every generalization action, is there really an equal and opposite reaction? analysis of the conservation law for generalization performance. In *Proceedings of the 12th International Conference on Machine Learning* (p. 471-479).

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

Saffran, J., & Thiessen, E. (2003). Pattern induction by infant language learners. *Developmental Psychology*, *39*, 484–494.

Samuelson, L., & Smith, L. (1999). Early noun vocabularies: Do ontology, category structure, and syntax correspond? *Cognition*, *73*, 1–33.

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

Schaffer, C. (1994). A conservation law for generalization performance. In *Proceedings of the 11th International Conference on Machine Learning* (p. 259-265).

Smith, L., Jones, S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

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

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(2), 629–641.

Thompson, S., & Newport, E. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*, 1–42.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *135*, 1124–1131.

Van Erven, T., Grünwald, P., & de Rooij, S. (2009). Catching up faster in Bayesian model selection and model averaging. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Information Processing Systems* (Vol. 20, p. 417-424).

Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech. *Computational Linguistics*, *27*(3), 351–372.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2009). One and done? Optimal decisions from very few samples. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (p. 148-153).

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

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

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012–5015.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.

Zipf, G. (1932). *Selective studies and the principle of relative frequency in language.* Cambridge, MA: Harvard University Press.