Computational models of early language acquisition

Michael C. Frank

Department of Psychology, Stanford University

Abstract

How do children acquire the sounds, words, and structures of their native language? A wealth of recent evidence suggests that probabilistic learning mechanisms play a role in language acquisition. Nevertheless, the structure of these mechanisms is controversial and it is still unknown how broadly they apply to the tasks faced by language learners. Computational models can serve as formal theories of probabilistic learning by instantiating proposals about the learning mechanisms available in early language acquisition. However, fulfilling this promise requires that models be evaluated on two grounds: their sufficiency—whether they are able to learn aspects of language given appropriate input—and their fidelity—whether they fit the patterns of success and failure shown by human learners. I review experimental and computational evidence for the application of probabilistic learning across a range of acquisition tasks and argue that models of probabilistic learning succeed when they use expressive representations to capture complex regularities in the input and when they implement a parsimony bias.

Introduction

How do children learn their native language? Over a handful of years, infants who know almost nothing about any language become children who can express their thoughts fluently in one language in particular. Though the broad developmental course of language acquisition is well-established, there is virtually no consensus on the psychological mechanisms by which the different aspects of language are acquired. Are substantial aspects of linguistic structure innately given (Lenneberg, 1967; Chomsky, 1981; Pinker, 1995)? Or are infants endowed only with more general probabilistic learning mechanisms that can be applied to a broad class of tasks (Rumelhart, McClelland, & the PDP Research Group, 1986; Elman et al., 1996)? Since the birth of the field of language acquisition, the use of formal or computational tools to give a description of the machinery necessary to acquire a language has been recognized as an important strategy for answering these questions (Chomsky, 1975; Pinker, 1979).

In recent years, exciting empirical results on infant learning abilities (Saffran, Aslin, & Newport, 1996; Marcus, Vijayan, Bandi Rao, & Vishton, 1999; Gómez, 2002; Gerken, Wilson, & Lewis, 2005) combined with promising computational results (Vallabha, McClelland, Pons, Werker, & Amano, 2007; Goldwater, Griffiths, & Johnson, 2009; Goldsmith, 2001; Albright & Hayes, 2003; Perfors, Tenenbaum, & Regier, 2006; Alishahi & Stevenson, 2008; Bannard, Lieven, & Tomasello, 2009) have together suggested that probabilistic learning is important for language acquisition.

I have chosen the term *probabilistic learning* to describe a phenomenon that is observed across a very broad range of experiments: that learners are able to acquire information even from observations that are individually ambiguous between a number of different hypotheses. I use this term instead of the more common term *statistical learning* because of the association that statistical learning has with a particular set of paradigms (Saffran, Aslin, & Newport, 1996; Saffran, Johnson, Aslin, & Newport, 1999; Fiser &

Aslin, 2002), and the contrast that has been made by some authors between statistical learning and other kinds of learning (Marcus et al., 1999; Pena, Bonatti, Nespor, & Mehler, 2002; Endress & Bonatti, 2007). In addition, a recent review gives the definition that statistical learning "involves no overt reinforcement or direct feedback but rather operates by mere exposure or observation" (Aslin & Newport, 2008). Although the vast majority of the computational theories reviewed here are unsupervised—they receive no feedback or reinforcement—probabilistic learning encompasses a broader range of possible learning mechanisms.

It is now widely accepted that general probabilistic learning mechanisms plays a large role in tasks like identifying the phonetic units of a language or identifying words from fluent speech (e.g. Kuhl, 2000, 2004; Saffran, Aslin, & Newport, 1996), but the nature and number of these mechanisms and their application to more complex aspects of language learning is still controversial. Although probabilistic learning mechanisms are acknowledged to play a part in word learning, they are often thought of as only one "cue" to identifying word meanings (Waxman & Gelman, 2009). In the acquisition of syntax, the ability of probabilistic mechanisms to identify language-specific rules has been even more controversial (Pinker, 1979; Wexler & Culicover, 1983; Pearl & Lidz, 2009). Thus, despite the progress that has been made, many questions regarding the role of probabilistic learning in language acquisition are still unanswered.

Creating computational models of individual tasks in acquisition—from sound category learning to syntactic rule learning—allows researchers to instantiate proposals about how a particular learning mechanism might perform in a particular domain. This strategy has led to a profusion of models in recent years, but it is not always clear how they should be compared or what generalizations emerge across different areas of acquisition. Thus, this article is a review of progress in modeling different areas of language acquisition. The review has three goals: (a) to describe criteria for evaluating

models on their adequacy as theories of language acquisition, (b) to survey computational models of early language acquisition across the full range of acquisition challenges and evaluate them on these criteria, and (c) to describe similarities between the most successful models across a range of tasks.

Summarizing the conclusions of the review, I argue that models should be evaluated on two criteria: sufficiency (learning the same thing as human learners with the same amount of input) and fidelity (making the same mistakes along the way). Application of these criteria to models of different acquisition tasks suggests that a tremendous amount of progress has been made in modeling early acquisition—sound category learning, word segmentation, and word learning. An important goal is consolidating this progress: systematic evaluations of existing models, extension of these models to incorporate other information sources, and testing of the models on novel predictions. In contrast, models of more complex acquisition tasks—word class learning, morphology learning, and syntactic rule learning—have further to go. Although there have been successes in these domains, there has been overall less convergence on assumptions that are shared across successful models.

As Goldsmith (2010) notes, if you dig deep enough into any task in acquisition, it will become clear that in order to model that task effectively, a model of every other task is necessary. While most of the models we review are models of a single domain, the review of different tasks concludes with a section on synergies between acquisition tasks (M. Johnson, 2008b). I finish by describing two broad generalizations that can be drawn from the models that succeed across all classes of tasks: first, that representations within these models should be efficient compressions of input data at the desired level of analysis, and second, that models should include some bias towards parsimony in the representations they learn.

Criteria for assessing proposed learning mechanisms

How can we assess whether a hypothesized learning mechanism (LM) truly plays a role for children in solving a particular challenge of language acquisition? Pinker (1979) proposed six conditions:

1. Learnability: LM must be able to acquire a language in the limit

2. Equipotentiality: LM must be able to learn any human language

3. Time: LM must be able to learn within the same amount of time as the child

4. Input: LM must be able to learn given the same amount of input as the child

5. Developmental: LM must make predictions about the intermediate stages of learning, and

6. Cognitive: LM must be consistent with what is known about the cognitive abilities of the child.

Although these conditions provide a detailed specification for evaluating a potential LM, several are difficult to evaluate. In particular, the time and cognitive conditions present a clear challenge for evaluation. How should the amount of time taken by a child to learn their native language limit a computer model of this process? The mapping between developmental time and computation cycles in a serial, digital computer is unknown (and the question in fact may not be well-formed). The number of computation cycles used by a computer program is a product of many conditions, including the operation set of the processor it is run on, the compiler or interpreter used to run it, and low-level algorithmic decisions in the code. None of these should be germane to the decision about whether it provides useful insight into language acquisition.

Likewise, although the cognitive condition appears compelling, and although we have an intuitive grasp of what can be consciously computed or remembered, we know very little about the true computational abilities of human learners. In the study of memory, it is continually surprising both how much we can (Brady, Konkle, Alvarez, &

Oliva, 2008) and cannot (Cowan, 2001) remember. With respect to computations, the computations hypothesized for perception (Marr & Poggio, 1979; Pouget, Dayan, & Zemel, 2000; Ma, Beck, Latham, & Pouget, 2006) or motor control (Koerding & Wolpert, 2004; Todorov, 2009) are often far more complex than those hypothesized to be difficult or impossible for language learners (Yang, 2004). Applied indiscriminately, the cognitive condition uses researchers' intuitions to limit the kinds of models that we consider. If those intuitions are incorrect, we may fail to consider appropriate mechanisms.

There has been recent interest in the distinction between *incremental* and *batch* models of learning. Batch models must have all of their input data present in order to perform their learning algorithm, while incremental models process their input example by example. Incremental learning is often argued to instantiate some version of Pinker's cognitive condition (Fazly, Alishahi, & Stevenson, 2008; McMurray, Aslin, & Toscano, 2009; McMurray, Horst, & Samuelson, under review). Incremental models thus describe a situation where memory for input data is sharply limited. Surely in the limit, human learners do not remember all the data they are exposed to, but should incremental models of human learning be preferred in every situation? Models that are too profligate with memory resources seem intuitively unappealing, but the fully incremental alternative is not always preferable.

Fully incremental learning prevents backtracking or re-evaluation of hypotheses in light of earlier data. This issue reappears throughout work on inference in computer science. For example, beam search is a standard search method for pruning low-probability search paths to decrease memory requirements; but the tradeoff involved in beam search is that of accidentally pruning a correct answer that seems unlikely at some time during search. The same issue arises in Bayesian modeling. The particle filter is a sequential Monte-Carlo method for finding the posterior distribution in complex statistical models (Doucet, Godsill, & Andrieu, 2000). Particle filters are an incremental

inference method, with the number of particles represented during inference corresponding to the number of hypotheses that are maintained during inference. With a very small number of particles, the memory demands of this inference method are limited, but like beam search schemes, the particle filter is limited in its ability to store initially low-probability hypotheses that turn out to be correct.

The general problem of a learning system with a short memory can be thought of as a problem with "Sherlock Holmes"-type inferences. As Holmes says, "when all other contingencies fail, whatever remains, however improbable, must be the truth," (Doyle, 1930). This kind of inference is impossible when the improbable possibilities have been forgotten ("pruned," in the language of search algorithms). From our current state of knowledge about the mind and brain, it seems potentially reasonable to assume that incremental learning is most useful for continuous, perceptual learning problems in which these Sherlock Holmes inferences rarely arise, while maintaining more of the input data—at least the most useful or perplexing examples—during learning is most useful in higher-level, more cognitive tasks. Nevertheless, this is an argument from intuition, not from empirical facts, and could be false or misleading. Because of these considerations I am unsure of the utility of the Cognitive and Time conditions.

The other four of Pinker's conditions on learning can be summarized easily. An LM should be able (a) to succeed in learning the appropriate parts of (b) any language (c) given the amount of input that children receive, and (d) it should make the same mistakes along the way that children make. These conditions can be consolidated into two criteria (Frank, Goldwater, Griffiths, & Tenenbaum, 2010): *sufficiency*—learning the right thing from the data—and *fidelity*—making the same mistakes along the way. These conditions are also easily mapped onto empirical tests of a proposed LM that is instantiated in a computational model. First, the model should converge to the right answer (whether it is an appropriate set of phonetic categories, a correct set of word-object mappings, or a set

of interpretations for sentences) given an appropriate sample of data—ideally from any one of a number of languages. Second, the model should fit human performance across a wide variety of experimental conditions, reproducing the different patterns of performance shown by children at different ages when it is given corresponding amounts of input data.[1]

In the review that follows, I use the conditions of sufficiency and fidelity as a guide in our evaluation of models and how they succeed and fail. However, there are two caveats to even these rough conditions for evaluation. First, models are most often proposed to capture a single learning task that children face, rather than to learn the entirety of a language. This decomposition of the learning task can be a useful tool—though it runs the risk of failing to take advantage of possible synergies between tasks (I return to this issue briefly at the end of the review)—but it can create situations where evaluating the output of a model is difficult. When a model solves a task that is intermediate along the way to a larger goal it may be difficult to evaluate a model on either of the proposed criteria. How are researchers to know what kind of performance would either be sufficient in the limit or faithful to human performance? The growing literature on artificial language learning tasks provides one partial solution to this issue, allowing models to be tested on their fit to what learners (often—somewhat problematically—adults, but sometimes children) can acquire from miniature languages. Thus, I include in this review a brief discussion of relevant artificial language results where appropriate.

Second, a model does not need to supersede previous work on these metrics to be important. In the words of the statistician George Box, "all models are wrong, but some are useful" (Box & Draper, 1987). The criteria proposed here should not be taken as a

[1]These conditions are similar but not identical to the conditions of *descriptive adequacy* and *explanatory adequacy* proposed by Chomsky (1965); for example, explanatory adequacy was not necessarily intended to encompass the pattern of successes and failures during learning. To minimize confusion, I have avoided using these terms.

suggestion that the only direction for new work should be in the positivistic improvement of performance or empirical coverage. Some models are good beginnings, providing framework ideas that can later be expanded dramatically; others are good demonstrations of principles that are simple enough to understand.

Computational models of language acquisition

In the following sections, I outline how the approach described above can be applied to models across a variety of domains of language acquisition and summarize the level of progress that has been achieved. A full review of progress in all areas of modeling language acquisition would be prohibitively long, so this review is necessarily somewhat selective. Wherever possible I have attempted to focus on those proposals that show particular promise in learning from corpus data or matching empirical work. I have divided the broader task of language acquisition into a list of sub-tasks: sound category learning, word segmentation, word learning, word-class learning, morphology learning, and syntactic rule learning. This classification is both ad-hoc—it reflects divisions in categories of models rather than being based on either clear psychological claims of modularity or the relative separability of tasks—and incomplete—it leaves out important areas such as prosody, pragmatics, and formal semantics because of the paucity of models in these areas. Nevertheless, it captures the majority of the active areas of modeling work.

Sound category learning

Learning the sound categories of their native language is the first step that children take towards acquiring their native language. Although human infants (Eimas, Siqueland, Jusczyk, & Vigorito, 1971) and other mammals (Kuhl & Miller, 1975) are sensitive to some of the consonant distinctions across the world's languages due to basic properties of their auditory system, as infants gain exposure to their native language, a variety of work has documented how they acquire language-specific vowel (Kuhl, Williams, Lacerda, Stevens,

& Lindbloom, 1992) and consonant (Werker & Tees, 1984) distinctions that other animals do not learn.[2] In addition, longitudinal studies have suggested that this learning process lays the groundwork for future language learning achievement (Tsao, Liu, & Kuhl, 2004).

Computational systems for recognizing sound categories in human speech have a long history and have grown quite sophisticated (Flanagan, 1972; Rabiner & Juang, 1993), but the canonical approach to speech recognition involves using *supervised* input—input that is tagged with category labels. In contrast, because learners do not know ahead of time which contrasts are meaningful, models of the acquisition of sound categories must be *unsupervised*: they must derive the categories from the data without being given labels by a designer. Hence, designers of computational models for sound categorization must look to developmental data for ideas about the learning process.

Experimental work provides the suggestion that infants can induce categories from the distribution of exemplars in acoustic space, via some type of probabilistic learning. Maye, Werker, and Gerken (2002) presented infants with phonetic tokens across a continuum and found that infants who heard unimodally-distributed stimuli for a short familiarization did not discriminate exemplars at the endpoints of the continuum, while those who heard bimodally-distributed exemplars did. Followup work suggested that this same paradigm could be used to facilitate 8-month-olds' discrimination of non-native contrasts (Maye, Weiss, & Aslin, 2008).

Recent models of sound category learning have made great progress in building on the available empirical data by investigating probabilistic learning mechanisms that use this clustering intuition (Boer & Kuhl, 2003; Vallabha et al., 2007; McMurray et al., 2009). Strikingly, all of these models have used variants of a mixture-of-Gaussians model:

[2]There are important distinctions between the perception of consonants and vowels—e.g., consonants are perceived categorically while vowels do show reduced discrimination at category boundaries but are perceived continuously—but for the purpose of this review we focus on similarities between the two.

a statistical technique that assumes that training data are generated independently from a set of Gaussian distributions and tries to recover the parameters of these distributions. It is highly unusual that nearly all of the modeling work on a particular phenomenon should converge on the same representational format, and may be indicative of a valuable convergence.

With respect to the sufficiency of these systems, the model of Vallabha et al. (2007) is the best example of a system that has the potential to scale up to the larger acquisition task. The Vallabha mixture model succeeded in learning vowel categories from multi-dimensional input generated from acoustic measurements of actual speakers. This success provides a valuable proof-of-concept and suggests the possibility of extensions to learning a larger space of sound categories from naturally-occurring data (rather than synthetic examples generated from natural measurements). In terms of fidelity, statistical models of this type have also shown considerable promise. Feldman, Griffiths, and Morgan (N. H. Feldman, Griffiths, & Morgan, 2009a) have used a related model of category discrimination that captures a large number of experimental findings on the perception of variable acoustic stimuli (including the "perceptual magnet effect," the tendency of vowel exemplars to be perceived as closer to the category center than they truly are; Kuhl et al., 1992). Work by Lake, Vallabha, and McClelland (2009) similarly applies the Vallabha model to the perceptual magnet effect.

To summarize work in this domain, there is a convergence of recent work on a single model class: variants of Gaussian mixture models. These models appear to have the potential to learn appropriate category structure (but for an account of possible weaknesses of these models on data with large numbers of overlapping categories, c.f. N. H. Feldman, Griffiths, & Morgan, 2009b). In addition, they appear to be able to account for a number of findings in the developmental and adult literatures on speech perception. Finally, a number of authors have explored online and neural net

approximations to these mixture models (Shi, Griffiths, Feldman, & Sanborn, in press; McMurray et al., 2009; Vallabha et al., 2007). Thus, in this domain computational models provide a guiding framework theory that strongly constrains hypotheses about the nature of early sound category learning.

Word segmentation

Although the boundaries between words are not marked by silences, there are a variety of language-specific cues such as stress, allophonic variation, and phonotactic constraints that are informative about where words begin and end (Jusczyk, 2000). Since these cues vary between languages, one proposal for a language-general strategy for segmentation is the use of statistical regularities in the occurrences of phoneme or syllable strings to find consistent linguistic units (Harris, 1951; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996). Work by Saffran, Aslin, and Newport (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996) suggested that infants and adults were able to identify frequent sequences of syllables from streams of continuous, monotonic speech with no prosodic cues.

Since they were first reported, findings on "statistical learning" have framed discussions of the role of probabilistic learning in language acquisition. In recent years the literature on statistical segmentation and related paradigms has blossomed, providing evidence that the same kind of segmentation is possible across a wide variety of domains and modalities (Kirkham, Slemmer, & Johnson, 2002; Fiser & Aslin, 2002; Saffran et al., 1999; Conway & Christiansen, 2005) and that species as diverse as tamarin monkeys (Hauser, Newport, & Aslin, 2001) and rats (Toro & Trobalon, 2005) can succeed in similar tasks. The sequences used in these tasks are typically designed to be simple enough for an experimental participant to learn the appropriate units after only a few minutes of exposure. As a consequence they can be solved effectively by a wide variety of possible

computational mechanisms (Frank et al., 2010).

In an early paper, Saffran, Newport, and Aslin (1996) suggested that success in statistical learning tasks could be accomplished via the computation of sequential transition probabilities (TPs) and the detection of local minima in TP. This proposal was supported by the suggestion that infants are able to succeed in statistical learning tasks even when raw frequencies do not distinguish coherent from incoherent sequences (Aslin, Saffran, & Newport, 1998). Since then, many authors have assumed (implicitly or explicitly) that pairwise statistics like TP, backwards TP (Perruchet & Desaulty, 2008), or mutual information (Swingley, 2005) are the mechanisms underlying human performance in statistical learning tasks (e.g. E. Johnson & Jusczyk, 2001; Tyler & Cutler, 2009; Endress & Mehler, 2009) and other language-learning phenomena (Thompson & Newport, 2007).

However, pairwise statistics like TP perform poorly on tests of sufficiency where they are evaluated on natural language corpus data (Brent, 1999a; Yang, 2004; Swingley, 2005). This poor performance has led some authors to suggest that statistical learning itself is unlikely to play a large part in language acquisition (Yang, 2004; Endress & Mehler, 2009) or may be a kind of bootstrapping cue that allows for the identification of more reliable cues (Swingley, 2005). Yet neither of these inferences, nor the inference that the computation of pairwise statistics underlies performance on statistical learning tasks, is supported by the data (Saffran, 2009). In principle, the computations underlying success in these simple paradigms could be much more complex, and hence could be sufficient for success in far more difficult tasks. The fact that a particular computational proposal (like sequential TP estimation) performs poorly may be evidence against that proposal, rather than evidence against the more general suggestion of a probabilistic computation in word segmentation.

Since these initial proposals, a range of other models of segmentation have been

described, including a heuristic, information-theoretic clustering model (Swingley, 2005); Bayesian lexical models (Brent, 1999b; Goldwater et al., 2009); and PARSER, a memory-based lexical model (Perruchet & Vinter, 1998, 2002). Both PARSER and the Bayesian models learn by assuming that unsegmented input was generated by combining a discrete set of words (a lexicon), which must then be recovered. Each provides different ways of balancing between the two edge-case lexicons: a too-long lexicon that includes each sentence in the input as a separate word, and a too-parsimonious lexicon that includes only the atoms of the language as words. PARSER uses memory mechanisms to make this tradeoff, retaining items in the lexicon that occur frequently enough not to decay out of usage. In contrast, Bayesian approaches apply a statistical tradeoff between the length of the lexicon itself (how many words there are) and the lengths of each individual word.[3] Nevertheless, they are highly related and perform similarly in comparisons to human data (Frank et al., 2010), suggesting a significant underlying unity.

Assessments of the sufficiency of these models in learning from corpus data have favored Bayesian lexical models (Brent, 1999b; Goldwater et al., 2009); models using this approach have also been adopted in the computational linguistic literature as the state of the art in segmentation across languages (e.g. Liang & Klein, 2009; M. H. Johnson & Goldwater, 2009). Investigators have also begun to examine the fidelity of different models of segmentation to human performance. Experiments in the auditory domain (Giroux & Rey, 2009) and in the visual domain (Orbán, Fiser, Aslin, & Lengyel, 2008) have both provided support for models of segmentation—like PARSER or the Bayesian lexical models—that posit the learning of discrete chunks (words in the auditory domain, objects in the visual) rather than transitions between syllables. In addition, Frank et al. (2010)

[3]These current models follow from an older tradition of work that addressed similar questions using heuristic techniques related to Minimum Description Length (Olivier, 1968; Wolff, 1975). For a summary of related approaches to segmentation and morphology learning, see Goldsmith (2010).

assessed the fidelity of a variety of models to experimental data in which systematic features of the speech input were varied (sentence length, number of word types, number of word tokens). They found that while all current models succeeded in learning the simple artificial languages, no models provided good fit to data without the imposition of memory constraints that limited the amount of data that was being considered. Supporting the importance of understanding the role of memory in segmentation, other results suggest that learners may not store the results of segmentation veridically, falsely interpolating memories that they have heard novel items that share all of their individual transitions with a set of observed items (Endress & Mehler, 2009).

Taken together, this work suggests an emerging consensus that lexical models—models that look for a small set of explanatory chunks—are most effective in segmentation. Challenges for future work are the extension of the current set of models (which mostly deal with transcribed data) to operate over noisy, acoustic data, and the integration of statistical models with other cues for segmentation. This second challenge is especially important. Unlike in sound category learning, where distributional cues are assumed to be the primary source of information, in segmentation a rich body of empirical work suggests not only that probabilistic learning interacts with language-specific acoustic cues like stress (E. Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003), but also that these language-specific cues can be acquired from a remarkably small amount of data (Thiessen & Saffran, 2007). Thus, a key part of creating successful models of word segmentation is understanding what part probabilistic learning plays in the acquisition and use of language specific cues.

Word learning

Given that many of the most successful grammatical frameworks in linguistics (Pollard & Sag, 1994; Steedman, 2000; Bresnan, 2001) and natural language processing

(Collins, 2003) are lexicalized (contain syntactic information that is linked to individual word forms), the majority of language acquisition could be characterized as "word learning." Inferring the meanings of individual lexical items—especially open-class words like nouns, adjectives, and verbs—is an important early challenge in language acquisition. As Quine (1960) observed, for any body of evidence about the use of a word, there are an infinite possible range of meanings that could fit the evidence exactly. Work on early word learning has recognized many ways that learners can overcome the problem of referential indeterminacy, from conceptual heuristics (Markman, 1991), to explicit social signals (Bloom, 2002), syntactic cues (Gleitman, 1990), or cross-situational associations (Yu & Ballard, 2007). Yet reviews of the state of the art in word learning sometimes read like a list of ad-hoc strategies, ending with an acknowledgment that future work needs to bring together disparate proposals into a more coherent framework (Snedeker, 2009). Computational models are a promising method for uniting these proposals. I describe models that address two aspects of the word learning problem: (a) matching words and meanings under ambiguity, and (b) generalization of word meanings from a limited set of examples.

Matching words with their meanings. A body of recent work has focused on the problem of matching words and their meanings when there are many words uttered and many candidate meanings present. Models of this problem have made use of the repeated observation of the co-occurrence of words and their meanings ("cross-situational observation"). An early model by Siskind (1996) provided a compelling demonstration that word meanings could be guessed by repeated observation and the application of deductive principles. Although other authors had speculated about the utility of cross-situational observation as a method for vocabulary acquisition (e.g. Gleitman, 1990; Pinker, 1984), Siskind's model provided a first quantification of the utility of this strategy. A body of experimental and theoretical work persuasively argues that this strategy is

most appropriate for learning nouns and that learning relational terms like verbs may require additional linguistic information (Gleitman, 1990; Gillette, Gleitman, Gleitman, & Lederer, 1999; Snedeker & Gleitman, 2004). Although Siskind used his system to learn even complex, relational meanings, the majority of the recent work in this area has focused on learning concrete nouns.

Despite this limitation, the cross-situational, cross-situational word learning models appear to be a promising framework for the integration of different information sources and strategies for word learning (Yu & Smith, 2007; Frank, Goodman, & Tenenbaum, 2009). Though the first model of this process used artificial data (Siskind, 1996), several models following up on that initial work have been applied to natural or naturalistic data (Roy & Pentland, 2002; Yu, Ballard, & Aslin, 2005). One recent model applied a machine translation algorithm for matching words across different languages to the problem of mapping words to objects that are present in the learner's field of view (Yu & Ballard, 2007). They coded the objects present in videos from CHILDES (MacWhinney, 2000) of mothers and children playing and used their system to match these with words in the sentences being spoken by the mothers. They found that such a system successfully identified correct word-object mappings. In addition, when they integrated manually-identified prosodic and social cues into their model, they found that these cues substantially increased their model's accuracy (Yu & Ballard, 2007).

A concern about these computational proposals is the possibility that they require memory and processing resources that are unavailable to human learners, especially infants. However, recent empirical investigations by Yu, Vouloumanos, and colleagues (Yu & Ballard, 2007; L. B. Smith & Yu, 2008; Vouloumanos, 2008; Vouloumanos & Werker, 2009) have given evidence that both adults and young children can use cross-situational exposure to learn associations between words and their meanings. This work provides a proof-of-concept that cross-situational learning is possible, and work is now emerging that

also attempts to characterize the mechanisms underlying these inferences in more detail (Yurovsky & Yu, 2008; Ichinco, Frank, & Saxe, 2009; Kachergis, Yu, & Shiffrin, 2009; Yurovsky, Fricker, Yu, & Smith, 2010). Analytical explorations have also provided some evidence for the viability of this type of strategy for acquiring large-scale lexicons (K. Smith, Smith, & Blythe, in press). Thus, it seems possible that the kind of cross-situational learning described by these computational models is not out of reach for human learners.

Although the models of cross-situational word learning described above show considerable promise in accounting for a range of phenomena, they neglect a crucial aspect of early word learning: its intentional character. Classic experiments demonstrate that from a very early age, word learners do not simply associate the words they hear with the objects in front of their eyes, but instead mediate these associations with their best guess at the speaker's intended referent (Baldwin, 1993; Akhtar, Carpenter, & Tomasello, 1996). To capture this idea, my colleagues and I introduced a probabilistic word learning model relying on the idea that learners are jointly trying to infer speakers' referential intention (what object they are talking about) and the meanings of the words that speakers utter (Frank, Goodman, & Tenenbaum, 2009). We found that our intentional model showed considerable improvement in the precision of the lexicons that were learned compared with the associative models. While the associative models learned many incorrect lexical associations, for example between function words and objects, the intentional model correctly rejected these spurious pairings. The intentional model also showed promise in accounting for empirical results. The model successfully predicted human performance in cross-situational word learning (Yu & Ballard, 2007), intentional word learning (Baldwin, 1993), and object individuation (Xu, 2002) experiments. These results suggest that integrating social aspects of word learning with the machinery to make statistical inferences could have the potential to account for a large variety of experimental data.

The phenomenon referred to as "mutual exclusivity" (ME) has been a key part of discussions of cross-situational word-meaning mapping. In a standard ME experiment, children are presented by an experimenter with a pair of toys, one novel and one familiar, and asked by the experimenter "give me a dax," where "dax" is a novel name. Across a wide variety of experimental procedures, children from the middle of their second year onward choose the novel object to go with the novel word (Markman & Wachtel, 1988; Halberda, 2003; Markman, Wasow, & Hansen, 2003). Though this phenomenon has primarily been explained in terms of lexical principles (like an language-specific bias to prefer one-to-one mappings in lexicons, Markman, 1991; Mervis & Bertrand, 1994) or pragmatic principles (E. Clark, 1988; Diesendruck & Markson, 2001), it can also be accounted for via inferences within a variety of probabilistic learning mechanisms. For example, the intentional model described above predicts mutual exclusivity and can generate novel predictions about mutual exclusivity in cross-situational word learning paradigms (Ichinco et al., 2009). However, other models also predict the same result: both an associative, exemplar-based model (Regier, 2005), and an incremental, probabilistic model similar to the translation model mentioned above (Fazly et al., 2008; Fazly, Alishahi, & Stevenson, in press) both were able to produce the basic mutual exclusivity phenomenon. These converging results suggest that, without having to posit separate lexical or pragmatic principles, general probabilistic learning may account for data on mutual exclusivity.

To summarize this section: models of word-meaning mapping have made great progress in recent years both in learning from corpus data and in fitting developmental data. However, a major weakness of all current studies is the limited scope of the data used for demonstrations of sufficiency. No existing model has been evaluated on a natural corpus longer than 20 minutes—in part because no such corpus exists. Thus, one important direction for future work in this area is the creation and annotation of

necessary data for evaluating models on their ability to learn words from interactions between parents and caregivers.

Generalizing word meanings. The problem of generalization is both a fundamental problem in cognitive science (Margolis & Laurence, 1999; Murphy, 2004) and a core part of the problem of word learning. Simplifying assumptions of the models discussed above to the contrary, knowing which word goes with which object in the current discussion (reference) does not imply knowledge of the *meaning* of the word that is being used. For example, sub- and super-ordinate labels like "animal" or "dalmation" can co-occur with more common, basic-level labels like "dog." In Quine's famous example, an anthropologist observes a tribesman labeling a rabbit as it runs by. Although the anthropologist assumes the word means "rabbit," Quine points out that the tribesman could equally well be saying "un-detached rabbit parts" or "momentary temporal part of an enduring rabbit." Work in infant development suggests that strong, shared notions of objecthood (Spelke, Breinlinger, Macomber, & Jacobson, 1992) and supports the notion that even very different human populations will likely share much of the same conceptual framework. Nevertheless, Quine's problem applies in more run-of-the-mill situations as well. Even if learners do not consider hypotheses like "un-detached rabbit parts" they still must consider sub- and super-ordinate labels like "white-tailed hare" and "animal."

An influential theory of generalization for early word learning suggested a set of principles which initially constrain children's inferences about the meanings of novel words, including considering only whole objects (rather than parts or properties) and considering only taxonomic categories (rather than e.g. thematic categories; Markman, 1991). As in the case of mutual exclusivity described above, recent work has begun to investigate whether learners could use make probabilistic inferences about what generalization best unites examples of a novel word. Xu and Tenenbaum (2007) described a simple Bayesian model of word meaning generalization that relies on the notion of a

suspicious coincidence. To take Quine's example above, if you see a single rabbit and it is labeled "gavagai," the word could mean "rabbit" or "animal"; if you see three rabbits pointed out to you as "gavagai" examples, but no other animals, it starts to seem improbable that the tribesman just happened to pick three examples of an animal that are exactly similar. Xu and Tenenbaum formalized this inference using the "size principle"—the idea that individual datapoints are probable under more specific hypotheses. Under this explanation, a very general hypothesis like "animal" should be disfavored unless it is the only hypothesis that fits.

A second suggestion for overcoming difficulties in word learning is the possibility that learners build up expectations about the kinds of ways that labels relate to concepts (L. Smith, 2000; L. Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). Children tend to use shape as the criterion for generalizing novel nouns (S. S. Jones, Smith, & Landau, 1991). Although this "shape bias" could be given innately, evidence suggests that it may be a learned expectation about category organization. L. Smith et al. (2002) trained children on novel categories that were organized around shape and found that at the end of training not only were the children able to use shape-based generalization in other novel categories, but their noun vocabulary had also increased considerably with respect to controls who did not receive the training. Both Bayesian (Kemp, Perfors, & Tenenbaum, 2007) and connectionist models (Colunga & Smith, 2005) predict these findings, suggesting that "overhypotheses"—distributions on the hypotheses likely to be true of a new set—like the shape bias can be learned quickly from data by learners with the appropriate representational capacity.

Work on word learning that uses statistical principles to overcome problems of generalization and referential indeterminacy has blossomed in the past decade. Increasingly, though, models of word learning have gone beyond simple associations between words and referents—or even words and concepts—to include social (Frank,

Goodman, & Tenenbaum, 2009; Yu & Smith, 2007), prosodic (Yu & Smith, 2007), and conceptual (Xu & Tenenbaum, 2007) information. The success of models taking these steps supports the view that probabilistic learning is not separate from the factors that have previously been identified. Instead, probabilistic learning may be the glue which holds these disparate kinds of information together and allows them to be used together in the service of learning words.

Word-class learning

Although tremendous progress has been made in tasks from learning speech sounds to learning words, progress at the highest levels of acquisition has been somewhat slower. Understanding how the structure of natural languages can be learned is a difficult challenge for both theories of language acquisition and for the applied fields of machine learning and natural language processing (NLP). Nevertheless, there has been considerable progress on important sub-problems like syntactic category learning, morphology learning, and verb argument-structure learning, and several recent systems show promise in more challenging fields, like grammatical inference.

A first step in acquiring a grammar is the extraction of syntactic categories (e.g. nonterminal categories in a context-free grammar like noun and verb). Despite the increasing emphasis on lexicalization—the use of syntactic representations that are tied to lexical content (Goldberg, 1995; Tomasello, 2003)—syntactic categories of some kind are largely agreed to be useful abstractions in characterizing the productivity of adult language. Evidence from adult language processing paradigms like syntactic priming supports the psychological reality of such categories (Bock, 1986) and recent evidence provides some support for this view in early child language (Thothathiri & Snedeker, 2008).

In NLP, learning syntactic categories from supervised (hand-labeled) data is largely

considered a solved problem (Jurafsky, Martin, & Kehler, 2000; Manning & Schütze, 2000), with performance very high on most measures. However, unsupervised learning of syntactic categories is not as simple. The output of such systems is a clustering of words into categories, but evaluating these categories is non-trivial. Although there have been a number of proposals for linking the gold-standard categories created by annotators to the categories found by unsupervised systems, there is no reason to assume that the output of such systems would be maximally correct if they did precisely match human annotations. Because we do not know the precise form of syntactic abstractions, we cannot say what the correct standard for the sufficiency of such a system should be.

Nevertheless, following initial suggestions by Maratsos and Chalkley (1980), a number of systems have addressed the challenge of unsupervised category induction using distributional information. For example, Redington and colleagues used a hierarchical clustering system that grouped words on the basis of their distributional context and recovered clusters that shared strong qualitative similarities with linguistic categorizations (Redington, Crater, & Finch, 1998). Other work has suggested that a number of different strategies, including minimum-description length clustering (Cartwright & Brent, 1997), clustering based on frequent contexts (Mintz, Newport, & Bever, 2002; Mintz, 2003), and Bayesian approaches (Goldwater & Griffiths, 2007; Parisien, Fazly, & Stevenson, 2008) all show relatively similar performance (Goldwater, 2007), suggesting that—at least at the highest level of granularity—word categories are relatively over-determined by the distributional data and can be learned through a number of different strategies.

In contrast, human results on "unsupervised" syntactic category learning have been mixed. Artificial language paradigms which should be amenable to simple distributional analyses have proven to be difficult for human learners. For example, the classic "MN/PQ" paradigm asks learners to acquire an artificial language whose sentences have either the form MN or PQ, where each letter represents an arbitrary class of nonsense

words or syllables. While this kind of learning is trivial for nearly any statistical model that posits word classes (e.g. a hidden Markov model), human learners tended to learn positional regularities (e.g. that M and P words came first in the sentence) rather than the abstract relation between categories (K. Smith, 1966).

Human learners only seem to succeed in finding category structure when there are multiple cues available. Braine (1987) showed evidence that distributional learning strategies could succeed in this task, but only when they were supplemented by additional referential or morpho-phonological information. Mintz (2002) showed that multiple distributional cues to category membership (e.g., a frame of two words rather than a single word) would allow learning (Mintz, 2002). Gerken et al. (2005) showed that 17-month-olds could learn a part of the Russian gender marking system, though only when some portion of the training stimuli were double-marked. Lany and Saffran (2010) even showed successful learning of categories by 18-month-olds using coordinated distributional and syllabic cues. One open explanation for this set of findings is that the pattern of failures in MN/PQ-style tasks may be due to learners' memory limitations and that adding coordinated cues may simply give extra cues for encoding (Frank & Gibson, 2011).

Although syntactic category acquisition has been a paradigm case for distributional learning (Maratsos & Chalkley, 1980), progress in this area has been hindered by the fact that a gold standard for syntactic categories is necessarily theory-based and cannot be uncontroversially derived from data. In addition, human experimental data are equivocal about whether distributional category learning is easy for human learners. One way in which distributional models can be evaluated more directly, however, is through the use of syntactic categories as an intermediate representation in another task (ideally one that can be compared directly to an uncontroversial gold standard). Thus, we suspect that further progress in this area will likely come through the use of categories in word-meaning mapping, the use of semantic information to extract sub-classes of words

(Alishahi & Stevenson, 2008), or the joint induction of categories and syntactic rules (Bannard et al., 2009).

Morphology learning

Morphological generalization has long been accepted as one of the methods for productivity in natural languages (Berko, 1958). Even before artificial language experiments demonstrated the plausibility of distributional learning strategies for aspects of language acquisition, the suitability of distributional strategies for morphological generalization was a topic of intense debate in studies of the English past tense. Early investigations using neural network models suggested that regularities in the frequencies of English verbs supported appropriate generalizations to novel forms (Rumelhart & McClelland, 1986). This work came under heavy criticism for its representational assumptions, generalization performance, and match to the empirical data, however (Pinker & Prince, 1988). Following on that initial investigation, Plunkett and Marchman investigated a broader range of connectionist systems for learning past tense forms (Plunkett & Marchman, 1991, 1993, 1996), which again elicited criticism for their match to empirical data (Marcus, 1995).

The controversy over the form of mental representations of past-tense morphology has had several positive outcomes, though, including an increased focus on fidelity to empirical data and a move towards the direct computational comparison of symbolic and associative views. Work by Albright and Hayes (2003) compared a purely analogical model of past-tense inflection to a model which used multiple stochastic rules of varying scopes. They found that the multiple-rules approach provided a better account of human generalizations in a nonce-word task than a pure similarity approach. The multiple-rules approach allowed the generalization system to capture the widely-varying scope of different rules (from non-generative exceptions like *went* to the fully general rules that

allow for regular inflection in novel forms like *googled*). The multiple-rule learner, though heuristic in nature, had a strong similarity to standard probabilistic clustering methods that can be used to model artificial language data as well, suggesting that this kind of expressive, rule-based approach might be broadly useful for modeling the acquisition of simple morphology-style regularities (Frank & Tenenbaum, 2010).

The unsupervised learning of morphological systems more generally has been a topic of interest in NLP. Given the wide diversity in morphological marking in the world's languages, computational systems for parsing in isolating languages like English will have only limited success when applied to morphologically-rich languages like Turkish. Systems for the induction of a morphological grammar from text thus play an important role the broader project of parsing text from these languages. Minimum-description length (MDL) formalisms have been used successfully for the induction of general morphological grammars (de Marcken, 1996). Goldsmith (2001) described a model based on this principle which searched for suffix morphology and identified linguistically-plausible analyses across a range of European languages. Unfortunately, although the MDL approach is highly general, full search for solutions in this formalism is intractable and so implemented systems must rely on a set of heuristics to find good descriptions.

Probabilistic systems using inference techniques such as Markov-chain Monte Carlo may offer a better alternative by allowing a full search of the posterior distribution over solutions. Recent work by Goldwater, Johnson, and colleagues (Goldwater, Griffiths, & Johnson, 2006; M. Johnson, Griffiths, & Goldwater, 2007; M. Johnson, 2008a) has made use of non-parametric Bayesian techniques to model the different processes underlying the generation of morphological rule types and the individual word tokens observed in the input. This dissociation of types and tokens allows for a more accurate analysis of the morphological rules that govern tokens. These new techniques present the possibility of unifying earlier work on the past tense with the broader project of learning morphology

from un-annotated data (Frank & Tenenbaum, 2010; O'Donnell, Tenenbaum, & Goodman, 2009).

Thus, the pattern in morphology learning is similar to those in other fields of acquisition. Initial computational work this area focused on simple, exemplar-based models of learning and generalization that computed simple statistics over the relations between datapoints. Issues with these strategies led more recent work to move towards a probabilistic framework that attempts to infer a parsimonious set of morphological descriptions within an expressive representational space (Albright & Hayes, 2003; M. Johnson et al., 2007). This work is still in its infancy, however, and little work has attempted to unite models which show sufficiency on larger corpora (Goldsmith, 2001) to those that show fidelity to human learning and generalization performance (Albright & Hayes, 2003).

Syntactic (and semantic) rule learning

Although there is still much work to be done, extracting the elements of language—phonemes, morphemes, words, and word categories—from distributional information is now largely assumed to be possible using statistical models. From the perspective of cognitive modeling, the major open challenge in this field is linking these statistical proposals to the abilities of human learners. In contrast, it is still unknown whether the *structural* features of language can be learned in the same way, or whether distributional learning mechanisms must be supplemented with other sources of information. Our last section reviews some of the heterogeneous literature on the learning of structured representations.

A large literature on learnability discusses the *a priori* possibility of a model that fulfills the sufficiency criterion for natural language syntax without assuming a large amount of structure.. The original learnability results in this field were by Gold (Gold et

al., 1967), with further investigation by a number of others (Horning, 1969; J. Feldman, 1972) (reviewed in Nowak, Komarova, & Niyogi, 2002). Discussion of this large and complex literature is outside of the scope of the current review. However, given the importance of these arguments, we note that while the mathematical results are clear, their applicability to the situation of children learning their native language is far from obvious (MacWhinney, 2004; A. Clark & Lappin, 2010). To take one example, the assumption of Gold's theorem is of an adversarial language teacher, who can withhold crucial examples for an infinite amount of time in order to derail the process of language acquisition. This assumption is strikingly different from the relationship that is normally assumed to hold between children and their caregivers (A. Clark & Lappin, 2010). Even if parents do not explicitly teach their children, they are unlikely to be adversarial in their use of syntax. More generally, the growth of systems for grammar induction has been so rapid, and their relationship to the assumptions of traditional "learnability in the limit" models is so complex, that we believe work on grammar induction should not be discounted on the basis of theoretical arguments (but c.f. Berwick & Chomsky, 2009). If we accept the possibility of success, then the development of novel techniques is important regardless of the sufficiency of the individual systems that initially instantiate these techniques.

In this vein, one of the most compelling early demonstrations of the power of statistical learning was by Elman (1990), who created a recurrent connectionist network that learned regularities in sequential artificial language data by the errors it made in predicting upcoming material. This network was only able to learn from small languages, but some work has attempted to translate these insights directly to much larger-scale systems with mixed results (Rohde, 2002). Although connectionist architectures have not generally proven efficient for large-scale language processing, the interest provoked by this proof of concept was considerable.

Unsupervised grammar induction has been a topic of persistent interest in NLP as well. Although the specific challenge of learning a set of correct rules from written, adult corpora is not directly comparable to the task of syntactic rule learning for children, this field still has the potential to contribute important insights. Early experiments for learning context-free grammars (CFGs) from plain-text representations were not highly successful (Carroll & Charniak, 1992; Stolcke & Omohundro, 1994), underperforming simple baselines like purely right-branching structures (Klein & Manning, 2005). More recent work has made use of related formalisms. Klein and Manning (2005) explored a model which induced constituency relationships (clusters) rather than dependencies between words and found increases over baseline. Clark and colleagues (A. Clark & Eyraud, 2006, 2007) have introduced efficiently-learnable formalisms that cover a large subset of the CFGs. The ADIOS system is related to both of these approaches via its clustering of related contexts; it uses a heuristic graph-merging strategy to perform scalable inferences over relatively large corpora (Solan, Horn, Ruppin, & Edelman, 2005). Taken together, these results suggest that it may be possible to circumvent learnability-in-the-limit results via formalisms that do not map directly to levels of the Chomsky hierarchy.

An interesting formal similarity exists between a number of these methods. Models by Klein and Manning (2005), Solan et al. (2005), and A. Clark and Eyraud (2006) all use similarities in distributional context to infer properties related to substitutability. The basic insight is that if strings x and y both occur in the context a_b, then they are at least partially substitutable and may be syntactically identical. These systems build on work on syntactic category acquisition (Redington et al., 1998) that uses distributional methods to merge items that occur in similar contexts. Work using these methods for sequence learning typically builds relatively item-specific syntactic rules (or analogous sequential regularities) and then merges them together based on various definitions of substitutability. These systems have shown a number of suggestive results, although more

work is necessary to understand how they relate to human performance given failures in comparable learning situations (e.g. the "MN/PQ" scenario described above).

Other work has attempted to unify insights from NLP with work on child language acquisition. For example, Perfors, Tenenbaum, and Regier (2010) used a Bayesian model-comparison approach to compare parsers of different formal expressivities on their overall complexity and fit to data when trained on a corpus of child-directed speech. They found that a CFG provided a smaller representation of the grammar than a finite-state grammar while still parsing sentences appropriately, suggesting that even a relatively small amount of input could allow a learner to conclude in favor of a more expressive formalism like a CFG over a simple linear representation of syntax. Although the Perfors system gave evidence in favor of such expressive representations, progress in learning grammars directly from child-directed speech has been limited.

Using insights from construction-based grammatical formalisms—which assume that children's initial syntactic representations may be centered around individual verbs rather than fully abstract grammars (Tomasello, 2003)—several recent systems have been applied to create models of children's productions. MOSAIC, an incremental, memory-based system, approximates this type of learning by memorizing parts of input strings that are congruent with primacy and recency factors (Freudenthal, Pine, Aguado-Orea, & Gobet, 2007). In addition, Borensztajn, Zuidema, and Bod (2008) and Bannard et al. (2009) created sophisticated probabilistic models that learn item-specific regularities, but both are complex systems incorporating novel formalisms. In the case of the Bannard et al. (2009) model, for example, the authors found that perplexity (prediction error) on children's early productions was decreased more by an item-based probabilistic context-free grammar (PCFG) and a traditional PCFG that did not contain item-specific information. The authors did not, however, present evidence that both learning models had actually learned adequate representations of the input data (in other words, that the

authors' search procedure had converged). Given the results described above describing the difficulties in unsupervised PCFG induction, more evidence will be necessary that both concrete and item-based grammars can be learned effectively from data.

Several psycholinguistically-inspired models have also attempted to link syntactic and semantic information, though these models have typically been more limited in the kinds of representations they posit. Early work on this topic was done by Kawamoto and McClelland (1987), who used a supervised neural network to identify the thematic roles associated with words in sentences. More recent work on this topic has been inspired by systems for semantic-role labeling in NLP, using animacy, sentence position, and the total number of nouns in a sentence to classify nouns as agents or patients (Connor, Gertner, Fisher, & Roth, 2008, 2009). Incorporating richer representations than the feature vectors used in previous work, a system by Alishahi and Stevenson (2008) learned verb classes and constructions from artificial corpora consisting of utterances and their associated thematic role information. Mirroring the development of children's productive use of verbs (Tomasello, 2003), they found that constructions gradually emerged through the clustering of different frames for using verbs. In addition, their model was able to simulate the generalization of novel verbs across a variety of experimental conditions. This body of work raises the intriguing possibility that children's early learning of language structure can be described better via semantic acquisition rather than the acquisition of fully-general syntactic rules.

Following this same idea, a number of groups have attempted to model natural language syntax and semantics jointly. Mooney and colleagues (Kate & Mooney, 2006; Wong & Mooney, 2007) have presented models based on discriminative learning techniques (e.g. support-vector machines) that attempt to learn parsers that directly translate natural language sentences into database queries. Other recent work has made use of combinatorial categorical grammar (CCG: Steedman, 2000), a linguistic framework

by which word order and logical forms are jointly derived from the same grammar. A series of systems now exist for learning CCG parsing systems that similarly identify the logical forms of natural language sentences (Zettlemoyer & Collins, 2005, 2007, 2009). Although these systems were not applied to data from acquisition (in large part due to the challenges of designing appropriate representations for sentential meaning in unrestricted contexts), they show considerable promise in unifying syntactic and semantic information in the service of sentence interpretation.

Recent work by Kwiatkowski and colleagues has taken further steps towards using child-language data (Kwiatkowski, Goldwater, & Steedman, 2009; Kwiatkowski, Zettlemoyer, Goldwater, & Steedman, 2010). The system presented in this work induces a CCG representation of the data and shows promising performance in parsing sentences from previous database query corpora and from sentences of child-language with their corresponding logical forms. Nevertheless, the logical forms underlying sentences in child language were derived directly from previous syntactic parses of these sentences; thus, the system assumes that the child *already* has access to some kind of valid syntactic trees corresponding to the input. This system is among the most promising current models of syntactic acquisition, but future work must solve the issue of appropriate training data before further progress can be made.

The human literature on learning rule-based structures in artificial languages is large and mixed, and unfortunately has made relatively little contact with the computational literature on grammar induction. On the one hand, there is a large recent literature on the ability of infants to learn identity-based regularities over short strings (e.g. "ABB", where A and B are distinct syllable classes; Marcus et al., 1999; Saffran, Pollak, Seibel, & Shkolnik, 2007; Marcus, Fernandes, & Johnson, 2007; Frank, Slemmer, Marcus, & Johnson, 2009). Although the representations necessary for success in these experiments are relatively impoverished, they nonetheless represent evidence that young

children can make inferences of the same type as those made by more sophisticated models of morphology and grammar learning (Frank & Tenenbaum, 2010).

On the other hand, there is an extensive literature on artificial grammar learning (AGL); although the majority of this work has been carried out with adults (Reber, 1967), some has also been conducted with infants (Gómez & Gerken, 1999; Saffran et al., 2008). The learning mechanisms underlying AGL have been studied for decades, and a full discussion of this literature is beyond the scope of this review (for more detailed discussion and an argument that statistical learning of the sort described in the section on word segmentation and AGL are parallel tasks, see Perruchet & Pacton, 2006). It is unknown whether the general mechanisms underlying AGL are involved in linguistic rule learning, though this point has been heavily debated (Lieberman, 2002). To date relatively few models of language acquisition have been applied directly (but cf. Perruchet & Vinter, 1998, 2002), though there is a parallel literature of models that apply only to AGL and not to language learning (Cleeremans & Dienes, 2008). An important task for future research is the application of models of language acquisition to AGL stimuli—a model that not only captured aspects of natural language learning but also the idiosyncratic phenomena of AGL would be an important advance in understanding the shared mechanisms of learning underlying success in these tasks.

To summarize: although work in the unsupervised learning of language structure is still in its infancy, there has nevertheless been a tremendous amount of progress in the last ten years. Recent developments have suggested that moving away from grammatical formalisms like CFGs to frameworks that fit natural language more closely can result in impressive progress (e.g. Kwiatkowski et al., 2009). Unfortunately, this work has not been as tightly connected to children's language acquisition or to artificial language results as work on sound category learning and word segmentation (for some exceptions, see e.g. Bannard et al., 2009; Alishahi & Stevenson, 2008; Kwiatkowski et al., 2009; Perfors et al.,

2010). Thus, important goals for future work on syntactic rule learning should be (1) the development of systems and experimental paradigms which allow direct links between human data and the learning performance of models, and (2) the creation of syntactically- and semantically-annotated corpora.

Synergies between tasks

The vast majority of the work that we have described here is confined to a single task like word segmentation or morphology learning. But there is no reason to believe that learners perform only one task at a time. In fact, it is very likely that children are learning over multiple timescales and across multiple tasks and representations. Our models, by focusing on a single timescale or a single task, may miss important synergies between tasks: opportunities where learning about one aspect of a problem may help in finding the solution to another (M. Johnson, 2008b).

Although work of this type is still in its infancy, there is some evidence that synergies in acquisition do exist. For example, N. H. Feldman et al. (2009b) created a model which both learns a set of lexical forms and learns speech categories. They found that these two tasks informed one another, such that performance in speech-category learning was considerably improved by the ability to leverage contrasting lexical contexts. Their work suggested that the space of English vowels may not be learnable via pure distributional clustering alone (e.g. mixture models like Toscano & McMurray, 2010), but instead may require this kind of joint lexicon learning.

A second example of using these kinds of synergies comes from work by Mark Johnson and colleagues, who proposed models that simultaneously segmented words from unsegmented input and learned the correspondences between words and objects. Compared with a text-only segmentation model, the joint model achieved better segmentation performance on referential words due to the ability of the model to cluster

those words based on their common referents (B. Jones, Johnson, & Frank, 2010). In addition, even greater synergies were found by a model that included the constraint that collocations (statistically coherent sequences) should include at most one referential word (M. Johnson, Demuth, Frank, & Jones, 2010).

Different tasks also operate over different timescales. Recent work on word learning has used two tasks to inform each other: sentence interpretation—which happens in the moment-by-moment of online interaction—and word-object mapping—which involves the aggregation of information over many different interactions. Models of word-object mapping that study the interplay between these two kinds of situations (Frank, Goodman, & Tenenbaum, 2009; McMurray et al., under review) suggest that synergies between the two timescales allow for better word learning and better fit to developmental phenomena such as the ability to use words for object individuation (Xu, 2002) and the decrease in reaction times in spoken word recognition across development (Fernald, Pinto, Swingley, Weinbergy, & McRoberts, 1998).

Although relatively little work to date has examined synergies of these types, research in this field is among the most important because it bridges across traditional boundaries between tasks in acquisition. These synergies also provide a crucial argument against approaches that make use of simple descriptive statistics: model-free statistics like co-occurrence are not able to capture how two independent tasks can nevertheless mutually inform one another. Together these findings suggest that there may be many more synergies between acquisition tasks that provide powerful leverage for language learners.

Expressive representations and parsimony biases

Chater and Vitányi (2003) consider how human learners solve the problem of induction: that any dataset is consistent with an infinity of possible generalizations. They

propose that the principle of simplicity, that "the cognitive system should prefer that pattern that gives the shortest code for the data," accounts for human inductive biases. Instantiating this simplicity principle mathematically leads to two closely related formalisms. The first is the minimum description length formalism, in which representations are preferred that are both themselves short and also efficiently compress the input data (Rissanen, 1983; Li & Vitányi, 2008).[4] The second is the Bayesian formalism, in which model complexity is balanced with the fit of the model to the input data (Tenenbaum & Griffiths, 2001; Gelman, 2004). Both formulations include a tradeoff between two terms: one that favors complex, expressive representations that compress the data with high efficiency, and one that favors parsimony in the representations that are learned.

Although models of language acquisition are stated in a wide range of formalisms, these two elements—expressive representations and parsimony bias—figure into nearly all successful models. For work in computational linguistics and machine learning, they are close to universal (and describing them comes close to describing common sense). Despite this, theorizing about probabilistic learning in cognitive science still often starts from the premise that what is important in probabilistic learning is "counting something": keeping track of frequencies or co-occurrences. This premise is not necessarily consistent with the elements above and may lead to poor inferences, given what we know about successful models (reviewed above).

Successful models of language acquisition tend to use representations that are sufficiently expressive to be able to compress the input data effectively. This notion of compression requires introducing some concepts from information theory. In information theory, any variable—e.g., the identity of the next word in a sentence or the position of

[4]These ideas have been applied directly to language acquisition, e.g. Brent and Cartwright (1996), Chater and Vitányi (2007), Goldsmith (2001).

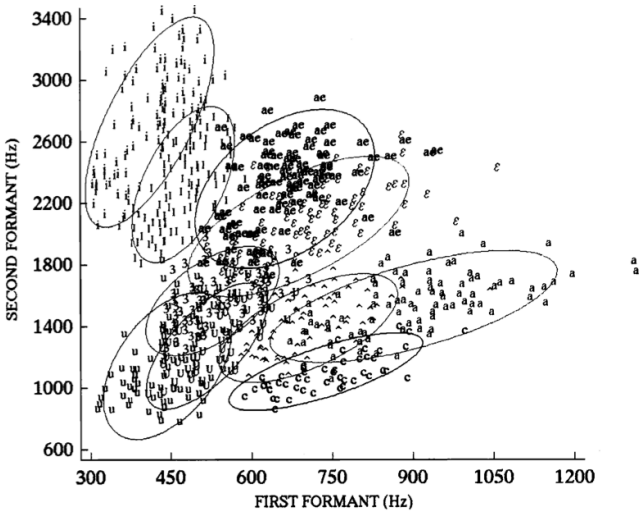

Figure 1. The distribution of vowel tokens in formant space. From "Acoustic Characteristics of American English Vowels" by J. Hillenbrand, L. A. Getty, M. J. Clark, & K. Wheeler, 1995, Journal of the Acoustical Society of America, 97, p. 3103.

the next phoneme in acoustic space—has an associated entropy. Entropy quantifies the uncertainty of that variable. If it is perfectly random, it has high entropy; if perfectly predictable, its entropy is zero. The flip side of entropy is redundancy: the lower the entropy of a variable is, the higher its redundancy. Redundancy can be eliminated via compression. A code (a re-representation of samples from that variable) can be used to write a series of samples from that variable in a more compact form. Codes can either be lossless, preserving all the information in a variable, or lossy, preserving only some information (Shannon, 1948; MacKay, 2003).

Under this construal, learning is the name we give to the process of finding a code for a particular variable. For example, the distribution of phonemes in acoustic space is highly non-random. Tokens of vowels plotted by their first and second formant energies are shown in Figure 1. If vowel tokens were perfectly random, they would cover this space

uniformly; instead each vowel has its own defined region. Learning phonetic categories can be thought of as learning a code that compresses samples from this acoustic space into a small alphabet of phonetic symbols. This code then allows for phonetic categorization and production in new situations.[5] Learning a grammar for how words are put together can likewise be thought of as learning a code for sentence structure. With such a code, it becomes possible to generate new sentences and to encode sentences more efficiently. At all levels of organization, language is non-random: it is characterized by a high degree of redundancy and hence there is a lot of room for compression (Shannon, 1951).

It is somewhat unusual to consider the products of language acquisition—phonetic categories, a lexicon, or a set of syntactic rules—as codes. In part this may be because these codes are not used independently from one another. For example, it seems more psychologically normal to compress sentences with respect to their meaning as opposed to their syntax. Despite this intuition, early experiments in the information theoretic paradigm did find effects of compression based on syntactic predictability (reviewed in Attneave, 1959) and recent psycholinguistic studies have revived this paradigm as a model of linguistic complexity (Levy, 2008). In addition, models of language acquisition describe the search for lossy representations of language structure (phonemes, words, grammars). The most successful of these models use codes that are highly expressive and hence allow for efficient compression of the data.

The generalization that good models of human learning make use of compressive representations seems close to tautological. Yet often discussions of language acquisition models implicitly assume representations that do not compress the input efficiently. For example, as reviewed above, many models of word segmentation presuppose the

[5]We give phoneme recognition as a simplified example. In practice, producing and recognizing meaningful phonetic strings requires duration and spectral information as well as information about co-articulation with other phonemes.

representation of large state-transition matrices keeping count of the transitions from syllable to syllable. These matrices can be justified as an algorithmic step along the way to the desired representation (a set of words in the language). Since such a matrix is actually a relatively poor compression of the regularities of syllable patterns, claiming such a step is akin to claiming that human learners use a sub-optimal representation for word segmentation. Such transition-probability based algorithms do not learn effectively from corpus data (Brent, 1999a) and they do not fit human performance (Frank et al., 2010), suggesting that they succeed on neither of the criteria outlined above. Instead, codes based on a list of words (a lexicon)—and perhaps even their dependency relationships— seem to be much more effective (Brent, 1999a; Goldwater et al., 2009).

There is a tradeoff between the complexity of a code and how well it can compress data of a particular type. For example, grammar-based codes can be highly efficient because of the complexity of the regularities in the data that they can capture (Kieffer & Yang, 2000, 2002). But for data that are known to have only regularities of more limited complexity, efficient coding can be achieved using a simpler formalism. The observation that successful models tend to use compressive representations should not be taken to mean that all models should have a degree of complexity in their representations beyond what is necessary to achieve optimal compression. In fact, successful models tend to incorporate some bias towards parsimony that limits the complexity of the representations that are learned.

A parsimony bias is in its essence, the imposition of some cost on learning such that if one thing is learned, another will not be. This kind of bias can be implemented in many different ways. In a Bayesian formalism, it is often implemented through the imposition of a prior probability distribution favoring simpler hypotheses (Goodman, Tenenbaum, Feldman, & Griffiths, 2008); in the minimum description length framework, a cost is assigned directly to the length of the coded representation (Goldsmith, 2001). In a

connectionist framework, some version of such a bias can be implemented through competition between hidden units (Rumelhart & Zipser, 1985). A parsimony bias can even be implemented directly in a co-occurrence matrix of the type mentioned above (Dayan & Kakade, 2000). The key is, however, that adding extra complexity to a particular instance of a code comes at a cost.

This tradeoff between appropriately expressive representations and a bias towards parsimony can be instantiated in a wide variety of different formalisms, but it is key to the success of many of the models reviewed above. For example, the Vallabha et al. (2007) analysis of phonetic category learning attempts to find a set of Gaussian categories that fit the observed data; it does so both by "assigning" datapoints to categories but also by using a competitive mechanism to prune categories that do not account for much of the data. Using a different formalism but a very similar principle, the Frank, Goodman, and Tenenbaum (2009) word learning model is able to avoid learning spurious word-object pairings by imposing a prior probability distribution on the size of the lexicon such that only those pairings which increase the model's ability to predict the corpus data are added. At yet a higher level of representational abstraction, Albright and Hayes (2003)'s minimal generalization learner uses a highly compressive representation (so-called "SPE" rules for describing inflectional morphology regularities) but selects specific sets of rules based on their scope (the number of cases they cover) and reliability (their accuracy on those cases).

In all three cases, these models are successful because they posit a relatively rich description of the input data that is therefore highly compressed. This compression is achieved both representations which go beyond the storage of individual datapoints (or even single summary statistics on those datapoints) and by the imposition of a bias to eliminate elements of the representation that do not serve to compress the data further.

Conclusions

We began by asking how children are able to learn the elements and structures of their native language. There is now a substantial body of experimental and computational evidence that statistical inference mechanisms play an important part in both of these tasks. Our review focused on the nature of the statistical inferences that best describe different aspects of language acquisition. Across the spectrum of learning tasks involved in language acquisition that we reviewed above, the models that performed best at learning from corpus data (sufficiency) and fitting human performance (fidelity) were not those that were framed in terms of simple distributional statistics. Instead, models that framed the problem as learning a parsimonious set of explanatory regularities like words, morphemes, categories, or rules—expressive units that allowed for efficient compression—were more successful.

References

Akhtar, N., Carpenter, M., & Tomasello, M. (1996). The role of discourse novelty in early word learning. *Child Development*, *67*, 635-645.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in english past tenses: a computational/experimental study. *Cognition*, *90*(2), 119-161.

Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science: A Multidisciplinary Journal*, *32*(5), 789–834.

Aslin, R. N., & Newport, E. (2008). What statistical learning can and can't tell us about language acquisition. In J. A. Colombo, P. McCardle, & J. Freund (Eds.), *Infant pathways to language: methods, models, and research directions.* Lawrence Erlbaum.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321-324.

Attneave, F. (1959). *Applications of information theory to psychology: A summary of basic concepts, methods, and results.* Holt, Rinehart and Winston.

Baldwin, D. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental psychology*, *29*(5), 832–843.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*(41), 17284.

Berko, J. (1958). The child's learning of english morphology. *Word*, *14*, 150–177.

Berwick, R., & Chomsky, N. (2009). *'Poverty of the stimulus' revisited: Recent challenges reconsidered.*

Bloom, P. (2002). *How children learn the meanings of words.* Cambridge, MA: MIT Press.

Bock, J. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387.

Boer, B. de, & Kuhl, P. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*(4), 129–134.

Borensztajn, G., Zuidema, W., & Bod, R. (2008). Children's grammars grow more abstract with age: evidence from an automatic procedure for identifying the productive units of language. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 47–51).

Box, G., & Draper, N. (1987). Empirical model-building and response surfaces.

Brady, T., Konkle, T., Alvarez, G., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325.

Braine, M. (1987). What is learned in acquiring word classes: A step toward an acquisition theory. *Mechanisms of language acquisition*, 65–87.

Brent, M. R. (1999a). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*(1), 71-105.

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

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*(1-2), 93–125.

Bresnan, J. (2001). *Lexical-functional syntax*. Wiley-Blackwell.

Carroll, G., & Charniak, E. (1992). Two experiments on learning probabilistic dependency grammars from corpora. In *Working notes of the workshop statistically-based nlp techniques* (pp. 1–13).

Cartwright, T., & Brent, M. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition*, *63*(2), 121–170.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in cognitive sciences*, *7*(1), 19–22.

Chater, N., & Vitányi, P. (2007). Ideal learning of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology*, *51*(3),

135–163.

Chomsky, N. (1965). *Aspects of the theory of syntax*. The MIT press.

Chomsky, N. (1975). *The logical structure of linguistic theory*. Springer.

Chomsky, N. (1981). Principles and parameters in syntactic theory. *Explanation in linguistics: The logical problem of language acquisition*, 32–75.

Clark, A., & Eyraud, R. (2006). *Learning auxiliary fronting with grammatical inference.*

Clark, A., & Eyraud, R. (2007). Polynomial identification in the limit of substitutable context-free languages. *Journal of Machine Learning Research*, *8*, 1725–1745.

Clark, A., & Lappin, S. (2010). *Linguistic nativism and the poverty of the stimulus*. Oxford, UK: Wiley Blackwell.

Clark, E. (1988). On the logic of contrast. *Journal of Child Language*, *15*, 317–335.

Cleeremans, A., & Dienes, Z. (2008). Computational models of implicit learning. *Cambridge handbook of computational psychology*, 396–421.

Collins, M. (2003). Head-driven statistical models for natural language parsing. *Computational linguistics*, *29*(4), 589–637.

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

Connor, M., Gertner, Y., Fisher, C., & Roth, D. (2008). *Baby SRL: Modeling early language acquisition.*

Connor, M., Gertner, Y., Fisher, C., & Roth, D. (2009). *Minimally supervised model of early language acquisition.*

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology Learning Memory and Cognition*, *31*(1), 24-3916.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, *24*(01), 87–114.

Dayan, P., & Kakade, S. (2000). Explaining away in weight space. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14.* Cambridge, MA: MIT Press.

de Marcken, C. (1996). Unsupervised language acquisition. *Arxiv preprint cmp-lg/9611002.*

Diesendruck, G., & Markson, L. (2001). Children's avoidance of lexical overlap: A pragmatic account. *Developmental Psychology*, *37*(5), 630–641.

Doucet, A., Godsill, S., & Andrieu, C. (2000). On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, *10*(3), 197–208.

Doyle, A. C. (1930). *The complete sherlock holmes.* Doubleday Books.

Eimas, P., Siqueland, E., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303.

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

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

Endress, A., & Bonatti, L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, *105*(2), 247–299.

Endress, A., & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, *60*(3), 351–367.

Fazly, A., Alishahi, A., & Stevenson, S. (2008). *A probabilistic incremental model of word learning in the presence of referential uncertainty.*

Fazly, A., Alishahi, A., & Stevenson, S. (in press). A probabilistic computational model of cross-situational word learning. *Cognitive Science.*

Feldman, J. (1972). Some decidability results on grammatical inference and complexity.

Information and control, *20*(3), 244–262.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009a). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological review*, *116*(4), 752–782.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009b). *Learning phonetic categories by learning a lexicon.*

Fernald, A., Pinto, J., Swingley, D., Weinbergy, A., & McRoberts, G. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*, *9*(3), 228.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, *99*(24), 15822-15826.

Flanagan, J. (1972). *Speech analysis, synthesis and perception.* Springer-Verlag New York.

Frank, M. C., & Gibson, E. (2011). Overcoming memory limitations in rule learning. *Language Learning and Development*, *7*(2), 130–148.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition.*

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578–585.

Frank, M. C., Slemmer, J. A., Marcus, G. F., & Johnson, S. P. (2009). Information from multiple modalities helps five-month-olds learn abstract rules. *Developmental science*, *12*(4), 504.

Frank, M. C., & Tenenbaum, J. (2010). Three ideal observer models for rule learning in simple languages. *Cognition.*

Freudenthal, D., Pine, J., Aguado-Orea, J., & Gobet, F. (2007). Modeling the developmental patterning of finiteness marking in english, dutch, german, and

spanish using mosaic. *Cognitive Science*, *31*(2), 311–341.

Gelman, A. (2004). *Bayesian data analysis*. CRC press.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, *32*(02), 249–268.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135–176.

Giroux, I., & Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive Science*, *33*, 260-272.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 3–55.

Gold, E., et al. (1967). Language identification in the limit. *Information and control*, *10*(5), 447–474.

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational linguistics*, *27*(2), 153–198.

Goldsmith, J. (2010). Segmentation and morphology. In C. Fox, S. Lappin, & A. Clark (Eds.), *The handbook of computational linguistics and natural language processing.* Wiley-Blackwell.

Goldwater, S. (2007). Distributional models of syntactic category acquisition: A comparative analysis. In *Workshop on psychocomputational models of language acquisition.* Citeseer.

Goldwater, S., & Griffiths, T. (2007). *A fully bayesian approach to unsupervised part-of-speech tagging.*

Goldwater, S., Griffiths, T., & Johnson, M. (2006). Interpolating between types and tokens by estimating power-law generators. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems 18* (pp. 459–466).

Cambridge, MA: MIT Press.

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

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, 431–436.

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

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*(1), B23–B34.

Harris, Z. S. (1951). *Methods in structural linguistics*. Chicago, IL: University of Chicago Press.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a human primate: statistical learning in cotton-top tamarins. *Cognition*, *78*, B53-B64.

Horning, J. (1969). *A study of grammatical inference*. Unpublished doctoral dissertation, Dept. of Computer Science, Stanford University.

Ichinco, D., Frank, M., & Saxe, R. (2009). Cross-situational word learning respects mutual exclusivity. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society.*

Johnson, E., & Jusczyk, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*(4), 548–567.

Johnson, M. (2008a). *Unsupervised word segmentation for sesotho using adaptor grammars.*

Johnson, M. (2008b). *Using adaptor grammars to identify synergies in the unsupervised*

acquisition of linguistic structure.

Johnson, M., Demuth, K., Frank, M. C., & Jones, B. K. (2010). Synergies in learning words and their referents. In *Advances in neural information processing systems.*

Johnson, M., Griffiths, T., & Goldwater, S. (2007). Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. *Advances in Neural Information Processing Systems*, *19*, 641.

Johnson, M. H., & Goldwater, S. (2009). Improving nonparameteric bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics* (p. 317-325).

Jones, B., Johnson, M., & Frank, M. C. (2010). Learning words and their meanings from unsegmented child-directed speech. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.*

Jones, S. S., Smith, L., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child development*, *62*(3), 499–516.

Jurafsky, D., Martin, J., & Kehler, A. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* MIT Press.

Jusczyk, P. (2000). *The discovery of spoken language.* The MIT Press.

Kachergis, G., Yu, C., & Shiffrin, R. (2009). *Frequency and contextual diversity effects in cross-situational word learning.*

Kate, R., & Mooney, R. (2006). *Using string-kernels for learning semantic parsers.*

Kawamoto, A. H., & McClelland, J. (1987). Mechanisms of sentence processing: Assigning roles to constituents of sentences. In *Parallel distributed processing, Vol. 2: Psychological and biological models* (pp. 195–248). Lawrence Erlbaum Associates.

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

Kieffer, J. C., & Yang, E. (2000). Grammar-based codes: a new class of universal lossless source codes. *IEEE Transactions on Information Theory*, *46*(3), 737.

Kieffer, J. C., & Yang, E. (2002). Structured grammar-based codes for universal lossless data compression. *Communications in Information and Systems*, *2*, 29–52.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, *83*, B35-B42.

Klein, D., & Manning, C. D. (2005). Natural language grammar induction with a generative constituent-context model. *Pattern Recognition*, *38*, 1407–1419.

Koerding, K., & Wolpert, D. (2004). Bayeisan integration in sensorimotor learning. *Nature*, *217*, 244–247.

Kuhl, P. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(22), 11850.

Kuhl, P. (2004). Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, *5*(11), 831–843.

Kuhl, P., & Miller, J. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, *190*(4209), 69.

Kuhl, P., Williams, K., Lacerda, F., Stevens, K. N., & Lindbloom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*, 606–608.

Kwiatkowski, T., Goldwater, S., & Steedman, M. (2009). Computational grammar acquisition from childes data using a probabilistic parsing model. In *Workshop on psycho-computational models of human language acquisition, at the 31st annual meeting of the cognitive science society.*

Kwiatkowski, T., Zettlemoyer, L., Goldwater, S., & Steedman, M. (2010). Inducing

probabilistic ccg grammars from logical form with higher-order unification. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1223–1233).

Lake, B., Vallabha, G., & McClelland, J. (2009). Modeling unsupervised perceptual category learning. *IEEE Transactions on Autonomous Mental Development*, *1*(1), 35-43.

Lany, J., & Saffran, J. (2010). From statistics to meaning. *Psychological Science*, *21*(2), 284.

Lenneberg, E. H. (1967). *Biological foundations of language*. New York: Wiley.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Li, M., & Vitányi, P. (2008). *An introduction to kolmogorov complexity and its applications*. Springer-Verlag New York Inc.

Liang, P., & Klein, D. (2009). Online EM for Unsupervised Models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 611–619).

Lieberman, P. (2002). *Human language and our reptilian brain: the subcortical bases of speech, syntax, and thought*. Cambridge, MA: Harvard University Press.

Ma, W., Beck, J., Latham, P., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. Third Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2004). A multiple process solution to the logical problem of language acquisition. *Journal of Child Language*, *31*(04), 883–914.

Manning, C. D., & Schütze, H. (2000). *Foundations of statistical natural language processing*. MIT Press.

Maratsos, M., & Chalkley, M. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. *Children's language*, *2*, 127–214.

Marcus, G. F. (1995). The acquisition of the english past tense in children and multilayered connectionist networks. *Cognition*, *56*(3), 271–279.

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

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77.

Margolis, E., & Laurence, S. (1999). *Concepts: core readings*. The MIT Press.

Markman, E. M. (1991). *Categorization and naming in children: Problems of induction*. The MIT Press.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*, 121–157.

Markman, E. M., Wasow, J., & Hansen, M. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, *47*(3), 241–275.

Marr, D., & Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *204*(1156), 301–328.

Maye, J., Weiss, D., & Aslin, R. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*(1), 122.

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

McMurray, B., Aslin, R., & Toscano, J. (2009). Statistical learning of phonetic categories:

insights from a computational approach. *Developmental science*, *12*(3), 369.

McMurray, B., Horst, J. S., & Samuelson, L. K. (under review). Using your lexicon at two timescales: Investigating the interplay of word learning and recognition.

Mervis, C., & Bertrand, J. (1994). Acquisition of the novel name-nameless category (n3c) principle. *Child Development*, *65*(6), 1646–1662.

Mintz, T. (2002). Category induction from distributional cues in an artificial language. *Memory and Cognition*, *30*, 678–686.

Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117.

Mintz, T., Newport, E., & Bever, T. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science: A Multidisciplinary Journal*, *26*(4), 393–424.

Murphy, G. (2004). *The big book of concepts*. The MIT Press.

Nowak, M., Komarova, N., & Niyogi, P. (2002). Computational and evolutionary aspects of language. *Nature*, *417*(6889), 611–617.

O'Donnell, T. J., Tenenbaum, J. B., & Goodman, N. D. (2009). *Fragment grammars: Exploring computation and reuse in language* (Tech. Rep. No. CSAIL-TR-2009-013). Massachusetts Institute of Technology.

Olivier, D. (1968). *Stochastic grammars and language acquisition devices*. Unpublished doctoral dissertation, Ph. D. thesis, Harvard University.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*, 2745–2750.

Parisien, C., Fazly, A., & Stevenson, S. (2008). *An incremental Bayesian model for learning syntactic categories.*

Pearl, L., & Lidz, J. (2009). When domain-general learning fails and when it succeeds:

Identifying the contribution of domain specificity. *Language Learning and Development*, *5*(4), 235–265.

Pena, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*(5593), 604.

Perfors, A., Tenenbaum, J., & Regier, T. (2006). *Poverty of the stimulus? a rational approach.*

Perfors, A., Tenenbaum, J., & Regier, T. (2010). The learnability of abstract syntactic principles. *Cognition*.

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

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*(5), 233–238.

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

Perruchet, P., & Vinter, A. (2002). The self-organizing consciousness as an alternative model of the mind. *Behavioral and Brain Sciences*, *25*, 360–380.

Pinker, S. (1979). Formal models of language learning. *Cognition*, *7*(3), 217–283.

Pinker, S. (1984). *Language learnability and language development.* Cambridge, MA: Harvard University Press.

Pinker, S. (1995). *The language instinct: The new science of language and mind.* Penguin London.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Connections and symbols*, 73–193.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition*,

38(1), 43–102.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, *48*(1), 21–69.

Plunkett, K., & Marchman, V. (1996). Learning from a connectionist model of the acquisition of the english past tense. *Cognition*, *61*(3), 299–308.

Pollard, C., & Sag, I. (1994). *Head-driven phrase structure grammar*. University of Chicago Press.

Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, *1*(2), 125–132.

Quine, W. (1960). *Word and object*. The MIT Press.

Rabiner, L., & Juang, B. (1993). Fundamentals of speech recognition. *Englewood Cliffs, NJ*.

Reber, A. (1967). Implicit learning of artificial grammars1. *Journal of verbal learning and verbal behavior*, *6*(6), 855–863.

Redington, M., Crater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science: A Multidisciplinary Journal*, *22*(4), 425–469.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science: A Multidisciplinary Journal*, *29*(6), 819–865.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The Annals of statistics*, *11*(2), 416–431.

Rohde, D. (2002). *A connectionist model of sentence comprehension and production*. Unpublished doctoral dissertation, Carnegie Mellon University.

Roy, D., & Pentland, A. (2002). Learning words from sights and sounds: a computational model. *Cognitive Science*, *26*, 113–146.

Rumelhart, D. E., & McClelland, J. L. (1986). Learning the past tenses of english verbs:

Implicit rules or parallel distributed processing. In *Parallel distributed processing, Vol. 2: Psychological and biological models* (pp. 195–248). Cambridge, MA: MIT Press.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, MA: MIT Press.

Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning*. *Cognitive Science*, *9*(1), 75–112.

Saffran, J. R. (2009). What is statistical learning, and what statistical learning is not. In S. P. Johnson (Ed.), *Neoconstructivism: The new science of cognitive development.* Oxford University Press.

Saffran, J. R., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926.

Saffran, J. R., Hauser, M., Seibel, R., Kapfhamer, J., Tsao, F., & Cushman, F. (2008). Grammatical pattern learning by human infants and cotton-top tamarin monkeys. *Cognition*, *107*(2), 479–500.

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

Saffran, J. R., Newport, E., & Aslin, R. (1996). Word segmentation: The role of distributional cues. *Journal of memory and language*, *35*(4), 606–621.

Saffran, J. R., Pollak, S., Seibel, R., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition*, *105*(3), 669–680.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–665.

Shannon, C. (1951). Prediction and entropy of printed english. *Bell System Technical Journal*, *30*(1), 50–64.

Shi, L., Griffiths, T., Feldman, N., & Sanborn, A. N. (in press). Exemplar models as a mechanism for performing bayesian inference. *Psychonomic Bulletin and Review*.

Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39-91.

Smith, K. (1966). Grammatical intrusions in the recall of structured letter pairs: mediated transfer or position learning? *Journal of Experimental Psychology*, *72*, 580–588.

Smith, K., Smith, A. M., & Blythe, R. A. (in press). Cross-situational word learning: mathematical and experimental approaches to understanding tolerance of referential uncertainty. *Cognitive Science*.

Smith, L. (2000). Learning how to learn words: An associative crane. *Becoming a word learner: A debate on lexical acquisition*, 51–80.

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.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Snedeker, J. (2009). Word learning. In L. Squire (Ed.), *Encyclopedia of neuroscience* (pp. 503–508). Elsevier.

Snedeker, J., & Gleitman, L. (2004). Why it is hard to label our concepts. *Weaving a lexicon*, 257–294.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(33), 11629.

Spelke, E., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99*(4), 605–632.

Steedman, M. (2000). *The syntactic process*. MIT Press.

Stolcke, A., & Omohundro, S. (1994). Inducing probabilistic grammars by bayesian model merging. *Grammatical Inference and Applications*, 106–118.

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

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

Thiessen, E., & Saffran, J. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Developmental Psychology*, *39*(4), 706–716.

Thiessen, E., & Saffran, J. (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development*, *3*(1), 73–100.

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

Thothathiri, M., & Snedeker, J. (2008). Syntactic priming during language comprehension in three-and four-year-old children. *Journal of Memory and Language*, *58*(2), 188–213.

Todorov, E. (2009). Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, *106*(28), 11478.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Toro, J. M., & Trobalon, J. B. (2005). Statistical computations over a speech stream in a rodent. *Perception and Psychophysics*, *67*(5), 867-875.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*, 434–464.

Tsao, F., Liu, H., & Kuhl, P. (2004). Speech perception in infancy predicts language development in the second year of life: a longitudinal study. *Child Development*, *75*(4), 1067–1084.

Tyler, M., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *The Journal of the Acoustical Society of America*, *126*, 367.

Vallabha, G., McClelland, J., Pons, F., Werker, J., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*(33), 13273.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*(2), 729–742.

Vouloumanos, A., & Werker, J. (2009). Infants' learning of novel words in a stochastic environment. *Developmental psychology*, *45*(6), 1611–1617.

Waxman, S., & Gelman, S. (2009). Early word-learning entails reference, not merely associations. *Trends in cognitive sciences*.

Werker, J., & Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63.

Wexler, K., & Culicover, P. (1983). *Formal principles of language acquisition*. Cambridge, MA: MIT Press.

Wolff, J. (1975). An algorithm for the segmentation of an artificial language analogue. *British Journal of Psychology*.

Wong, Y., & Mooney, R. (2007). *Learning synchronous grammars for semantic parsing with lambda calculus.*

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, *85*(3), 223–250.

Xu, F., & Tenenbaum, J. (2007). Word Learning as Bayesian Inference. *Psychological*

Review, *114*, 245.

Yang, C. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences*, *8*(10), 451–456.

Yu, C., & Ballard, D. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*, 2149–2165.

Yu, C., Ballard, D., & Aslin, R. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science: A Multidisciplinary Journal*, *29*(6), 961–1005.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.

Yurovsky, D., Fricker, D., Yu, C., & Smith, L. B. (2010). The active role of partial knowledge in cross-situational word learning. In *Proceedings of the 32nd annual conference of the cognitive science society.*

Yurovsky, D., & Yu, C. (2008). *Mutual exclusivity in crosssituational statistical learning.*

Zettlemoyer, L., & Collins, M. (2005). *Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars.*

Zettlemoyer, L., & Collins, M. (2007). *Online learning of relaxed ccg grammars for parsing to logical form.*

Zettlemoyer, L., & Collins, M. (2009). Learning context-dependent mappings from sentences to logical form. In *Proceedings of the Association for Computational Linguistics* (pp. 976–984).