Appears in *Histoire, Épistémologie, Langage* XXXI(1): 127-144, 2009.

Edward P. Stabler

Mathematics of language learning

Abstract. This paper surveys prominent mathematical approaches to language learning, with an emphasis on the common fundamental assumptions of various approaches. All approaches adopt some restrictive assumption about the nature of relevant causal influences, with much ongoing work directed to the problem of discovery and justification of these assumptions.

There has been a long-standing debate among philosophers, psychologists and linguists about what knowledge, or what biases, human language learners have at the outset. Are there languages that people could never learn? Empiricists sometimes suggest that every pattern can be learned (in some relevant sense of 'learn'), while nativists suggest that humans are biased or constrained to consider only certain kinds of hypotheses. With the flourishing mathematical linguistics and learning theory in the last 50 years, a significant body of clear and uncontroversial mathematical results has emerged, resolving some of the confusions of the earlier informal debates and revealing some clearer questions. Many basic relations among the various formalisms used to frame a learner's hypotheses about language are now rigorously understood, from Markov models and neural nets to certain formal versions of Chomskian grammars: some formalisms are strictly weaker than others, and others are notational variants in the sense that they define exactly the same patterns (or in probabilistic models, exactly the same probability measures over patterns). In mathematical approaches, given a proposed learning method, it is common to consider what kinds of languages the method can successfully learn, where success is sometimes modeled as convergence on a grammar for the language the learner is exposed to (sometimes called *exact identification* of the *target language*), or where success consists in getting arbitrarily close to the target language in a probabilistic sense. We can also consider whether the learner can succeed just by observing example utterances, and whether the learner can succeed when the data is incomplete, systematically biased, or noisy.

In a clear sense, scientific laws apply only in idealized settings, when there are no extraneous disrupting influences. From Newtonian mechanics

UCLA Department of Linguistics

to Mendel's or Fisher's evolutionary models of reproducing populations, the simple relations which hold in idealized settings are distorted to some extent in real applications by sometimes complex and often poorly understood factors. The revolution in theoretical linguistics since the 1950's was precipitated by the introduction of scientific and mathematical methods in the study of language structure, methods that similarly require abstraction from irrelevant factors. This point needs emphasizing again and again because, first, the idealizations of linguistic theory are less familiar than the 'frictionless planes,' 'ideal gases,' and Mendelian 'genes' that every schoolchild hears about. In the second place, people read linguistic theory expecting to hear about the 'language' familiar from common sense, or from ethnic and literary studies, when in reality the science must address something more abstract.[1] As even the brief review provided below should make clear, many different kinds of abstractions are involved in the various theoretical proposals about human language and learning.

Although language learning has been a central focus in theoretical linguistics, mathematical work on learning has had a life of its own. Note, in particular, that while linguists are primarily concerned with linguistic structure, the language learner must get the first clues about the language without knowing what that structure is. The learner must infer the hidden structure of language (which Chomsky calls the internal 'I-language') from audible and visible external ('E-language') clues provided in linguistic settings. Along with many others, Chomsky (1981, p.10) has emphasized this point, saying that for the learning problem, perceptible features of the language have an epistemological priority; the learner must begin with the aspects of language that can be identified before the grammar of the language is known. So in order to define a collection of fixed learning problems that can be studied, we must adopt some (simplified and idealized) assumptions about the perceived evidence available to the learner (e.g. sequences of sounds uttered in certain contexts) and about the 'target' grammars to be learned. With an understanding based on such assumptions, we can then investigate whether, when confounding factors are controlled to the extent possible, there is evidence of similar learning in people.

Linguistics and mathematical learning theory have had a significant impact on each other, and many questions about human language learning have become much clearer, but versions of the old debates remain, as will become clear below. We will see that all of the mainstream mathematical models adopt restrictive assumptions about the range of identifiable patterns, as the nativist suggested, but the empiricist idea that the learner can be regarded as a rational agent, constructing the most probable explanation of the data, is also supported.

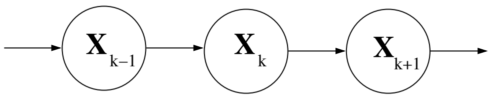
This paper briefly reviews some of the main applications of mathematical learning theory to language, emphasizing points of consensus that have emerged, and the nature of some remaining controversies.

1. Markov models

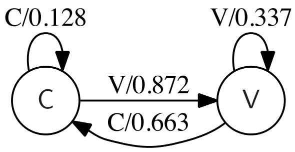
Although the connection between signs and what they signify seems arbitrary at first, and although it seems we can say whatever we like, pronouncing our words as we please, even superficial inspection reveals regularities that are preserved across utterances, across speakers, and even across languages. One class of these can be quantified with the simple models introduced by Markov (1906, 1908). These models regard linguistic events as the 'states' or 'outcome events' of a random variable at each moment of utterance, subject to a very strict independence condition:

> an infinite sequence $X_1, X_2, \ldots, X_k, X_{k+1}, \ldots$, of variables connected in such a way that X_{k+1} for any k is independent of $X_1, X_2, \ldots, X_{k-1}$, in case X_k is known.

That is, we can consider each word in a sequence of words, or each phoneme in a sequence of phonemes, as an event (the outcome of a variable) that depends only on the previous event and not on any earlier ones. Now called a 'Markov chain', a sequence with this simple kind of dependency is sometimes depicted with a graph like this, (Jordan et al., 1999):



Markov (1913) used a chain of this kind, with a random variable taking two states to model the vowel and consonant sequences in the first 20,000 symbols of the Pushkin poem *Eugene Onegin*. When the transition probabilities do not vary from one variable X_{k-1} to the next X_k, the probabilities of each state transition can be represented in a finite state diagram or, more succinctly, in a matrix M. Markov found the following transition probabilities on average:



$$M = \begin{pmatrix} .128 & .872 \\ .663 & .337 \end{pmatrix}$$

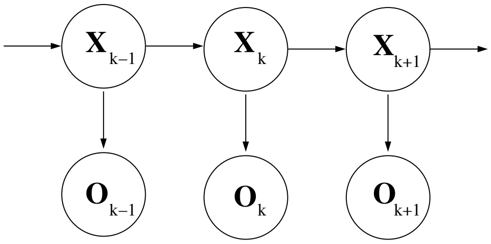
In the matrix, the rows and columns correspond to the consonants C and vowels V, respectively, so that row 1 column 2 represents the probability of going from state C of X_k to state V of variable X_{k+1} (for any point in time k). So for example, this model indicates that in the data, after hearing a C, there is a 12.8% chance that the next symbol will be a C, and a 87.2% chance of V. Elaborating these predictions a little more, the model tells us that after hearing a C, the probability of hearing CCCC is $0.128^4 = 0.000268435$, while VCVC is more than 1000 times more likely, $0.872 \times 0.663 \times 0.872 \times 0.663 = 0.334241$. The values in the matrix are called the *parameters* of the model, and are sometimes represented as *weights* rather than probabilities.[2] The learner can set these parameters according to the relative frequencies of these transitions in the data – the simplest kind of learning strategy. Markov shows that this simple model predicts quite well the number of vowels and consonants in the text.

While the particular probabilities of consonant-vowel transitions vary across speakers and languages, the general tendency to avoid long consonant clusters and vowel combinations is universal (e.g. Greenberg, 1978; Zec, 1995). Notice that the range of patterns that can be modeled with this 2 state Markov chain is very limited. For example, an 'avoid coda' preference for syllable initial versus syllable final consonants cannot be represented. This limitation comes from the basic structure of the model, from the number of states and Markov's independence assumption; the question of whether you are at the beginning or end of a syllable cannot be defined simply by the previous sound. Obviously no amount of training the model to get a more accurate probability matrix can overcome the bounds imposed by the choice of model. The interplay here between the stipulated model structure and rational setting of model parameters is reminiscent of the different emphases of nativists and empiricists.

The simple structure of Markov chains which makes them unable to capture many of the dependencies among elements in language is also what makes them so useful as a first approximation, and so they are still very extensively used in models of language and of language learning, from the classic beginnings of information theory (Shannon, 1948) to recent work on language learning and evolution Niyogi (2006). In this review, focusing on the linguists' interest in finding models of language and learning that *can* model the structures found in human language, we consider some of the most important steps in that direction.

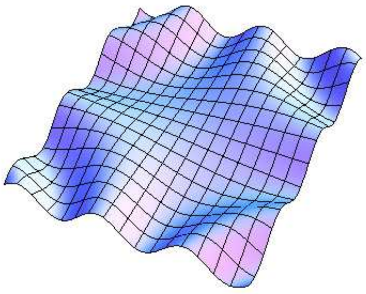
1.1. Hidden Markov models (HMMs)

The bounds of any particular Markov chain can be enlarged by allowing more states, and an additional degree of freedom can be obtained by allowing that the state is not perfectly represented by the utterance (or any other observable data). Hidden Markov models (HMMs) take these steps, and are probably the most widely used models for the sound patterns of speech (Cappé, Moulines, and Rydén, 2005; Jelinek, 1999; Rabiner, 1989). In these models, the states of the observable random variables $\ldots, O_k, \ldots$ depend on a Markov chain $\ldots, X_k, \ldots$ that is hidden in the sense that its states cannot be directly observed, with the dependencies indicated by the diagram (Jordan et al., 1999),

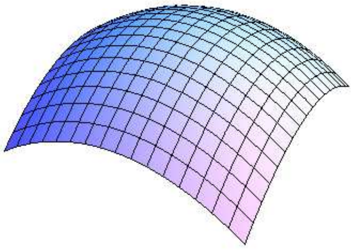


This model is a step in the right direction. For example, we could model how a language learner, observing a sequence of consonants and vowels, learns that these are pronounced at the beginning, middle, and ends of syllables even though the syllable structure is not explicitly given. Then outputs C,V would be the observable events $\mathbf{O}_i$, and the associated positions in syllable structure would be the hidden values of $\mathbf{X}_i$. Similarly, a learner hearing a sequence of words may need to learn that a given kind of word $\mathbf{O}_i$ is associated with certain implicit, hidden positions $\mathbf{X}_i$, e.g., the beginning of a noun phrase. HMMs can be specified by (i) a finite matrix M like the one above, specifying the probabilities of each event $\mathbf{X}_{k+1}$ based on the previous one $\mathbf{X}_k$, together with (ii) a matrix O indicating the probability of each possible output given the current state, and (iii) a matrix I indicating the probabilities of starting in each state. So for example, a given range of vowels or consonants will be more or less probable outputs at each point in the production of parts of a syllable, which are not explicitly given, but hidden, structural aspects of the input that the learner must discover. (In speech recognition, the acoustic signal is often sampled in much smaller, overlapping slices $\mathbf{O}_i$, thousands per second.)

One reason for the popularity of HMMs is the fact that when the transition probabilities of the hidden states and the output probabilities are unknown, it is possible to adjust these probabilities in a way that makes the data more probable. This fits with the very general empiricist conception of learning (and of scientific investigation) as the identification of a model that best explains for the data (Jaynes, 2003). These learning methods are sometimes called gradient or variational, since they adjust the model in a way guaranteed to improve the fit with the data. So again, these systems have fixed bounds on the range of patterns that can be represented, coming from the number of states in the model and the independence assumptions, but within those bounds, a certain kind of learning is possible. One important wrinkle enters the picture here, coming from the careful study of the complexity of various learning methods. Within the range delimited by the fixed number of states, we can ask whether a gradient learning method will always succeed in finding the transition probabilities giving the best fit with the data. For standard methods, the answer is that these methods will not, in general, find the best fit. In particular, the methods will fail when there are local maxima, points that are not the best setting, but which are surrounded by points that fit less well. We could abandon these gradient



non-convex function of 2 dimensions



convex function of 2 dimensions

methods in favor of learning strategies that are guaranteed to succeed, but these are intractable in the general case (Terwijn, 2002).

So we have this rather surprising situation: the tractable learning methods cannot be guaranteed to find the best fit, even within the limits imposed by the structure of the model, and so the performance of the various heuristic methods for parameter setting then becomes a central research concern.

Expectation maximization (EM) methods (Dempster, Laird, and Rubin, 1977) are most commonly used: each adjustment in weights is made to increase the expected fit, and then the weights are re-estimated, until the adjustments needed are very small. If the adjustments get smaller and smaller, approaching a particular point, we say the learner is *converging.* Ideally, the learner will converge on the optimum fit, the best possible model of the data, but gradient methods like EM will sometimes converge on points that are locally but not globally optimal. Furthermore, EM is sometimes very slow to converge, so many other learning methods are explored, including the Newtonian methods more commonly used in other approximation problems (MachLachlan and Krishnan, 1997; Cappé, Buchoux, and Moulines, 1998; Minami, 2004)

1.2. Maximum entropy Markov models

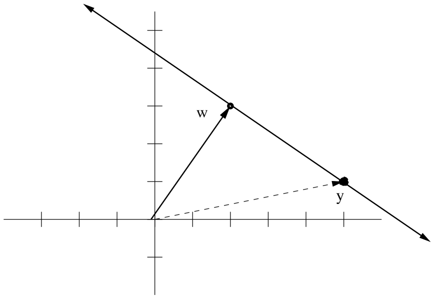
There are some problems where the gradient methods will never get stuck on a non-optimal hypothesis, a hypothesis from which all immediate changes result in even less optimal ones. If the fit with a model varying along n dimensions is convex,[3] then a class of relatively well-understood gradient 'convex optimization' methods may apply (Boyd and Vandenberghe, 2004). In a convex space, any move towards a more probable model is a move towards the global optimum. So why not just define our language models in such a way that the fit with the data varies as a convex function of variation in model parameters? Berger, Della Pietra, and Della Pietra (1996) show how convex model spaces can be constructed for the probabilities of finitely many features, and McCallum, Freitag, and Pereira (2000) extend the strategy to maximum entropy Markov models (MEMMs). which are carefully restricted so that the maximum likelihood is convex in the weights associated with each such transition. Some other linguistic models with convex search spaces have been explored too, including, for example, assigning finitely many parts of speech to the words of English sentences (Ratnaparkhi, 1996), parameter setting in probabilistic context free grammars (Chi, 1999; Geman and Johnson, 2003), and the maxent phonotactics of Wilson and Hayes (2008). Apart from the details of each of these proposals, there are general empirical questions about whether the space of hypotheses available to human learners is really one in which, from any point, the global optimum can be reached by successive improvements, climbing along the gradient, and about whether it is methodologically appropriate for linguists to seek grammar formalisms guaranteeing this. These are topics of ongoing research.

2. Neural models

Although Markov's work on quantitative analysis of event sequences dates from 1906, HMMs with parameters inferred by computationally intensive EM and other related methods have become widely used only since Dempster, Laird, and Rubin (1977). So it is no surprise that Rosenblatt's (1958) work on the perceptron is often cited as the first serious work in learning theory.[4] Inspired by the behavior of single neurons, a simple perceptron is a function f with an associated 'weight vector' of real numbers $\mathbf{w}$ and a 'threshold constant' b, mapping vector $\mathbf{x}$ to 1 if and only if the dot product of $\mathbf{w}$ and $\mathbf{x}$ is greater than b,[5]

$$f(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{w} \cdot \mathbf{x} > b \\ 0 & \text{otherwise.} \end{cases}$$

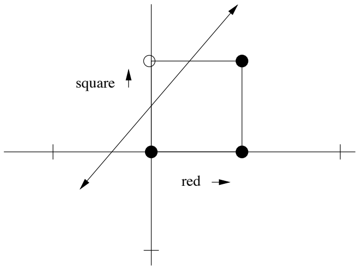
The idea here is very simple. Like a neuron that fires or not depending on the particular weighting and stimulation of its input dendrites, this function will become 1 according to the particular values of the coordinates of $\mathbf{w}$ and the constant b. For example, two weights or probabilities in an HMM could be representated by a particular point in the 2 dimensional plane, and the firing of the neuron, telling us whether the point is in the concept or not, is then determined by the constant b. The remarkable mathematical fact, the fact that explains the particular form of this definition of the perceptron f, is that in the 2-dimensional Cartesian plane, a point $\mathbf{w}$ and a constant b define a line: let the line be all the points $\mathbf{x}$ such that $\mathbf{w} \cdot \mathbf{x} = b$. Consider for example the line in the following diagram:



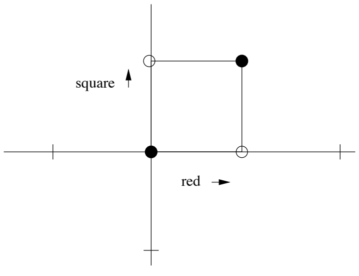
$\mathbf{w} = (2, 3)$ and
$b = 13$ determine a line
that includes, e.g. $\mathbf{y} = (5, 1)$

With the values of $\mathbf{w}$ and b shown in the diagram, f is defined as the function which maps everything above the line to 1, and everything on the line or below to 0. The definition of a perceptron f in terms of a point $\mathbf{w}$ and a constant b generalizes immediately to situations where $\mathbf{w}$ is a point from 3-dimensional or even higher dimensional spaces. Geometrically, corresponding to a line in 2 dimensions, a *hyperplane* is a set of points $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{w} \cdot \mathbf{x} = b$, and so the perceptron f, defined above, maps all points on one side of the hyperplane to 0, and all other points to 1.

Although the perceptron may seem very artificial and mathematical at first, it turns out that many concepts can be regarded as perceptron functions, even concepts given by discrete, symbolic values. For example, to learn a Boolean shape-color concept defined by a propositional formula like red or not square, we can think of the propositions red and square as the dimensions of $\mathbb{R}^2$, but where the concept actually takes values only at the points where the propositions are either true (with value 1) or false (with value 0). Then the concept red or not square can be depicted as on the left below, showing the points where the concept is true with solid circles, linearly separable from the points where the concept is false which are drawn with open circles,

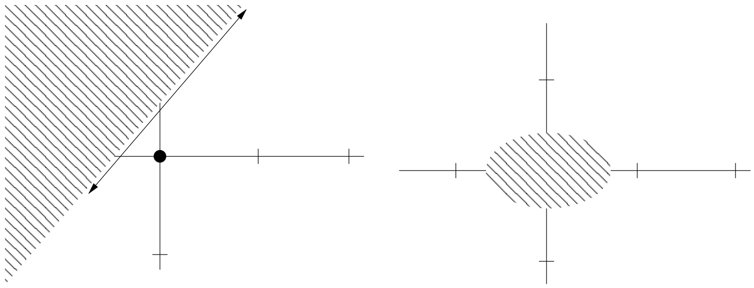


red $\vee\neg$ square



red $\oplus\neg$ square

The concept red or not square, but not both, which uses the exclusive-*or* ($\oplus$), obviously does not allow any line to separate the positive from the negative points (Minsky and Papert, 1971). For geometrical concepts on the plane (or higher dimensions), there are of course infinitely many concepts that are not linearly separable, including such simple concepts as an ellipse in the plane:



linearly separable

not linearly separable

Perceptrons are limited in this way. They can only describe concepts that allow the positive and negative examples to be separated by a line (or hyperplane in higher dimensions).

The particular importance of the perceptron stems mainly from the fact that it allows an extremely simple learning strategy. Suppose you begin with some initial guess about $\mathbf{w}$ and b, but you are given example points $\mathbf{x}$ and told whether they are examples of the concept or not. If $\mathbf{x}$ is an example of the concept and your current guess about $\mathbf{w}$ and b is such that $f(\mathbf{x}) = 1$, then your guess already fits this data point and no change is needed. But if $f(\mathbf{x}) = 0$, Rosenblatt proposed a very simple adjustment strategy. Roughly, it suffices to add $\mathbf{x}$, or some fraction of it, to $\mathbf{w}$, moving the line towards the data point.[6] This adjustment seems like something that fairly simple neural mechanisms might be able to realize, and a guarantee that this kind of method will learn any linearly separable concept was established by Novikoff (1962).[7]

Many facts about language can be regarded as Boolean and hence at least sometimes separable by hyperplanes. For example, can objects precede the verb? Can the subject pronoun be unpronounced? But it seems that many concepts in grammar are not naturally given as half of an n-dimensional space. For example, does it even make sense to think of defining the grammar of syllable structure this way, or the lexicon and syntax of French or English? Surprisingly, it turns out that, by transforming a grammar or lexicon into other (sometimes numerical, high-dimensional) forms, a remarkable range of these options *can* be given a clear sense, allowing linear learners in domains where they seemed impossible. This strategy is very briefly mentioned again in §2.2 below.

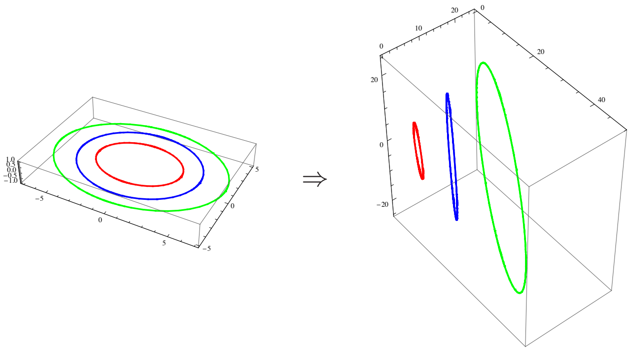
2.1. Neural networks with hidden states

One strategy for extending neural models beyond the perceptron introduces layers of 'hidden' perceptron-like units (neurons) between the n inputs from $\mathbb{R}^n$ and the output. Learning strategies for the functions described by such finite neural nets were not discovered until more than two decades after the work on perceptrons, when Rumelhart, Hinton, and Williams (1986) noticed a simple way to adjust all the weights of a network appropriately, using a gradient strategy called 'back-propagation'. This method propagates a correction to each neuron back from the output layer, moving the network towards a local minimum error with an adjustment similar to the one made by the perceptron learner. Rumelhart and McClelland (1986) use neural nets to model the learning of the English past tense, capturing the fact that learners often go through a phase of over-generalization. This project stimulated a great deal of controversy and follow-up work (Pinker and Prince, 1988). Another seminal neural network model learns categories: Guenther and Gjaja (1996) use neural networks to model the 'perceptual magnet' effect (Kuhl, 1991; Feldman and Griffiths, 2007), that is, the tendency for discriminability of sounds to be reduced when they are similar to prototypical linguistic sounds, so that small deviations from prototypical sounds are not noticed. However, standard neural network learners use gradient learning methods, and so can fail to find points with globally minimum error, making the representation of the problem and initial values critical. And the precise adjustment in weights required for back-propagation is not biologically realistic (Mazzoni, Andersen, and Jordan, 1991), prompting interest in the feasibility of simpler 'reinforcement' learning strategies (Sutton and Barto, 1998).

Optimality theoretic (OT) phonology and syntax may be seen as emerging from neural network fundamentals (Smolensky and Legendre, 2006), with a special notion of 'optimization' emerging at the symbolic level from a probabilistic, optimizing architecture. These grammars have very nice learnability properties when explored at the discrete symbolic level discussed in §3 below (Tesar and Smolensky, 1998; Riggle, 2009). Discrete OT grammars can be seen as a special case of probabilistic ('harmonic') grammars with continuous probability measures, but it remains unclear how the known grammar learning methods for these discrete systems should be implemented in a neural network architecture. Some interesting proposals are being explored recently (Goldrick and Daland, 2009; Magri, 2008).

2.2. Support vector machines (SVMs)

A different kind of response to the weakness of perceptron models was proposed by Aizerman, Braverman, and Rozoner (1964), and developed recently by (Vapnik, 1998) and many others. When a n-dimensional concept is not linearly separable, sometimes it is easy to encode the concept in a higher dimensional space in such a way that, in the larger space, it is linearly separable. For example, a 2 dimensional ellipse in the plane can be projected into 3 dimensions in such a way that just its boundary is on a plane, so that the positive and negative points can be separated.



And notice that to find the best line (or hyperplane), what matters are the examples that are closest to it; these points are sometimes called *support vectors*. Increasing the dimension of the space the learner must consider can involve a significant increase in complexity, but it need not always do so. In some cases, it is possible for the learner to adjust the higher dimensional hypothesis using feasible computation on their lower dimensional representations (Cristianini and Shawe-Taylor, 2000). This method has found useful application even in problems that are known to be intractable in principle, like learning all the Boolean functions (including exclusive-*or*) (Sadohara, 2001; Khardon and Servedio, 2004), and an increasing range of language learning problems (Kontorovich, Cortes, and Mohri, 2006; Clark, Florêncio, and Watkins, 2006; Kontorovich, Cortes, and Mohri, 2008). In the transformed representations of these learning problems, perceptron-like linear learning can be applied to identify complex concepts. It is quite conceivable that illuminating extensions of these recent methods to human language learning problems will be found.

3. Model selection and the nature of linguistic abstractions

The previous sections have ignored a fundamental (one might even say: *the* fundamental) problem in language learning, sometimes called the 'model selection' problem: what kinds of models should be explored? In HMMs, practical applications often require poorly understood decisions about how many states are needed, and how they can depend on each other (Smyth, Heckerman, and Jordan, 1997; Jordan et al., 1999). Similarly, in neural nets, learning depends on the network size and topology. And in SVMs, everything depends on the particular encoding of the problem. Many of the traditional debates about initial biases and the rationality of language learning really center on how this fundamental part of the learning problem is handled.[8]

An empiricist tends to emphasize respects in which the learner's response to the data is rational and free from a priori bias, but searching all possible models to find the best fit with the data is not a solution to this problem. First, model fit (which can be quantified and estimated with measures of 'mutual information', etc.) is not a good criterion, since the fit with the data typically improves with larger models (Rasmussen and Ghahramani, 2001), but simpler models often generalize better (‘Occam’s razor’). Second, even if the standard of model comparison is chosen, searching through all the options is impossible, and gradient-based stepwise models do not guarantee good results. These issues have been approached in many different ways, from asymptotic methods (Akaike, 1974; Schwarz, 1978), to bootstrap and boosting methods (Efron, 1979; Efron and Tibshirani, 1994) and a wide range of interesting Bayesian approaches (Raftery, 1995; Sato, 2001; Bishop, 2008; Beal, Ghahramani, and Rasmussen, 2002; Teh et al., 2006; Fox et al., 2008; van Gael et al., 2008).

Model selection can also sometimes be done non-probabilistically. Sometimes it can be determined that a current hypothesis simply does not admit the data; that is, the data has probability 0 given the hypothesis, signalling that a different, sometimes more complex set of hypotheses should be considered. This is the basic idea behind many non-probabilistic, discrete learning methods (Gold, 1967; Jain et al., 1999), methods which can be applied successfully even when there are infinitely many models, some of which have infinitely many states. The Chomskian tradition of generative syntax – and perhaps all of the mainstream traditions in linguistic theory – can be regarded as a model-selection effort. Their goal is a class of hypotheses which can be fit to any human language by parameter setting methods.

3.1. Model selection from example sequences

Many familiar languages can only be recognized by systems with infinitely many states. For example, propositional logic is often written with parentheses to avoid ambiguity, so that $\mathsf{red\ or\ not\ square,\ but\ not\ both}$ is $((\mathsf{red} \lor\neg \mathsf{square})\land\neg(\mathsf{red} \land\neg\mathsf{square}))$. Another notation that avoids ambiguity without the use of parentheses is the prefix notation, $\land\lor\mathsf{red}\neg\ \mathsf{square}\neg\land\mathsf{red}\neg\mathsf{square}$. The language for this prefix notation can be described by a simple ‘context free’ rewrite grammar like this one:

$$\begin{array}{llll} \mathrm{S}\rightarrow\land\mathrm{SS} & \mathrm{S}\rightarrow\lor\mathrm{SS} & \mathrm{S}\rightarrow\neg\mathrm{S} & \\ \mathrm{S}\rightarrow\mathsf{red} & \mathrm{S}\rightarrow\mathsf{square} & \mathrm{S}\rightarrow\mathsf{big} & \ldots \end{array}$$

That is, the propositions red, square, big,…have the sentence category S, and a conjunction is not (S∧S) but ∧SS. The sentences of this language are obtained by rewriting the category symbol S repeatedly in any way allowed by the rules, as in

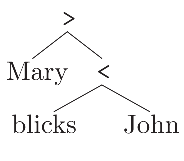
$$\mathrm{S} \Rightarrow \lor\mathrm{SS} \Rightarrow \lor\mathsf{red}\mathrm{S} \Rightarrow \lor\mathsf{red}\neg\mathrm{S} \Rightarrow \lor\mathsf{red}\neg\mathsf{square}.$$

Notice that this language includes $\land\mathsf{red\ red}$ and $\land\land\mathsf{red\ red\ red}$ but not $\land\mathsf{red}$ or $\land\land\mathsf{red\ red}$. Consider any device that can recognize this language, distinguishing grammatical sentences from nonsense strings. For any two different numbers n, m, the state of any recognizer after seeing $\land^n$ needs to be different from the state of the recognizer after seeing $\land^m$, since these two situations impose different requirements on what follows. Intuitively, the recognizer needs to count the number of coordination symbols ∧ that begin any sentence, and since the grammar imposes no limit on how many there can be, the recognizer cannot be finite. That is, no 'finite state model' can represent this language.[9] Standard models of this language also impose certain kinds of dependencies among the states – only certain states can be reached from any given one. These dependencies among states are really what is of interest in practical applications, since real world devices are not infinite; they are described as if they were infinite in order to understand the dependencies among their states.

The prefix notation for propositional calculus is a 'very simple' context free language in a sense defined by Yokomori (2003). A very simple context free language is one that can be defined with rules of the form $A \to wB_1 \ldots B_n$ for $n \geq 0$, where for each pronounced symbol w, there is exactly one rule. It is easy to see that the propositional language defined above is very simple in this sense, and that, for any finite alphabet of more than one pronounced word w, there are infinitely many very simple languages. It turns out that there is a learning algorithm which maps every initial sequence of example sentences to a very simple grammar for a language that includes that sample in such a way that, as the number of samples approaches the whole language, the algorithm will eventually exactly identify the language (Yokomori, 2003). That means, given certain finite sequences of examples, the learner will construct a grammar that generates the example sentences *together with infinitely many other sentences that have not been seen*, and this learner will eventually converge on a grammar that is exactly right. In particular, given the examples red and $\neg$red, the learner will already realize that the target language must contain $\neg^n$red for all n. The human language learner must generalize in something like this way too, realizing at some point that not only are funny and really funny good adjective phrases, but so is really^{n}funny for any n.

3.2. Model selection from example structures

Empirical studies of human learning support the common sense observation that children pay attention not only to the order of words in utterances, but also to the meanings of words that are plausible in the discourse context (Hirsh-Pasek and Golinkoff, 1996; Tomasello, 2003; Trueswell and Gleitman, 2004). That is, the evidence available to human language learners includes more than example word sequences. Learners hearing *Mary blicks John* with a novel word *blick* are likely to adopt a hypothesis about this word that allows it to combine with the subject and object in the way other words do. So instead of simply adjusting the grammar to allow *Mary blicks John*, the learner may adjust the grammar to allow the structure



where the arrows < and > ‘point’ to the expression that takes the other as argument. That is, *blicks* takes *John* as an argument, and *blicks John* takes *Mary* as an argument. Kanazawa (1996) has rigorously demonstrated that, if lexical ambiguity is suitably restricted, a learner getting this kind of structural evidence can identify a context free language, in the sense that, as the example structures of the language approach the whole set, the learner will, at some finite point, converge on a grammar that generates the whole language, exactly.

Notice that the learner described here is in effect noticing that *blicks* is in a ‘substitution class’ of elements that can occur in the context *Mary* ___ *John*. Kanazawa’s (1996) learner generalizes this kind of structuralist inference. This kind of learning is familiar in computer science too, since it is similar to the type inference in programming languages. In many programming languages where $0, 1, 2, \ldots$ are integers, if we define f to be the “is equal to 1” function by saying that fx is $true$ if $x = 1$ and otherwise $false$, the system will infer that f has the type $int \to t$, where t is the type of truth values. In a similar way, a learner who knows that *John* and *Mary* denote entities e and that sentences denote truth values t, might conclude from *John blicks Mary* that *blicks* denotes a function of type $e \to e \to t$.

Human languages are widely thought to require grammars that are more expressive than context free grammars. Joshi (1985) has suggested that the grammars required for human languages are ‘mildly context sensitive’, just slightly more expressive than context free grammars. Kanazawa’s result extends easily to some of the larger classes of languages (Retoré and Bonato, 2001; Stabler et al., 2003; Fulop, 2007). The general strategy of considering learners that make use of convergent syntactic, prosodic, semantic and discourse cues (data with more relevant ‘dimensions’) brings us closer to models of the human learner’s predicament, but extending the formal results to human languages requires relaxing the non-ambiguity restrictions, and restricting the data to structures that the learner could plausibly get evidence for.

Notice that all these learners succeed only on a limited range of languages (as the nativist would be inclined to emphasize), but the limited range can still include infinitely many languages, where many of those languages each require infinite state recognizers. Even in these cases, the learning methods can be successful, and feasible.

4. The future

A number of results mentioned in this survey are relatively uncontroversial, with consequences for every approach to language:

- A wide range of mathematical models of language learning are being explored, involving different assumptions about the most important or most relevant dependencies among the states responsible for the sequences of linguistic elements.

- In many classes of models, the problem of calculating the parameter settings that maximize the probability of the evidence is intractable, so various heuristic methods are used.
- A central goal of linguistic theory has been to determine which kind of model is most appropriate (how many states, with what dependencies). This model selection problem is typically done 'by hand', and is still rather poorly understood.

The consensus around these points has been noted before (Pereira, 2000). But certainly the last point is the most interesting. Recent theoretical linguistics is focused on model selection, on determining what kinds of models are appropriate for describing human languages (roughly speaking: how many states, and how can they depend on each other). Finding linguistic universals, the bounds of variation, has been a central goal of generative grammar at least since Chomsky (1965), and this methodological stance has explicitly been related to the learning problem. Turning to the content of the proposals that have emerged in generative grammar, though, the impact of mathematical learning models on mainstream linguistic theory has been more marginal (though still significant, as noted for example at the ends of §§1.2, 2.1, 2.2). This is no surprise, since model selection is exactly where mathematical approaches have had the least to offer, and it is also no suprise that this is where empiricist and nativist rhetoric remains most prominent. But active and ongoing research in the interplay between model selection and parameter tuning has illuminated much that had been obscure, and this is certainly where the most important developments will be.

Notes

[0] For helpful suggestions, I am grateful to Jeff Heinz, Greg Kobele, Katya Pertsova, Kie Zuraw, Jason Riggle, and the editors of this issue.

[1] The importance of abstraction is emphasized for example in Chapter 1 of Chomsky's (1965) *Aspects of the Theory of Syntax.* In recent work Chomsky (1996, pp.7,15) is still repeating these fundamental themes and emphasizing the necessity: "Idealization, it should be noted, is a misleading term for the only reasonable way to approach a grasp of reality."

[2] Typically, for probability p, the weight $w = -\log p$.

[3] A function $f : \mathbb{R}^n \to \mathbb{R}$ is 'convex' or 'concave up' if and only if for any $x, y \in \mathbb{R}^n$ and any $0 \leq \alpha \leq 1$, $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$ (Boyd and Vandenberghe, 2004, §1.1). A function is 'concave down' if its negation is concave up. Notice that since the axes are not labeled in the figures, we cannot tell whether the function on the right is 'concave up' or 'concave down.' Convex optimization methods can of course apply in either case.

[4] For example, Vapnik (2000, p.1) and Duda, Hart, and Stork (2001, p.333).

[5] Recall that for two vectors, $\mathbf{x} = \langle x_1, \ldots, x_n \rangle$ and $\mathbf{y} = \langle y_1, \ldots, y_n \rangle$, the dot product $\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i$, the sum of the products of the respective components.

[6] See for example the rigorous formulation of this method in Anderson (1995, p.221) or Cristianini and Shawe-Taylor (2000, Table 2.1).

[7] For a modern text presentation see for example Anderson (1995, §8).

[8] The distinction between 'model selection' and 'parameter setting' is prominent in the literature, but it is an informal, qualitative distinction. It is of course possible to define parameters in such a way that they determine the required states of the model (thus, in effect, 'selecting the model'); and it is also possible to tune the parameters of infinite state systems: infinite neural nets (Hornik, Stinchcombe, and White, 1989; Neal, 1996), infinite HMMs (Beal, Ghahramani, and Rasmussen, 2002; Fox et al., 2008; van Gael et al., 2008), etc.

[9] This reasoning is formalized in the standard Myhill-Nerode theorem (Hopcroft and Ullman, 1979, §3.4).

References

Aizerman, M.A., E.M. Braverman, and L.I. Rozoner. 1964. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–825. Translated from *Automatika i Telemekhanika 25*(6): 917-936, 1964.

Akaike, Hirotugu. 1974. A new look at statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Anderson, James A. 1995. *An Introduction to Neural Networks*. MIT Press, Cambridge, Massachusetts.

Beal, M. J., Z. Ghahramani, and C. E. Rasmussen. 2002. The infinite hidden Markov model. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14: Proceedings of the 2001 Neural Information Processing Systems (NIPS) Conference*, pages 577–585, Cambridge, Massachusetts. MIT Press.

Berger, Adam L., Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22:39–72.

Bishop, Christopher M. 2008. A new framework for machine learning. In *Computational Intelligence: Research Frontiers, IEEE World Congress on Computational Intelligence (WCCI'08)*, Lecture Notes in Computer Science LNCS 5050, pages 1–24, NY. Springer.

Boyd, Stephen and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press, NY.

Cappé, O., V. Buchoux, and E. Moulines. 1998. Quasi-Newton method for maximum likelihood estimation of hidden Markov models. In *Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP)*, pages 2265–2268.

Cappé, O., E. Moulines, and T. Rydén. 2005. *Inference in Hidden Markov Models*. Springer, NY.

Chi, Zhiyi. 1999. Statistical properties of probabilistic context free grammars. *Computational Linguistics*, 25:130–160.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, Massachusetts.

Clark, Alexander, Christophe Costa Florêncio, and Chris Watkins. 2006. Languages as hyperplanes: grammatical inference with string kernels. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 90–101.

Cristianini, Nello and John Shawe-Taylor. 2000. *Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, NY.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38.

Duda, Richard O., Peter E. Hart, and David G. Stork. 2001. *Pattern Classification*. Wiley, NY.

Efron, Bradley. 1979. Bootstrap methods. *Annals of Statistics*, 7:1–26.

Efron, Bradley and R. J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman and Hall, London.

Feldman, N. H. and T. L. Griffiths. 2007. A rational account of the perceptual magnet effect. In *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society*.

Fox, E., E. Sudderth, M. I. Jordan, and A. Willsky. 2008. An HDP-HMM for systems with state persistence. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*.

Fulop, Sean. 2007. The learnability of type-logical grammars. *Research on Language and Computation*, 5(2):159–179.

Geman, Stuart and Mark Johnson. 2003. Probability and statistics in computational linguistics, a brief review. *Mathematical foundations of speech and language processing*, 138:1–26.

Gold, E. Mark. 1967. Language identification in the limit. *Information and Control*, 10:447–474.

Goldrick, Matthew and Robert Daland. 2009. Linking speech errors and phonological grammars: Insights from harmonic grammar networks. *Phonology*, 26.

Greenberg, Joseph. 1978. Some generalisations concerning initial and final consonant clusters. In Joseph Greenberg, editor, *Universals of Human Language*. Stanford University Press, Stanford, California, pages 243–279.

Guenther, F. H. and M. N. Gjaja. 1996. The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America*, 100:1111–1121.

Hirsh-Pasek, Kathy and Roberta Michnick Golinkoff. 1996. *The Origins of Grammar: Evidence from Early Language Comprehension*. MIT Press, Cambridge, Massachusetts.

Hopcroft, John E. and Jeffrey D. Ullman. 1979. *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, Reading, Massachusetts.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Jain, Sanjay, Daniel Osherson, James S. Royer, and Arun Sharma. 1999. *Systems that Learn: An Introduction to Learning Theory (second edition)*. MIT Press, Cambridge, Massachusetts.

Jaynes, E.T. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press, NY.

Jelinek, Frederick. 1999. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, Massachusetts.

Jordan, Michael I., Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. 1999. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Joshi, Aravind. 1985. How much context-sensitivity is necessary for characterizing structural descriptions. In D. Dowty, L. Karttunen, and A. Zwicky, editors, *Natural Language Processing: Theoretical, Computational and Psychological Perspectives*. Cambridge University Press, NY, pages 206–250.

Kanazawa, Makoto. 1996. Identification in the limit of categorial grammars. *Journal of Logic, Language, and Information*, 5:115–155.

Khardon, Roni and Rocco A. Servedio. 2004. Maximum margin algorithms with Boolean kernels. *Journal of Machine Learning Research*, 6:1405–1429.

Kontorovich, Leonid, Corinna Cortes, and Mehryar Mohri. 2006. Learning linearly separable languages. In J. L. Balcázar, P. M. Long, and F. Stephan, editors, *Algorithmic Learning Theory, 17th International Conference, ALT'06*, volume 4264 of *LNCS/LNAI*, pages 288–303. Springer.

Kontorovich, Leonid, Corinna Cortes, and Mehryar Mohri. 2008. Kernel methods for learning languages. *Theoretical Computer Science*, 405(3):223–236.

Kuhl, Patricia K. 1991. Human adults and infants show a 'perceptual magnet' effect for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics*, 50:93–107.

MachLachlan, Geoffrey J. and Thriyambakam Krishnan. 1997. *The EM Algorithm and Extensions*. Wiley, NY.

Magri, Giorgio. 2008. Linear methods in optimality theory: A convergent incremental algorithm that performs both promotion and demotion. In *Northeast Computational Phonology Meeting, NECPhon 2008*. http://pantheon.yale.edu/~gjs42/necphon08/.

Markov, Andrei Andreevich. 1906. Rasprostranenie zakona bol'shih chisel na velichiny, zavisyashchie drug ot druga (Extending the law of large numbers for variables that are dependent of each other). *Izvestiya Fiziko-matematicheskogo obshchestva pri Kazanskom universitete*, 15:124–156.

Markov, Andrei Andreevich. 1908. Rasprostranenie predel'nyh teorem ischisleniya veroyatnostej na summu velichin svyazannyh v cep'. *Zapiski Akademii Nauk po Fiziko-matematicheskomu otdeleniyu*, VIII,25(3). English translation, "Extension of the limit theorems of probability theory to a sum of variables connected in a chain" (translated by S. Petelin) in R. A. Howard (ed.), *Dynamic Probabilistic Systems, Volume 1*, Wiley, New York, 1971, pp. 552-576.

Markov, Andrei Andreevich. 1913. Primer statisticheskogo issledovaniya nad tekstom “Evgeniya Onegina”, illyustriruyushchij svyaz’ ispytanij v tsep’. *Izvestiya Akademii Nauk*, 7:153–162. An English translation, “An Example of Statistical Investigation of the Text *Eugene Onegin* Concerning the Connection of Samples in Chains” (translated by G. Custance and D. Link) appears in *Science in Context 19*(4): 591-600, 2006.

Mazzoni, Pietro, Richard A. Andersen, and Michael I. Jordan. 1991. A more biologically plausible learning rule for neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 88(10):4433–4437.

McCallum, Andrew, Dayne Freitag, and Fernando Pereira. 2000. Maximum entropy Markov models for information extraction and segmentation. In *Machine Learning: Proceedings of the Seventeenth International Conference (ICML 2000)*, pages 591–598, Stanford, California.

Minami, Mihoko. 2004. Convergence speed and acceleration of the EM algorithm. In Michiko Watanabe and Kazunori Yamaguchi, editors, *The EM Algorithm and Related Statistical Models*. CRC Press, NY, pages 85–94.

Minsky, Marvin L. and Seymour Papert. 1971. *Perceptrons*. MIT Press, Cambridge, Massachusetts.

Neal, Radford M. 1996. *Bayesian Learning for Neural Networks*. Springer, NY.

Niyogi, Partha. 2006. *The Computational Nature of Language Learning and Evolution*. MIT Press, Cambridge, Massachusetts.

Novikoff, A.B. 1962. On convergence proofs and perceptrons. In *Symposium on the Mathematical Theory of Automata 12*, pages 615–622.

Pereira, Fernando C. N. 2000. Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society*, 358:1239–1253. Reprinted in *The Legacy of Zellig Harris, Volume 2*, edited by Bruce F. Nevin and Stephen M. Johnson, John Benjamins, Philadelphia, 2002.

Pinker, Steven and Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28:73–193.

Rabiner, Lawrence R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286.

Raftery, Adrian E. 1995. Bayesian model selection in social research. In P. V. Marsden, editor, *Sociological Methodology*. Blackwell, Oxford, pages 111–196.

Rasmussen, C. E. and Z. Ghahramani. 2001. Occam’s razor. In V. Tresp T. Leen, T. Dietterich, editor, *Advances in Neural Information Processing Systems 13*. MIT Press, Cambridge, Massachusetts.

Ratnaparkhi, Adwait. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing*

Conference, Philadelphia, Pennsylvania.

Retoré, Christian and Roberto Bonato. 2001. Learning rigid Lambek grammars and minimalist grammars from structured sentences. In L. Popelínský and M. Nepil, editors, *Proceedings of the Third Learning Language in Logic Workshop, LLL3*, pages 23–34, Brno, Czech Republic. Faculty of Informatics, Masaryk University. Technical report FIMU-RS-2001-08.

Riggle, Jason. 2009. The complexity of ranking hypotheses in optimality theory. *Computational Linguistics*, 35(1).

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

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323:533–536.

Rumelhart, David E. and James L. McClelland. 1986. On learning the past tenses of English verbs. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models*. MIT Press, Cambridge, MA, USA, pages 216–271.

Sadohara, Ken. 2001. Learning of Boolean functions using support vector machines. In *International Workshop on Algorithmic Learning Theory, ALT'01*, Lecture Notes in Artificial Intelligence, pages 106–118, NY. Springer.

Sato, Masa-Aki. 2001. Online model selection based on the variational Bayes. *Neural Computation*, 13(7):1649–1681.

Schwarz, Gideon. 1978. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Shannon, Claude E. 1948. The mathematical theory of communication. *Bell System Technical Journal*, 127:379–423. Reprinted in Claude E. Shannon and Warren Weaver, editors, *The Mathematical Theory of Communication*, Chicago: University of Illinois Press.

Smolensky, Paul and Géraldine Legendre. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar, Volume I: Cognitive Architecture*. MIT Press, Cambridge, Massachusetts.

Smyth, Padhraic, David Heckerman, and Michael Jordan. 1997. Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, 9:227–269.

Stabler, Edward P., Travis C. Collier, Gregory M. Kobele, Yoosook Lee, Ying Lin, Jason Riggle, Yuan Yao, and Charles E. Taylor. 2003. The learning and emergence of mildly context sensitive languages. In W. Banzhaf, T. Christaller, P. Dittrich, J.T. Kim, and J. Ziegler, editors, *Advances in Artificial Life*. Springer, NY.

Sutton, Richard S. and Andrew G. Barto. 1998. *Reinforcement Learning*. MIT Press, Cambridge, Massachusetts.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.

Terwijn, Sebastian. 2002. On the learnability of hidden Markov models. In *Proceedings of the 6th International Colloquium on Grammatical Inference, ICGI'02*, pages 261–268, London, UK. Springer-Verlag.

Tesar, Bruce and Paul Smolensky. 1998. Learnability in optimality theory. *Linguistic Inquiry*, 29:229–268.

Tomasello, Michael. 2003. *Constructing a Language: A Usage-based Theory of Language Acquisition*. Harvard University Press, Cambridge, Massachusetts.

Trueswell, John and Lila Gleitman. 2004. Children's eye movements during listening: Developmental evidence for a constraint-based theory of lexical processing. In John M. Henderson and Fernanda Ferreira, editors, *Interface of Language, Vision, and Action: Eye Movements and the Visual World*. Psychology Press, NY.

van Gael, J., Y. Saatci, Y.-W. Teh, and Z. Ghahramani. 2008. Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th International Conference on Machine Learning*.

Vapnik, Vladimir N. 1998. *Statistical Learning Theory*. Wiley, NY.

Vapnik, Vladimir N. 2000. *The Nature of Statistical Learning Theory*. Springer, NY.

Wilson, Colin and Bruce Hayes. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39:379–440.

Yokomori, Takashi. 2003. Polynomial-time identification of very simple grammars from positive data. *Theoretical Computer Science*, 298:179–206.

Zec, Draga. 1995. Sonority constraints on syllable structure. *Phonology*, 12:85–129.