Early Word Learning Through Communicative Inference

by

Michael C. Frank

B.A. and B.S., Stanford University (2004)

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Cognitive Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2010

© Massachusetts Institute of Technology 2010. All rights reserved.

Author . . . .
Department of Brain and Cognitive Sciences
August 31st, 2010

Certified by . . . .
Edward Gibson
Professor of Cognitive Sciences
Thesis Supervisor

Accepted by . . . .
Earl K. Miller
Picower Professor of Neuroscience
Director, Brain and Cognitive Sciences Graduate Program

Early Word Learning Through Communicative Inference

by

Michael C. Frank

Submitted to the Department of Brain and Cognitive Sciences
on August 31st, 2010, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Cognitive Science

Abstract

How do children learn their first words? Do they do it by gradually accumulating information about the co-occurrence of words and their referents over time, or are words learned via quick social inferences linking what speakers are looking at, pointing to, and talking about? Both of these conceptions of early word learning are supported by empirical data. This thesis presents a computational and theoretical framework for unifying these two different ideas by suggesting that early word learning can best be described as a process of joint inferences about speakers' referential intentions and the meanings of words.

Chapter 1 describes previous empirical and computational research on "statistical learning"—the ability of learners to use distributional patterns in their language input to learn about the elements and structure of language—and argues that capturing this ability requires models of learning that describe inferences over structured representations, not just simple statistics. Chapter 2 argues that social signals of speakers' intentions, even eye-gaze and pointing, are at best noisy markers of reference and that in order to take advantage of these signals fully, learners must integrate information across time. Chapter 3 describes the kinds of inferences that learners can make by assuming that speakers are informative with respect to their intended meaning, introducing and testing a formalization of how Grice's pragmatic maxims can be used for word learning. Chapter 4 presents a model of cross-situational intentional word learning that both learns words and infers speakers' referential intentions from labeled corpus data.

Thesis Supervisor: Edward Gibson
Title: Professor of Cognitive Sciences

This thesis is dedicated to my fiancée Alison Kamhi: my partner, my supporter, and my love.

Acknowledgments

My parents, Lucy and Peter Frank, have given me their love and encouragement in everything that I've done over my entire life. Thank you both for being the best parents I can imagine.

My advisor, Ted Gibson, has been a mentor, advocate, and role model from the very first day of graduate school. Josh Tenenbaum has pushed my intellectual boundaries at the same time as he has been a fantastic collaborator. Rebecca Saxe has provided a sounding board for my crazy ideas and broadened my perspective immensely. All three have been unstintingly generous with their time and houses. I've felt incredibly lucky to have worked with faculty that have been both intellectual companions and friends.

I have had amazing mentors outside of MIT as well. Anne Fernald showed me the responsibilities of a scientist to the community and gave me an education in the history of developmental psychology over the course of three wonderful summers. Lera Boroditsky gave me my start and passed on her sense of just how cool psychology can be. Michael Ramscar pushed me in the right direction when I badly needed a push and got me hooked on language acquisition. Scott Johnson inspired me with his commitment to understanding development.

Many ideas in this thesis were developed over the course of runs, cups of coffee, and videoconferences with Noah Goodman. The broader "wurwur" project owes its existence to a couple of weeks spent hacking together in the summer of 2007. The other people who collaborated on parts of this broader project also deserve many thanks: Anne Fernald, Avril Kenney, Peter Lai, and especially Josh Tenenbaum. They have given a tremendous amount of time and intellectual energy to this work.

The last five years at MIT have been some of the most fun I've ever had. The folks in my cohort—Ed Vul, Talia Konkle, and Vikash Mansinghka especially—have inspired me and pushed me over and over again, and I am incredibly grateful for their support, help, and conversation over the years. I've learned more from you guys than from any class. I also want to thank my friends in the Cognitive group past and

present more generally for an incredible graduate experience. This list includes folks from Tedlab: Mara Breen, Ev Fedorenko, Steve Piantadosi, John Kraemer; from Cocosci: Charles Kemp, Amy Perfors, Lauren Schmidt, Liz Bonawitz, Pat Shafto, among others; and from other labs: Tim Brady, George Alvarez, Liane Young, Marina Bedny, and many others that I can't even fit here. Thanks for making BCS the wonderful community it's been and continues to be!

Contents

List of Figures

List of Tables

Chapter 1

The role of statistical inference in early language acquisition

1.1 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 (D. E. Rumelhart, McClelland, & 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; G. F. Marcus, Vijayan, Bandi Rao, & Vishton, 1999; Gómez, 2002; Gerken, Wilson, & Lewis, 2005) combined with promising computational results (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 plays a large part in language acquisition. 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 of these mechanisms and their application to more complex aspects of language learning is still controversial. Although statistical 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 is even more controversial (Pinker, 1979; Wexler & Culicover, 1983). Thus, despite the progress that has been made, many questions regarding the role of probabilistic learning in language acquisition are still unanswered.

The goal of this review is to survey computational studies of early language acquisition across the full range of acquisition challenges, from sound-category learning to syntactic rule learning. We describe criteria for evaluating these models on their adequacy as theories of language acquisition. Applying these criteria reveals that there are considerable similarities between many of the most successful models across a range of tasks. Probabilistic learning has often been characterized as implying that learners compute and apply simple descriptive statistics like co-occurrence and conditional probability. We find that simple statistics do not generalize well from task to task. We argue instead that the most successful computational proposals across tasks attempt to infer a parsimonious description of the data in a structured representational vocabulary.

1.2 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? This same question was taken up by Pinker (Pinker, 1979), who proposed six empirical conditions:

1. Learnability condition: LM must be able to acquire a language in the limit
2. Equipotentiality condition: LM must be able to learn any human language
3. Time condition: LM must be able to learn within the same amount of time as the child
4. Input condition: LM must be able to learn given the same amount of input as the child
5. Developmental condition: LM must make predictions about the intermediate stages of learning
6. Cognitive condition: 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, they are difficult to evaluate.

If the time condition is distinct from the input condition, it is not clear how it should be evaluated: The mapping between time and computation cycles in a serial computer is not direct, and the mapping function is unknown. Likewise, although the cognitive condition appears compelling, we know very little about the true computational abilities of children. If we were to take such a limitation seriously, then the kinds of computations that go on at every level of processing during tasks like object recognition or motor control would seem far out of reach of child learners (Pouget, Dayan, & Zemel, 2000; Todorov, 2009). Applied indiscriminately, this condition runs the risk of using researchers' intuitions to limit the kinds of LMs that we consider. If those intuitions were incorrect, researchers would not even consider the appropriate mechanism. Thus, although the time and cognitive conditions may have an eventual theoretical purpose, they should not be included in any comparison of current models.

The other four conditions can easily be summarized: a LM should be able to succeed in learning the appropriate parts of any language given the amount of input that children receive, and it should make the same mistakes along the way that children make. We consolidate these conditions into two criteria (Frank, Goldwater, Griffiths, & Tenenbaum, in press): *sufficiency*—learning the right thing from the data—and *fidelity*—making the same mistakes along the way. These conditions are 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.

Models are often proposed to capture a single learning task that children face, rather than to learn the entirety of a language. For example, a model of syntax acquisition is usually assumed to receive as input utterances that are already segmented into words (Bannard et al., 2009); a model of quantifier semantics is assumed to know some lexical semantics already (Piantadosi, Goodman, Ellis, & Tenenbaum, 2008). 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 (we 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 can acquire from miniature languages. Thus, we include in this review a brief discussion of relevant artificial language results where appropriate.

1.3 Application of statistical inference to language

In the following sections, we will 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 the review is necessarily highly selective. Wherever possible we have attempted to focus on those proposals that show particular promise in learning from corpus data or matching empirical work. We divide the broader task of language acquisition into a list of subtasks, including sound category learning, word segmentation, word learning, word class learning, morphology learning, and syntactic rule learning.

1.3.1 Sound category learning

A rich literature describes the progression of infants in learning the sound categories of 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, they gradually acquire language specific vowel (Kuhl, Williams, Lacerda, Stevens, & Lindbloom, 1992) and consonant (Werker & Tees, 1984) distinctions that other animals do not learn.

How do infants infer which particular sound contrasts are meaningful in their native language? Experimental work provided the suggestion that they could induce categories from the distribution of exemplars in acoustic space (Maye, Werker, & Gerken, 2002). Maye, Werker, and Gerken 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 computational modeling applies similar techniques for unsupervised category induction to cross-linguistic speech-category learning (Vallabha, McClelland, Pons, Werker, & Amano, 2007). This work suggests that by assuming some set of categories that each produce a distribution of exemplars within an acoustic feature space it is possible to work backwards and infer the particular categories present in some input set.[1] Feldman, Griffiths, and Morgan (N. Feldman, Griffiths, & Morgan, 2009a) have recently proposed a model of category learning that captures a large number of experimental findings on the perception of variable acoustic stimuli (including the "perceptual magnet effect" (Kuhl et al., 1992)). In addition, work by McMurray and colleagues (McMurray, Aslin, & Toscano, 2009; Toscano & McMurray, 2010) elaborates on the use of similar unsupervised mixture modeling techniques to learn phonetic categories incrementally from data. Both of these approaches focus on demonstrating both sufficiency—acquisition of correct, adult-like categories—and fidelity—fit to human perception results.

In the domain of speech category learning it has long been known that infants learn the structure of their native language from exposure to noisy phonetic tokens (Werker & Tees, 1984), but current research is beginning to provide descriptions of this process that map from observed input tokens to a category representation that explains effects in speech perception. Although previously described as a "warping" of the perceptual space (Kuhl, 2000), it now seems clear that the best-performing models assume some discrete category structure that is being recovered from noisy input. Rather than applying "simple statistics" to the input data, infants' categorization is

[1] 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.

best characterized as inference to a parsimonious set of explanatory variables.

1.3.2 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. They proposed that learners might compute probabilities between syllables and then look for dips in transition probability (TP) to signal word boundaries. Followup experiments by Aslin, Saffran, and Newport (R. N. Aslin, Saffran, & Newport, 1998) provided support for that idea by showing that infants were able to distinguish between statistically-coherent (high internal TP) sequences from incoherent (low internal TP) sequences, even when they were balanced for frequency.

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, Johnson, Aslin, & Newport, 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. Although many of these experiments were interpreted in terms of the computation of TPs, there are a range of computational proposals that explain segmentation performance, including a heuristic, information-theoretic clustering model (Swingley, 2005); Bayesian lexical models which look for a parsimonious lexicon that best explains the speech input (Brent, 1999; Goldwater et al., 2009); and PARSER, a memory-based model that sequentially incorporates chunks of input into a lexicon (Perruchet & Vinter, 1998, 2003). Assessments of the sufficiency of these models in learning from corpus data have favored the Bayesian lexical models (Brent, 1999; Goldwater et al., 2009) over transition probability-based approaches (Yang, 2004). Lexical approaches 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 how empirical evidence from statistical segmentation paradigms can decide between models of segmentation. 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, Goldwater, Griffiths, and Tenenbaum (Frank et al., in press) 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 paints a picture of segmentation—like sound category learning—as inference to a parsimonious set of items or categories. Again, simple statistical approaches like transition probability do not seem to fit performance or succeed in acquisition in the same way as more sophisticated models that assume inference to a set of explanatory variables (lexical items, in this case, rather than phonetic categories). One challenge for future work in this area is that unlike in sound category learning, none of the models under consideration deal well with noise or variability in exemplars, and none are incremental. A second challenge is the integration of statistical models with other cues for segmentation. A rich body of empirical work suggests not only that statistical information 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).

1.3.3 Word learning

Given the lexicalized form of many successful grammatical frameworks in both linguistic syntax (e.g. Pollard & Sag, 1994; Steedman, 2000; Bresnan, 2001) and natural language processing (e.g. Collins, 2003), the majority of language acquisition could arguably 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. One primary focus in this literature is overcoming problems of referential indeterminacy (Quine, 1960): the fact that any given use of a word in context, no matter how specific, can support many possible interpretations.

Experimental and computational work has focused on overcoming one aspect of this problem—figuring out which words in a sentence have which meanings—through repeated observation of the co-occurrence of words and their meanings ("cross-situational observation"). An early model by Siskind (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. Work by Gleitman and colleagues 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). Thus, 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.

Drawing on the competition-based lexical models of MacWhinney (MacWhinney, 1989), Regier (Regier, 2005) presented a connectionist, exemplar-based model that learned mappings of words and objects that were presented via distributed, featural representations. Though this model was not applied to learning words from annotated corpus data, the general computational principles he incorporated allowed the model to fit a number of results from the literature on early word learning. The connectionist models of Li and colleagues are also related, though they primarily focus on understanding the grouping dynamics of the growing vocabulary via the use of self-organizing maps (Li & Farkas, 2002; Li, Zhao, & Whinney, 2007).

Several models have also been applied to learning words from natural or naturalistic data. Roy and Pentland (Roy & Pentland, 2002) created a system for manipulating multi-modal input. Yu and colleagues (Yu, Ballard, & Aslin, 2005; Yu & Ballard, 2007) described a system based on a model of machine translation which learned mappings between words and objects (and was also applied to multimodal data processed by computer vision and speech processing algoriths). Notably, Yu and Ballard (2007) also took the step of integrating prosodic and social cues into their model, which they found increased word-object mapping performance considerably. Fazly and colleagues presented an incremental model of this process that they applied to a range of developmental results such as mutual exclusivity (Fazly, Alishahi, & Stevenson, 2008, in press). Frank, Goodman, and Tenenbaum created a Bayesian model which jointly solved two tasks: inferring what referent the speaker was talking about and learning word-referent mappings (Frank, Goodman, & Tenenbaum, 2009). This model unified the approaches of several previous systems, modeling competition-style inferences like mutual exclusivity while also learning effectively from annotated corpus data. The success of this system and others also suggests 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. Combined with the success of models like Regier (2005)'s in fitting developmental data, the suggestion from this literature is that findings like mutual exclusivity (E. M. Markman & Wachtel, 1988) can be accounted for via a variety of statistical inference mechanisms.

In addition to this computational work, recent empirical investigations by Yu, Vouloumanos, and colleagues (Yu & Ballard, 2007; L. Smith & Yu, 2008; Vouloumanos, 2008; Vouloumanos & Werker, 2009) have given evidence that both adults and young children can use the same kind of cross-situational exposure to learn associations between words and their meanings. The current body of evidence with respect to these tasks provides a proof-of-concept that learning is possible, though work is now emerging that also attempts to characterize the mechanisms underlying these inferences in more detail (Yurovsky & Yu, 2008; Kachergis, Yu, & Shiffrin, 2009; Ichinco, Frank, & Saxe, 2009). 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.

Despite the assumptions of the models discussed above, knowing the object that is being talked about does not imply knowledge of the meaning of the word that is being used. To take one of many examples, sub- and super-ordinate labels like "animal" or "dalmation" can co-occur with more common, basic-level labels like "dog." Although previous theories suggested that early word learners were constrained to consider only basic-level categories as targets for mappings (E. Markman, 1991), recent work has provided a theory for how learners could make principled statistical inferences about what level of categorization a word refers to (Xu & Tenenbaum, 2007b).

Xu and Tenenbaum described a simple Bayesian model that relies on the notion of a suspicious coincidence. The motivating intuition for this model is that single example of a dalmation could be randomly chosen as an example of a dog, but choosing three dalmations randomly would be a coincidence if they were being chosen from the set of all possible dogs. Those same three examples would be much more likely to be chosen as examples of the sub-ordinate category "dalmation." 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.

One other 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. Jones, Smith, & Landau, 1991). One possible hypothesis is that this “shape bias” is innately given, but another possibility is that the shape bias is a higher-level bias that is learned through experience. In support of this second hypothesis, Smith and colleagues (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) have been applied to 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, 2007b) information. The success of models taking these steps supports the view that statistical learning is not a separate associative mechanism or cue. Instead, a framework that posits statistical inferences over the available data for the purpose of inferring a discrete lexicon seems to provide better fit to the broad, integrative inferences that children make in learning the meanings of new words.

1.3.4 Word class learning

Although tremendous progress has been made in tasks from learning speech sounds to learning words and morphology, progress at the highest levels of acquisition has been somewhat slower. Understanding how the structure of natural languages can be learned is the most 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 a considerable amount of progress in several 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 learning grammars from un- or minimally-annotated input.

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 in linguistics, NLP, and acquisition (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 (Maratsos & 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). Braine (Braine, 1987) then showed evidence that distributional learning strategies could succeed in this task, but only when they were supplemented by additional referential or morpho-phonological information. More recently, studies that include multiple distributional cues to category membership (e.g., a frame of two words rather than a single word) have also been successful in allowing learning (Mintz, 2002), and there is some evidence that languages that simply rule out the possibility of positional regularities via optional words at the beginnings and ends of strings may allow for learning (Reeder, Newport, & Aslin, 2009). In addition, Gerken (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. Recent work by Frank and Gibson (Frank & Gibson, under review) suggests a reason for these effects: that the pattern of failures in these experiments may be due to learners' memory limitations (and that adding coordinated cues may simply give extra cues for encoding).

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 to learn words more effectively (Frank, Goodman, & Tenenbaum, 2009), the use of semantic information to extract sub-classes of words (Alishahi & Stevenson, 2008), or the joint induction of categories and rules over those categories (Bannard et al., 2009).

1.3.5 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 (D. 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 (G. 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 (Albright & Hayes, 2003) compared a rules-plus-exceptions model (Prasada & Pinker, 1993) to a purely analogical model and 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. This 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 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 (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 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, Ichinco, & Tenenbaum, 2008; 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).

1.3.6 Syntactic 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. Human learners have sharp limitations on memory and computation that leading to characteristic errors that are not always described by the current generation of models. 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 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 (E. 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 (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. Although this network was only able to learn from small languages, 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.

Recent work has attempted to unify insights from NLP with work on child language acquisition. For example, Perfors et al. (2006) 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. Applying 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 children's productions (Borensztajn, Zuidema, & Bod, 2008; Bannard et al., 2009). These systems have found evidence for the increasing abstraction of the units used in production, although they leave unanswered the question of the representations underlying early syntactic abstraction in comprehension (Gertner, Fisher, & Eisengart, 2006; Thothathiri & Snedeker, 2008; Arunachalam & Waxman, in press).

In addition to creating systems that learn pure syntactic relationships, 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 work by Zettlemoyer, Collins, and colleagues (Zettlemoyer & Collins, 2005, 2007, 2009) has made use of combinatorial categorical grammar (Steedman, 2000), a linguistic framework by which word order and logical forms are jointly derived from the same grammar. This group designed a series of systems for learning CCG parsing systems that similarly identify the logical forms of natural language sentences. Although these systems have not been 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.

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 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 (or even, as in (Borensztajn et al., 2008) and (Bannard et al., 2009), partially-specified) syntactic rules.

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 (G. F. Marcus et al., 1999; Saffran, Pollak, Seibel, & Shkolnik, 2007; G. F. 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, under review).

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, 2003), 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.

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. 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 (with some exceptions (Bannard et al., 2009; Alishahi & Stevenson, 2008)). Thus, an important goal for future work should be the development of systems and experimental paradigms which allow direct links between human data and the learning performance of models.

1.3.7 Synergies between tasks

Much 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. Feldman, Griffiths, and Morgan (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 B. Jones, Johnson, and Frank (2010), who proposed a model 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.

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, Horst, & Samuelson, 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.

1.4 Conclusions: Word learning as a case study

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

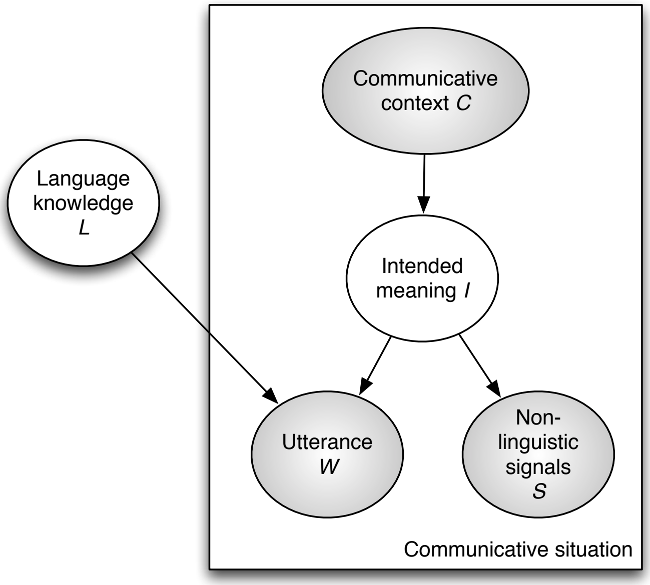

Figure 1-1: The graphical model representing dependence relations in our framework for language acquisition.

this process. 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 learning problem as induction of explanatory units like words, morphemes, categories, or rules were more successful.

The goal of this thesis is to apply the view described above—language acquisition as driven by statistical inference over structured representations that integrate multiple information sources—to the task of early word learning. Early word learning—which we define as lexical acquisition during the period of word learning during which the learner's production is still at the single word stage and vocabulary is still dominated by nouns and social terms—is an important test case for our view because it is both simple enough that we should be able to capture many of the important phenomena and complex enough that it interacts with many other areas of cognition including social cognition and conceptual development.

The schematic model that guides this view is shown in Figure 1-1. This model describes a communicative view of language: that words are used to refer to aspects of the physical context. This model thus sketches a view of word learning as "communicative inference": inferences about the meanings of words that make use of the way words are used to convey meaning in context.

The specific assumptions of this proposal are as follows. We assume that knowledge of language L, in this case a lexicon (a set of word-concept mappings), is built up across experience with a variety of communicative situations. In each situation, there is a context C that leads to the speaker forming some intention to communicate, with an intended meaning I, and on the basis of that intention uttering some utterance W and some non-linguistic, social signals S. Learners are assumed to observe a set of communicative situations, each with some context, language, and non-linguistic signals. Their challenge is to infer the two hidden variables L and I. The language (the actual knowledge that the speakers have, not the learner's guess about it) is assumed to be constant across time, while the speaker's intention is assumed to vary from situation to situation. These two hidden variables represent two timescales that are relevant for word learners (cf. McMurray, Horst, and Samuelson (McMurray et al., under review) for a different take on this issue).

Across our investigations of different aspects of early word learning the contents of these variables changes, but the general relationship between them that is pictured in Figure 1-1 is assumed to remain constant. The chapters of this thesis address distinct aspects of the difficult joint inference problem that the full model describes. Figure 1-2 shows graphically how the three substantive chapters of the thesis address the relationships between different variables in the larger model.

Chapter 2 motivates this framework for studying early word learning by investigating the relationship between physical context, intended referent, and caregivers' social signals. We describe a corpus study of English-speaking mothers interacting with children from six months to a year and a half old that investigates how well social signals like eye-gaze and pointing can be used to predict a speaker's referent in a very

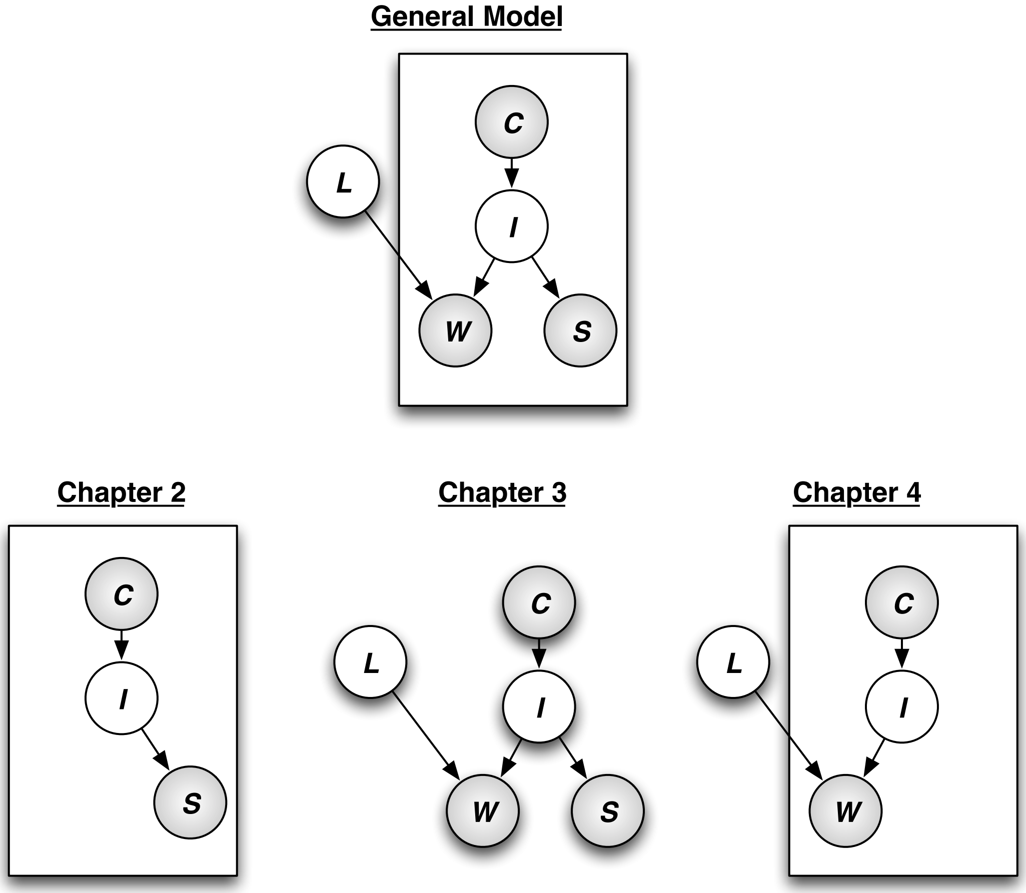

Figure 1-2: Instances of the general model that correspond to each chapter of the thesis.

restricted environment. We find that, even when there are only a small number of possible referents to talk about, social signals are at best noisy indicators of mothers' intentions. Thus, succeeding in making inferences about what is being talked about with any degree of consistency will require aggregating information across multiple referential episodes (as well as making use of linguistic context).

Chapter 3 investigates the kinds of complex inferences that are possible in our framework during a single interaction. We use our framework of communicative inferences to formalize what it would mean for a speaker to choose their message informatively given a particular context. We then apply this formalization to the problem of inferring the meaning of a novel adjective. Experiments with adults show a quantitative match between the strength of inferences about adjective meaning and the predictions of our model; an experiment with children suggests that they are capable of making this kind of inference as well. This chapter articulates why we describe our proposal as "communicative" rather than simply inferential: the kinds of inferences that are possible based on knowing that another agent is taking an action to communicate are much stronger than those that are possible when an action is caused physically.

Chapter 4 describes a computational model of communicative word learning that can be applied to both experimental and corpus data. We show that it outperforms purely associative models in learning words from annotated corpora and demonstrate how it predicts a range of results like mutual exclusivity and the use of labels for object individuation.

Chapter 2

Contributions of social and discourse information to the determination of reference[1]

In this chapter, we describe a study that takes steps towards quantifying the informativeness of cues that signal speakers' chosen referent, including their eye-gaze, the position of their hands, and the referents of their previous utterances. We present results based on a hand-annotated corpus of 24 videos of child-caregiver play sessions with children from 6 to 18 months old. Our analyses suggest that social cues must be combined to be effective in guessing reference, and that assuming continuity of reference—that the same referent is talked about over the course of an extended section of discourse—may be an important step in aggregating information from social cues over time. Within our broader communicative inference framework, Figure 2-1 shows the hypothesized relationships underlying this study: the dependence of communicative intentions I on context as well as the dependence of social cues S on those intentions; this relationship is iterated over time, as indicated by the plate over all three variables.

[1] Joint work with Anne Fernald. Parts of this chapter were published as Frank, M.C., Goodman, N. D., Tenenbaum, J. B., and Fernald, A. (2009). Continuity of discourse provides information for word learning. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society.*

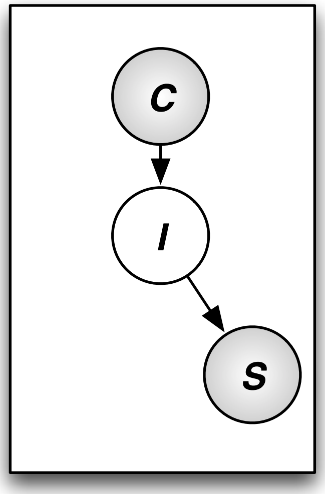

Figure 2-1: A graphical model showing the dependency relationships captured by the study of social and discourse cues in Chapter 2.

2.1 Introduction

How do children learn the meanings of concrete, referential words? Divergent lines of research have argued either that cross-situational associations between words and their referents can reveal word meanings over time (Locke, 1847; Yu & Ballard, 2007; L. Smith & Yu, 2008) or that ostensive social cues like eye-gaze and pointing can uniquely determine the referent of a word during a single naming event (St. Augustine, 397/1963; Carey, 1978; E. M. Markman & Wachtel, 1988; Markson & Bloom, 1997). Empirically, sometimes words are learned quickly from a few ostensive examples and retained effectively (Carey, 1978; Markson & Bloom, 1997). Sometimes meanings may be inferred quickly but are lost just as quickly (Horst & Samuelson, 2008). And in other cases, words appear to be learned slowly via extensive, incremental exposure (Yu & Ballard, 2007; L. Smith & Yu, 2008; Vouloumanos, 2008; Vouloumanos & Werker, 2009).

Recent computational proposals have suggested both processes are at work and that they interact with each other in productive ways (Yu & Smith, 2007; Frank, Goodman, & Tenenbaum, 2009; McMurray et al., under review). One perspective on these differing experimental results is that the presence of social information occasionally serves to raise the salience of particular referents to such a degree that only a small number of associations between word and referent are needed for learning (Yu & Smith, 2007). A contrasting perspective suggests that rather than directly raising the salience of word-object pairings, social information instead contributes to the process of understanding speakers' intentions (Frank, Goodman, & Tenenbaum, 2009).

Crucial to both viewpoints is the assumption that social signals are truly informative about speakers' referential intentions. This chapter addresses this assumption empirically, investigating the overall reliability of social cues as signals to reference in child-directed speech. The method for our investigation is a corpus study. Quantifying the information available from videotapes of caregiver-child interaction allows us to analyze the learning environment directly. We measure the informativeness of social cues to reference in a very simple, highly-constrained situation: a caregiver and a child playing with pairs of toys. We hand-coded what the caregiver and child are looking at, talking about, and pointing at; we also code, for each utterance, which object in the immediate environment the caregiver is referring to (if any).

The first parts of this chapter describe our corpus, our coding methods, and the results of our analyses of social cues. Our data suggest that either individually or taken together, social cues are very noisy. Although they carry information about caregivers' referents, even a learner who knew *which* social cues to attend to would still only be able to guess the speakers' referent correctly a fraction of the time. This result strongly suggests that learners aggregate information about reference across time. In the second part of the chapter, we explore one way that learners might compensate for the noisiness of referential cues: by assuming that reference is continuous and that what was being talked about is quite likely to be the same as what was being talked about in the previous utterance. We provide some intuitions about this assumption of referential continuity and show analyses that suggest that it is a good assumption to make about child-directed speech. We conclude by using a classification analysis to investigate how much information about speakers' referential intentions can jointly be extracted by combining information from social cues with an assumption of discourse continuity. Taken together, these results provide support for a view of early word learning as communicative or intentional inference: the aggregation of information about reference from non-linguistic social cues and from language in service of inferring word meanings.

2.2 Previous work

2.2.1 Social cues to reference

The general relationship between social cues and language development is well established. A large body of work examines the efficacy of social cues for word learning in restricted experimental situations (e.g. (Carey, 1978; Baldwin, 1993; Woodward, Markman, & Fitzsimmons, 1994; Hollich, Hirsh-Pasek, & Golinkoff, 2000)), and it is widely acknowledged that from around their first birthday or before, children can use explicit, ostensive signals to learn the connection between a word and an object. In addition, a wide variety of work has attempted to characterize the nature of caregiver-child interactions and their links to language development (e.g. (Stern, 2002; Bruner, 1975; M. Carpenter, Nagell, Tomasello, Butterworth, & Moore, 1998)). For example, work by Carpenter, Nagell, and Tomasello (M. Carpenter et al., 1998) investigated the relationship between markers of joint attention and social cognition, finding that infants who engaged in more joint attention communicated more, both linguistically and gesturally.

If we consider longitudinal studies like those of Carpenter et al. to be the equivalent of macro-economics—measurements of the relationships between global variables in the environment—there has been comparatively little work on the micro-economics of reference: which particular cues matter to determining reference in an interaction. One recent computational study is relevant, however. Yu and Ballard (Yu & Smith, 2007) created a model of the unsupervised learning of word-object associations from child-directed speech annotated with the objects that were present during each utterance. They then used social cues like eye-gaze and prosodic cues like focus as markers of which objects and which words were more salient; incorporating these cues to attention into their model resulted in gains in word learning performance. This study suggested that social cues were informative about reference, however it did not quantify this relationship independently from the particular associative learning model that was used. Our current work builds on this work by investigating the relationship of individual social cues to the speakers' chosen referent.

2.2.2 Continuity in child-directed speech

Discourse structure has been well-studied in psycholinguistics (P. Carpenter, Miyake, & Just, 1995; Graesser, Millis, & Zwaan, 1997; Wolf & Gibson, 2006), but the level of description of discourse primitives has been much more detailed and abstract than that which would be possible for a child learning the meanings of words. Instead of describing continuity of topic—an abstract notion which may be hard to pin down for the concrete, here-and-now interactions between parents and young children—here we consider a version of "talking about the same topic" that may be more appropriate to the situation early in language acquisition: talking about the same object. This approach to discourse, which we call "continuity of reference" is more likely to capture the kind of information available to even the youngest word learners (who have access to social markers of reference but might not be able to track continuity linguistically). For the rest of the paper, we will use the terms "continuity of reference" and "discourse continuity" interchangeably.

It is widely acknowledged that speech to children is highly repetitive and includes many partial repetitions of phrases and that these features may be useful to learners (e.g. (Snow, 1972; Hoff-Ginsberg, 1986, 1990)). Despite this fact, research on word learning has largely neglected the role of reference continuity as an information source for learning words. For example, although a number of recent computational models use cross-situational information about the co-occurrence of words and referents for word learning, most of these models assume that utterances are sampled independently from one another with respect to time, throwing away important information about the order of utterances (Siskind, 1996; Yu & Ballard, 2007; Frank, Goodman,

& Tenenbaum, 2009).

One counterexample to this generalization comes from work by Roy (Roy & Pentland, 2002). Their CELL model searched for speech segments that correlated with shapes in its visual input. In order to avoid learning spurious associations, they used a recurrence filter to exclude pairings that were not repeated nearby one another in the input. Although this strategy was included as a heuristic way of eliminating the majority of the incoming sensory input (rather than via a model of discourse as such), it serves the same purpose and effectively illustrates a mechanistic reason why discourse continuity might be useful in learning.

Slightly further afield, *variation sets* (Harris, 1951; Kuntay & Slobin, 1996; Onnis, Waterfall, & Edelman, 2008)—contrasts between temporally-contiguous sentences with only small changes to their content that may allow learners to make comparisons easily and extract relevant regularities—are a more specific idea than continuity, but the two concepts are nonetheless related. In one investigation using artificial language materials, Onnis et al. (Onnis et al., 2008) showed evidence that learning word segmentation from continuous materials and learning a simple artificial grammar were both facilitated by the presence of variation sets in the materials. Work by Waterfall (Waterfall, 2006) found significant correlations between the use of variation sets centered around particular constructions by caregivers and the production of those constructions by children, suggesting that concentrated exposure to a particular construction in an repeated, semantically-meaningful context can be important in acquisition.

Thus, although as far as we know there have been no direct investigations of the role of discourse or reference continuity in word learning, related work on repetition and variation suggests that discourse continuity is likely both to be present in child-directed speech and useful in vocabulary acquisition.

Figure 2-2: A sample frame from the FM corpus.

2.3 Corpus Materials

For our analyses, we chose our corpus based on two criteria. First, a potential corpus needed to include video as well as audio so that we could accurately identify both the speaker's referents and the other objects present in the physical context. Second, the corpus needed to be collected in a restricted enough context that it would be possible to code all of the objects that were present (so we could consider all of the alternative possible referents for a word).

We selected a corpus which fulfilled these requirements: a set of videos of object-centered play between English-monolingual, American mothers and children in their homes, originally reported in (Fernald & Morikawa, 1993). The children in these videos fell into three age groups: 6 months (N=8, 4 males), 11-14 months (N=8, 5 males), and 18-20 months (N=8, 4 males). All families were Caucasian. The corpus was collected by a pair of female observers who made visits to the homes of participants and audio- and video-recorded mother-child dyads as they played. After an introductory period, sets of standardized toy pairs were introduced, including a stuffed dog and pig, a wooden car and truck, and a brush and a box. The mother was given each pair of toys for 3-5 minutes and asked to play "as she normally would."

Towards the end of the session, the experimenter asked the mother to hide several of the objects and have the child search for them. Although the original study analyzed only 5 minutes of data from each video (due to the particular aims of the study), we used the full set of transcripts they produced, which included a considerably larger amount of interaction in some cases, including a small amount of parenthetical speech to the experimenter. Descriptive data for each mother-child dyad in the corpus are given in Table 2.1, including gender, age, length of the recording, and a variety of measures of the diversity of referents and utterances.

Table 2.1: Descriptive statistics for each file in the FM corpus.

Group	Code #	Gender	Age	Utterances	Video len.	Objects	Objects/utt.	Word tokens	Word types/utt.
6 mos	31	M	6	58	3:01	3	1.17	88	3.26
	32	F	6	66	4:18	5	1.26	117	3.39
	33	M	6	155	8:45	5	1.25	190	4.26
	35	F	6	197	10:04	4	2.00	200	3.79
	36	F	6	232	10:16	5	1.95	166	3.45
	38	M	6	189	12:20	4	1.80	224	3.72
	39	F	6	178	10:26	5	1.60	161	3.63
	40	M	6	397	12:42	4	2.00	185	3.38
12 mos	28	M	11	457	25:37	10	1.99	255	3.70
	2	M	12	216	19:01	9	1.66	163	3.69
	3	M	12	240	20:02	9	1.74	148	3.62
	4	M	12	327	27:08	9	1.74	244	4.07
	8	F	12	297	28:29	9	1.45	182	4.33
	12	F	13.5	428	23:25	10	1.25	234	3.79
	14	M	14	184	17:50	9	1.40	195	5.11
	16	F	14	546	23:14	14	1.99	273	3.82
18 mos	17	F	18	351	27:41	9	1.97	330	5.57
	18	M	18	276	14:27	12	1.23	193	3.72
	26	F	18	455	23:50	9	1.70	311	4.46
	29	M	18	352	22:36	8	3.54	214	3.77
	22	M	19	451	22:14	10	3.25	223	3.49
	19	F	20	244	18:17	9	3.06	201	4.09
	20	F	20	360	29:52	9	2.05	311	4.87
	21	M	20	382	24:00	9	2.28	285	3.69

For each utterance we first coded the mid-sized objects present in the field of view of the learner at the time of the utterance. A sample frame from the videos is shown in Figure 2-2. The only object judged to be in the field of view of the child at the time of the utterance most proximate to this frame was the `dog`. We also coded, for each utterance, the object or objects in the context that were being looked at, held, and pointed to by the mother. These cues were sparse: in many cases, no object was being looked at, held, or pointed to and so these fields were marked 'none.' This method of coding was chosen because it was practical for the large amount of video data we were working with (a total of approximately 7.3 hours of video). One potential downside of this coding method is that it does not make use of the temporal coordination between, e.g., eye-gaze and language production (Griffin & Bock, 2000). The use of eye-tracking during natural interaction is outside of the scope of the current study and may prove difficult more generally (but c.f. (Merin, Young, Ozonoff, & Rogers, 2007) for an interesting approach to the problem). In addition, a child observing a caregiver's eyes during natural interaction may be only slightly more accurate in identifying the object of their eye-gaze than a third-person observer who can watch the same video clip multiple times.

Though the data arguably are not a part of the same ideal observer analysis, we also coded two other social cues: the object or objects that were being looked at or held by the child. We use these data in the followup analyses we report on the reliability of individual social cues and return to them in the discussion section.

We next coded the speaker's *referential intention* for each utterance. We operationalized referential intention as an intention to refer linguistically to an object. We coded an utterance as referring to an object when the utterance contained the name of the object or a pronoun or onomatopoeia referring to that object. For example, in a sentence like "look at the doggie," the referential intention would clearly be to talk about the `dog`. Likewise, in an utterance like "look at his eyes and ears," (where the caregiver was pointing at the dog), the referential intention would also be the `dog`—though the coder would need to check the videotape to determine the pronoun reference. We did not mark the use of property terms like "red," super-/subordinate terms like “animal” or “poodle,” or part terms like “eye.” Exclamations like “oh” were not judged to be referential, even if they were directed at an object. Objects that were not present were still judged to be part of a referential intention, e.g., “do you like to read books” would be judged to have the intention `book` even if the child could not see a book or a book was not present in the scene at all.

The end product of this coding effort was a corpus of approximately 7k utterances and 28k words, for which each utterance was annotated with the objects present in the field of view of the learner, the referential intention(s) of the speaker, and the social cues given by the mother and child. The text and videos of this corpus are available at `http://www.stanford.edu/~mcfrank/materials`.

2.4 Social cues

The goal of the following analyses is to measure the efficacy of physical social cues given by mothers in revealing what objects they were referring to. We first use descriptive analyses to understand the basic distribution of social cues across objects for children in the different age groups. We then examine the timecourse of these cues.

2.4.1 Signal detection analyses

We first measured the independent reliability of each social cue. To perform this analysis we examined the transcripts from each video separately. For each of the three social cues we coded (mother’s eye-gaze, hand position, and point) we calculated three signal-detection measures: hits (e.g., pointing to an object and referring to it in the same utterance), misses (e.g., not pointing to an object but referring to it), and false alarms (pointing to an object but not referring to it). These measures assume that each social cue is a noisy, binary signal, whose general informativeness relative to the behavior of interest (reference) can be measured within the signal-detection framework.

From these three measures, we calculated two standard scores for summarizing

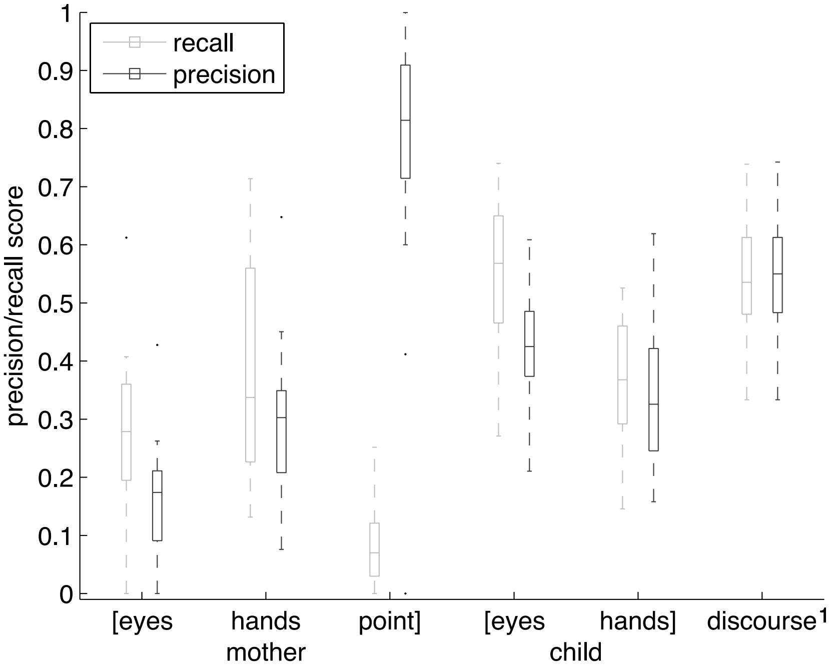

Figure 2-3: Boxplots showing precision and recall for social and discourse cues to reference provided by each mother in our study. Boxes show median and 25th and 75th percentiles, whiskers show approximately 2.7 standard deviations, and dots show outliers.

performance. The first was recall (hits / hits + misses) and the second was precision (hits / hits + false alarms). Intuitively, recall measures the proportion of correct references for which a particular cue to intention was present, while precision measures the proportion of the time in which the cue was correct. These two measures can be combined into F_0, their harmonic mean, for easy comparison.

To see how this analysis operates in practice, imagine a corpus that consists of only two scenarios. In the first, the mother looks at a toy, points to it, and says "you see this?"; in the second, she looks at the child and picks up the toy, saying "isn't it nice?" In this example, the toy is being referred to in both utterances (though both times by pronouns). The mother's point correctly identifies her referent when it was observed but was not present the second time the mother referred to the toy, giving it a precision of 1.0 (1/1) but a recall of .5 (1/2), for an F-score of .667. In contrast, the mother's look identified the toy in the first utterance but failed to identify the

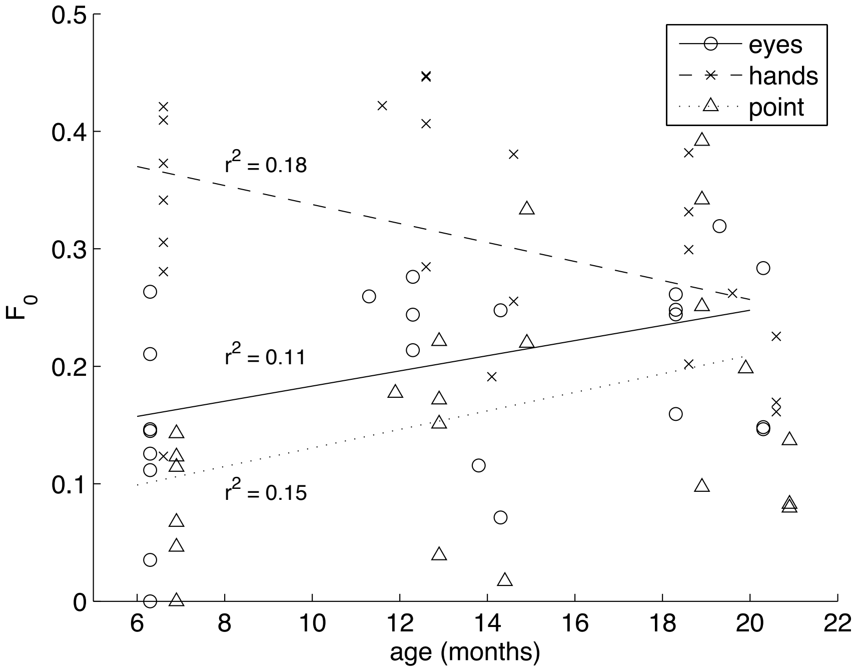

Figure 2-4: F_0 (the harmonic mean of precision and recall) for each of three reference cues plotted against individual infants' ages. Lines show the best-fitting linear model.

toy in the second utterance, giving it a precision of .5 (1/2) and a recall of .5 (1/2), for an F-score of .5.

Precision and recall boxplots for each information source are plotted in Figure 2-3. Neither the position of caregivers' eyes (mean precision = .16, mean recall = .27, mean F_0 = .20) nor their hands (precision = .39, recall = .30, F_0 = .32) were particularly good cues to reference. In contrast, caregivers' pointing gestures were low recall but high precision (precision = .77, recall = .08, F_0 = .15). Points were relatively few and far between, even in the kind of context that would be most open to ostenstive word teaching. But when they were present, points were very reliable cues that a particular object was being talked about.[2]

Figure 2-4 shows the relationship of the child's age to the relative reliability of information from the caregiver, plotting F_0 by the age of individual infants. The

[2] Note that although pointing is more precise than other cues, mothers still point in ways that selects objects other than those that are being named, e.g. "by the pumpkin! look, look!" (pointing at a flashlight that was lost).

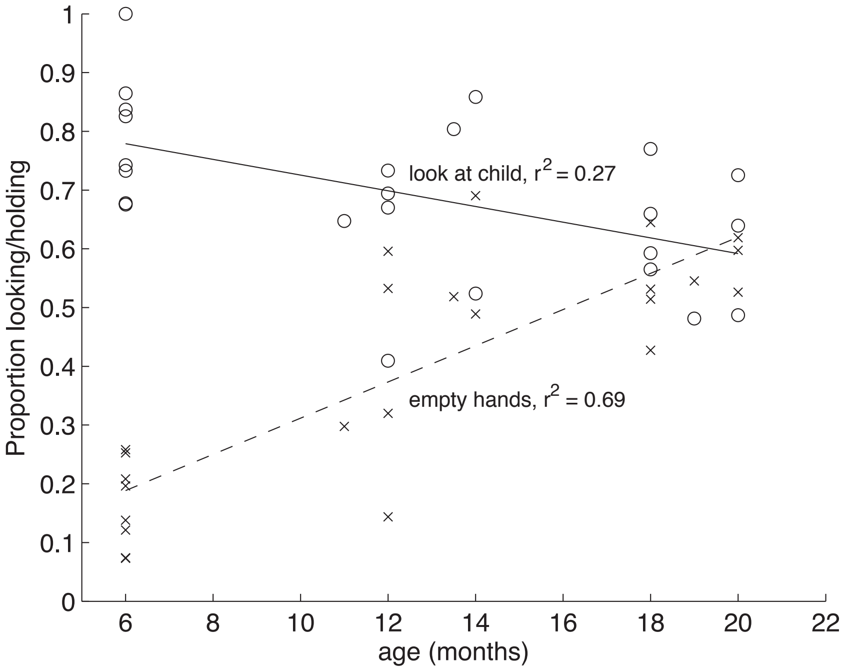

Figure 2-5: Proportion of utterances for which the mother was looking at the child and proportion of utterances for which the mother's hands were empty, both plotted by age. Lines show the best-fitting linear model for each variable.

reliability of caregivers' hands decreased significantly with age ($r^2 = .17$, $p = .05$), while the reliability of pointing and eye-gaze both tended to be higher for the mothers of older children ($r^2 = .15$, $p = .07$ and $r^2 = .11$, $p = .12$). These data suggest that caregivers may trade off particular cues to reference: when children are old enough to hold objects on their own reliably, caregivers are somewhat more likely to look at objects rather than their child and to indicate their referential intention with a point.

What do caregivers look at and hold, if they are not looking at or holding the item they are talking about? By and large mothers look at their children. Mothers in each of the three groups (6, 12, and 18mos) were coded as looking at their children 79%, 67%, and 61% of the time, respectively. Even for the oldest group, mothers looked at children the majority of the time, but this behavior did decrease significantly as children got older ($r^2 = .27$, $p < .01$). The contents of caregivers' hands also changed considerably across ages ($r^2 = .59$, $p < .0001$). Caregivers of the youngest children were generally holding something, while more than half the time (55%), the caregivers of the oldest children were not holding anything. Both of these trends are plotted in Figure 2-5.

Finally, in order to understand the relatively small amount of information conveyed by the caregivers' hands, we used the same signal-detection analyses to examine the informativeness of the child's eyes and hands in revealing what the speaker was talking about. Both of those markers of the child's attention had higher precision and recall than the equivalent cue for the mother (these information sources are also plotted in Figure 2-3); the position of the child's eyes was especially revealing (precision = .55, recall = .42, $F_0 = .47$) but the hands were also a good cue (precision = .35, recall = .36, $F_0 = .34$). Neither of these cues was modulated by age ($r^2 = .01$ for the child's eyes and $r^2 = .04$ for the child's hands).

Taken together, these analyses suggest that no individual social cue was consistent in revealing caregivers' intentions. Overall, a child attempting to figure out what her mother was talking about would do best to focus on what she herself was holding or looking at, rather than trying to interpret the same information from her mother. Only in the relatively rare cases when caregivers pointed were their actions highly informative about their referent.

2.4.2 Timecourse analyses

The goal of these analyses was to explore temporal dynamics in the use of particular social cues. In particular, we were interested in whether some cues were used more often at the beginning of talking about objects. For instance, one simple hypothesis might be that caregivers use points to introduce objects into the discourse.

In order to test this question, we performed an analysis of the probability of using each social cue over time. We performed the analysis as follows: we defined a "bout" of references to a particular object to be at least three continuous references to an object (results did not change significantly when we explored other reasonable values for this number). Next, for each age group we aggregated all bouts of talking about any object. We aligned each of these bouts so that they were numbered starting from

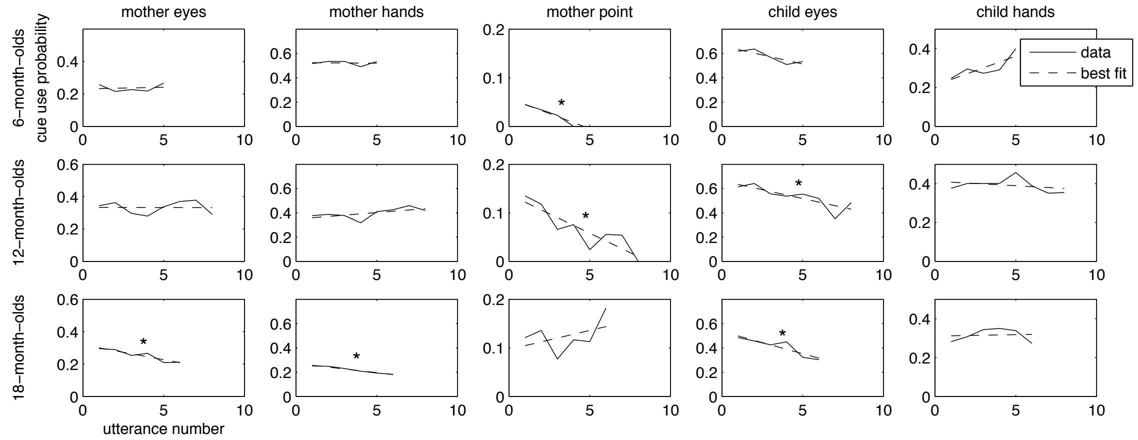

Figure 2-6: Each plot shows the probability of a particular social cue being used within a "bout" of speech about a particular object, plotted by the length of time the object had been talked about. Empirical data are shown with a solid line and the best linear fit to these data are shown with a dashed line. Significant linear trends are shown with a star. Each row of plots shows an age group and each column of plots shows a particular social cue.

their first utterance and averaged the probability of each social cue for the object that was being referred to. There were 89, 215, and 223 such bouts for the 6-, 12-, and 18-month-olds, respectively. Since relatively few bouts lasted longer than 5 or 10 utterances, data were too sparse to calculate social cue probabilities accurately for longer bouts. Therefore, we excluded bout lengths for which we had fewer than 20 bouts. Results are plotted in Figure 2-6, where each axis shows, for a particular cue and age group, the probability of observing that cue at each point during a bout.

For each cue and age group we performed a simple linear regression. We found that the probability of use for all cues stayed constant or decreased; no cues significantly increased in frequency as a function of time as bouts continued. The probability of the mother's eyes being on the object stayed relatively constant for all age groups except 18-month-olds, for whom it decreased slightly over time. The same result held true for the mother's hands. Though the base rate of the mother pointing to the object was low to begin with, the probability of a point decreased considerably for both the 6- and 12-month-olds as an object was talked about more. Interestingly, that generalization did not hold for the 18-month-olds, at least in part because some mothers were using points to pick out subordinate features of the objects. The probability of the child looking at the object also decreased as the object was talked about more, perhaps due to the increasing speed with which older children habituate (Hunter & Ames, 1988); this trend was significant for the two older age groups but trended in the same direction for the younger children. Finally, the probability of the child's hands being on the object that was being talked about stayed relatively constant.

This analysis suggests that the utilities of individual social cues often tend to decrease over the course of an episode of talking about a particular object. Because of this, identifying which object is being talked about may be easier to do early on after it has been introduced. Thus, this section identified two challenges for learners attempting to determine a caregiver's referent: first, that individual social actions do not always transparently signal reference, and second, that signals of reference tend to become sparser as an object has been talked about for longer. Our next set of analyses examine whether aggregating information over time may allow learners to overcome these issues.

2.5 Continuity information

Our next goal was to quantify the continuity of reference—the tendency of caregivers to talk about a single object for multiple utterances in a row—and to test whether it could be a source of information for predicting what objects were being referred to. We first develop a visualization of reference in child-directed speech. We next show some descriptive results about the magnitude and temporal dynamics of reference continuity. We end by comparing reference continuity to the social cues examined above using the same signal detection analyses.

2.5.1 Visualizing continuity of reference

The first step we took towards understanding the prevalence of discourse continuity was to visualize the results of coding the speakers' referential intention. We introduce what we call a "Gleitman plot": a visualization of a stretch of discourse based on

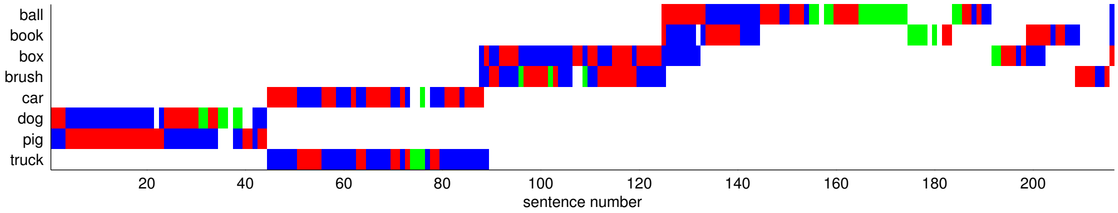

Figure 2-7: Example Gleitman plot. Each row represents an object, each column represents an utterance. A blue mark denotes that the object was present when the utterance was uttered but not mentioned; a green mark denotes that the object was mentioned but not present; and a red mark denotes that the object was present and mentioned. The streaks of red indicate bouts of continuous utterances referring to a particular object.

(1) what objects are present and (2) what objects are being referred to. Gleitman (Gleitman, 1990) was concerned with the relationship between what is present in a learner's experience and what is being talked about, hypothesizing that words like verbs that are often used when their referents are not clearly picked out in the physical context (and sometimes when their referents are not even present). Our visualization attempts to make the relationship between reference and the physical context transparent; a sample Gleitman plot for one file in our corpus is shown in Figure 2-7.

We can draw two anecdotal conclusions on the basis of viewing the Gleitman plots for each file in the corpus. First, within the corpora we studied, mothers talk primarily about objects that are present in the field of view of the children. This can be seen by examining the small amount of green within the plots. The largest bout of green is in the range around utterances 155 – 180, when the mother is playing a hiding game with several of the toys. Unsurprisingly, for a word learner guessing the meaning of a novel noun, the best guess will likely be that the word refers to an object that is present (Pinker, 1989; Yu & Smith, 2007; Siskind, 1996). (Although this generalization may be true for nouns, it is much less likely to be true for verbs, as Gleitman (Gleitman, 1990) pointed out).

Second, we can see clear evidence of discourse continuity (again, defined as continuity of reference). For example, in Figure 2-7, rather than being distributed evenly throughout the span of time when an object is present, references to an object are

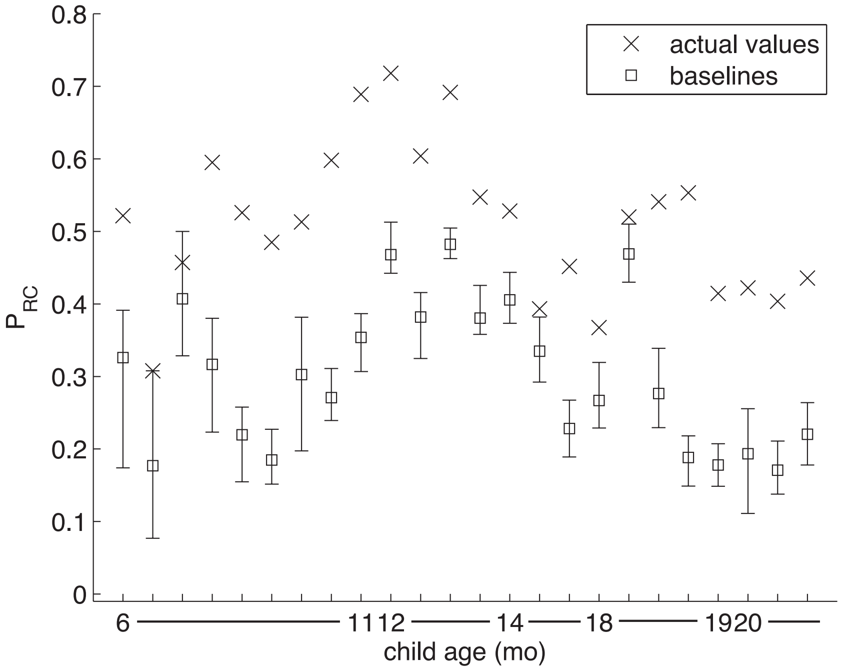

Figure 2-8: Probability of reference continuity for each child is shown with an "x". Each point on the horizontal axis shows a child (in age order). Squares with with error bars shows 95% confidence intervals for a permuted baseline.

"clumpy": they cluster together in bouts of reference to a single object followed by a switch to a different object. This can be seen for example in the `dog` / `pig` portion (first 45 utterances), where the mother alternates several times between the two objects, talking about each for several utterances before switching.

2.5.2 Measuring reference continuity

In our visualizations, we observed clumps of references to a particular object rather than a more uniform distribution of references over time. To quantify this trend, we first defined a quantitative measure of reference continuity, P_{RC}: the probability of referring to a particular object, given that it was talked about in the previous utterance.

We go into some detail about how this measure was calculated in order to be clear about how we calculated our baseline measure, since an appropriate baseline is crucial for determining whether P_{RC} is greater than chance. For an object o, we defined the reference function $R_t(o)$ as a boolean function returning whether or not that object was referred to at time t. We then defined $P_{RC}(o)$ (the probability of reference continuity for a particular object):

$$P_{RC}(o) = \frac{\sum\limits_t R_t(o) R_{t-1}(o)}{\sum\limits_t R_t(o)} \tag{2.1}$$

We calculated $P_{RC}(o)$ for each object for the times when it was present in the physical context. We then took an average of $P_{RC}(o)$ over all objects, weighted by the frequency of each object, to produce an average value for each file.

We then estimated a baseline value for P_{RC} via permutation analysis. Intuitively, this analysis asks what a "chance" value for P_{RC} would be if utterances were completely independent of one another. This analysis is important because the distribution of individual objects is very uneven in time and some objects are more likely to be talked about than others.

We calculated this baseline value for each corpus file by recomputing $P_{RC}(o)$ for 1000 random permutations of the times at which each object was talked about.[3] For the Gleitman plots in Figure 2-7, this analysis would be represented by randomly shuffling all the red and blue squares in each row so that the same overall set of squares were red and blue but their ordering was different.

The results of this analysis are shown in Figure 2-8: the value of P_{RC} for each child is plotted along with the permutation baseline for that child. As predicted based on the Gleitman plots, P_{RC} was outside of the 95% confidence interval on chance for all but 4 of the 24 files. A simple linear regression showed no relationship between P_{RC} and age ($r^2 = 0.061$, $p = .25$). Thus, it appears that reference is considerably more continuous than would be expected by chance in child-directed play situations of the type in our corpus: repeated reference was on average 1.85 times as likely as expected

[3] Excluding utterances during which an object was not present was important in calculating an accurate baseline; had we permuted all utterances, we would have artificially deflated the baseline by spreading references to o across the entire file even when o was not present.

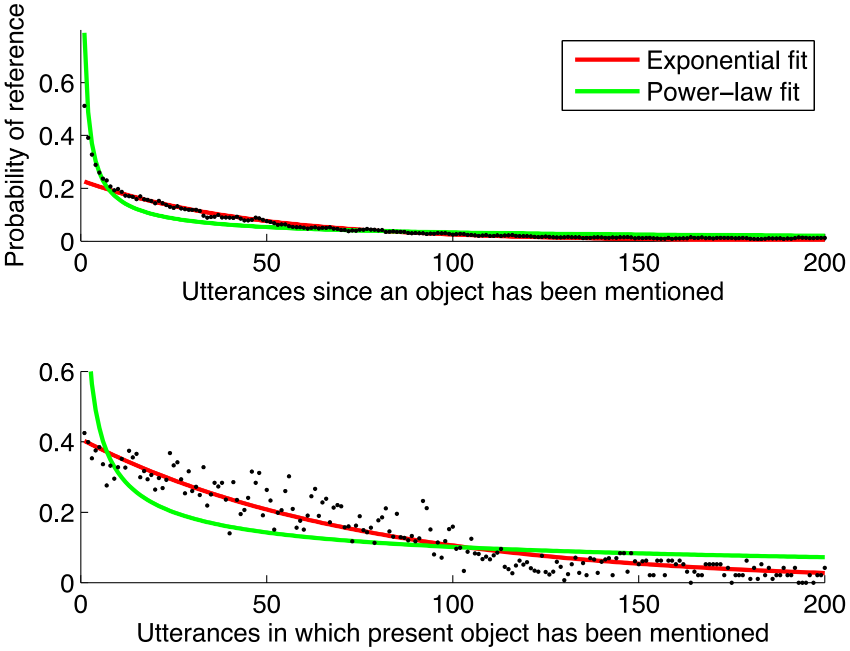

Figure 2-9: (top) Probability of reference given that an object was referred to N utterances ago, where n is plotted on the x axis. Dots show individual data points (jittered to avoid overplotting). Datapoints represent individual dyads. (bottom) Probability of reference given that an object is present and has been referred to m times over the course of the interaction. Datapoints have been aggregated across caregivers. For both plots, the red and green lines show the best two-parameter exponential and power-law fit, respectively.

by chance. In the absence of other information about what was being talked about in a particular situation, a good bet would be that moms in our corpus were still talking about the same thing they were a moment ago.

2.5.3 Temporal properties of reference

We next examined two temporal properties of discourse. The first is how recently an object has been talked about. This property can be thought of as a generalization of the analysis above; the new analysis asks about the probability of an object being talked about given that it was referred to some number of utterances ago. The second is the novelty of an object in context, here represented by how many times an object that is present in the context has been mentioned previously.

We conducted the first analysis simply by calculating a generalization of P_{RC} for each child-caregiver dyad. This new measure, P^n_{RC}, gives the probability of an object being referred to, given that it was referred to n utterances ago. Thus, P_{RC} is the same as P^1_{RC}, and we calculate it via an aggregation across objects, as before:

$$P^t_{RC}(o) = \frac{\sum\limits_t R_t(o) R_{t-n}(o)}{\sum\limits_t R_t(o)} \tag{2.2}$$

The result of this analysis are plotted in Figure 2-9, top. It is clear from this visualization first that very recent utterances are disproportionately correlated with the probability of referring again—this observation is captured by the previous analysis of discourse continuity. The influence of a particular object in discourse declines slowly, however, and even 50 sentences later there is some residual increase in the probability of talking about a particular object, given that it was talked about.

We attempted to quantify this pattern by fitting two functions to the resulting data, an exponential and a power-law. Both functions were fit by adjusting two parameters (intercept and decay) in order to minimize mean squared error. We found that the power-law (MSE = 1.11) fit considerably better than the exponential function (MSE = 1.66). This dynamic may be due to the more general phenomenon of power-law decays in human memory (Anderson & Schooler, 1990).

An object's novelty, both in the context and to a particular speaker, provides an additional factor governing how likely that speaker is to refer to an object. Intuitively, an object that is newly introduced to the context of the learner or a particular speaker is more likely to be talked about, and some empirical evidence suggests that children may be able to make use of this information to learn new words. (Akhtar, Carpenter, & Tomasello, 1996) found that two-year-olds were able to use the fact that an object was new to the experimenter (even though the children themselves had already played with it) to infer that the object was the experimenter's intended referent and hence was named by the novel word the experimenter produced.

To quantify the effects of discourse novelty on the probability of talking about an object, we calculated the probability that an object was being talked about (given that it was present) for the number of utterances for which the object had been present. We aggregated information as in the previous section (first across objects, then across dyads). The results are plotted in Figure 2-9, bottom. In contrast to the previous analysis, the resulting curve took longer to decay but had an overall lower intercept and was better-described by an exponential function (MSE = 3.16) than a power-law (MSE = 4.27), capturing the relatively steep drop-off in references to individual objects after they had been in play for a while.[4] The curve was so shallow, in fact, that it was fit almost as well by a linear function (MSE = 3.17); however, extrapolating out to 200 sentences we found that the exponential fit better (MSE = 5.44) than the linear function (MSE = 7.24).

2.5.4 Signal detection analysis

We conducted the same analysis for discourse continuity as cue to reference as we did for each social cue, analyzing the precision, recall, and F_0 of guessing that a particular utterance will refer to the same object that the utterance before did. Results are plotted alongside social cues in Figure 2-3, since the measures of evaluation are identical. We found that discourse continuity had the highest average F_0-score of any cue (precision = .57, recall = .54, F_0 = .55) and was not correlated with age ($r^2 = 0.00$, $p = .92$). In other words, a learner with perfect information about previous referents would do better guessing the current reference based on continuity than based on any individual social cue.

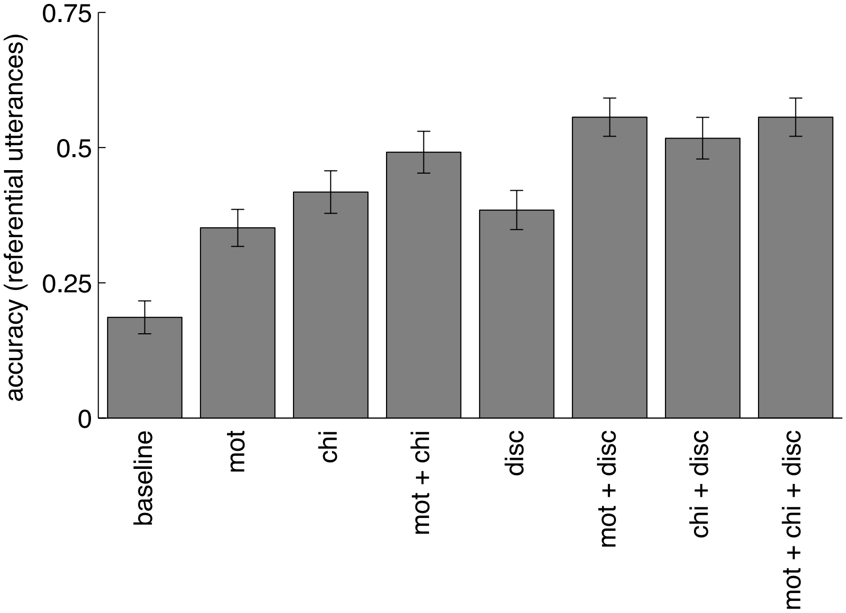

Figure 2-10: Classifier performance on those utterances for which there was an intention to refer to an object plotted by cues used in classification. Error bars show standard error of the mean across all children. "mot" = mother's social cues, "chi" = child's social cues, and "disc" = discourse cues.

2.6 Joint classification analysis

The goal of our last analysis was to measure how well an observer could guess which object a speaker was talking about, given the information available in the social and discourse cues just discussed. The idea behind this analysis is to use a supervised classification scheme to provide some measure of the total information available in these cues.

In order to carry out our classification analysis, we used a Naïve Bayes classifier (Hastie, Tibshirani, & Friedman, 2001) to combine each of the cues in order to make a judgment about what object was being talked about. We chose this classifier because

[4] This second result deserves a caveat since the loose structure of the instructions to mothers participating in the original (Fernald & Morikawa, 1993) study involved playing with several different pairs of objects. Thus, a larger and more diverse sample of corpus material will be needed to test the generality of this function.

it has the advantage of being simple and computationally efficient; unlike logistic regression it can classify datapoints across an arbitrary number of alternatives. Naïve Bayes makes the assumption that individual predictors are independent from one another; the simplicity of the approach allows us to measure when this assumption is violated by our data, however.

The classifier's task was to choose which of the m objects was being talked about in each utterance, based on weighted evidence from each of the cues C. We used a standard Naïve Bayes formulation:

$$p(O_m|C_1, ..., C_n) = \frac{1}{Z} p(O_m) \prod_i p(C_i|O_m) \tag{2.3}$$

where O_m denotes the object being talked about in a particular utterance (or "none"), C_i denotes a particular cue, and Z is a constant scaling factor. The Naïve Bayes classifier decomposes the posterior probability of an object into two terms: a prior and a likelihood. The term $p(O_m)$ is the prior, denoting the baseline probability (frequency) of a particular object being referred to; the term $p(C_i|O_m)$ is the likelihood of the cue given that the object is being talked about.

Because each caregiver/child dyad was given a separate set of objects (and also to ensure the generality of our results), we constructed a separate classifier for each dyad. The classifiers were evaluated using a tenfold cross-validation scheme in order to ensure that results were not due to overfitting. Results reported here are averaged across all ten test sets.[5]

Figure 2-10 shows the results of this analysis. The baseline probability of reference to an object (calculated as the proportion of utterances with a coded intention that was not "none") was relatively low in all three groups (6-month-olds, .41; 12-month-olds, .61; 18-month-olds, .52). We therefore report classification performance only for those sentences which had a referential intention. We evaluate classifiers created by fully crossing three sources of information: social cues exhibited by the mother—eyes, hands, and pointing; markers of the child's attention—the child's eyes and hands; and

[5] We experimented with a simple logistic regression classifier as well as regression-classification trees and found highly similar results for both alternative techniques.

discourse cues. All classifiers included baseline information about which objects were present in the field of view of the child.

We next constructed a regression model to understand the contributions of the different factors to classification accuracy. Classification accuracies were roughly normal in their distribution, so we constructed a multi-level linear model (Gelman & Hill, 2006). The form of this model included a separate intercept for each subject and age group as well as group-level effects of mother social cues, child social cues, and discourse information and their interactions. We then evaluated statistical significance using posterior simulation via Markov-chain Monte Carlo (Baayen, 2008). Coefficient estimates and significance are given in Table 2.2. Since each predictor is binary, coefficient weights are directly comparable.

We found highly significant and similar main effects of each of the three information sources, indicating that each one was effective for classification. We also found significant negative interactions between both mother and child predictors and child and discourse predictors. These two interactions both had approximately half the magnitude of the main effects, indicating that there was likely overlapping information about reference between the different sets of cues (hence the gain from having both was less than the gain expected from one plus the gain expected from the other). In the case of the mother/child social cue interaction, it seems likely that this overlap is due to cases of joint attention in which both participants are directly focused on a single object. The interaction between the child's social cues and discourse markers is less clear but may suggest that children's attention is "sticky," staying on the current focus of conversation and switching more gradually than the mother's reference.

Coefficients for age (a subject-level predictor in this model) were $\beta = .15$ for 6-month-olds, $\beta = .27$ for 12-month-olds, and $\beta = .09$ for 18-month-olds, indicating a slightly U-shaped trend in classification accuracy. We believe that this trend likely reflects the proportion of referential utterances (and hence the amount of data available to the classifier) as well as the diversity of the object set (which was greater for the older groups).

Summarizing this analysis, social cues and discourse together represent overlap-

Table 2.2: Coefficient estimates and significance for regression model predicting classifier accuracy on the basis of social and discourse information.

Predictor	Coefficient	95% CI	p-value
mot	0.22	0.17 – 0.27	< 0.001
chi	0.25	0.20 – 0.30	< 0.001
disc	0.19	0.14 – 0.24	< 0.001
mot:chi	-0.12	-0.19 – -0.05	< 0.001
mot:disc	-0.02	-0.09 – 0.06	0.89
chi:disc	-0.11	-0.19 – -0.04	< 0.001
mot:chi:disc	-0.02	-0.11 – 0.09	0.33

ping sources of information for determining what is being talked about. Taken together, these information sources allow for making surprisingly good guesses about the topic of an utterance without any additional linguistic information.

2.7 General Discussion

We began by asking whether social signals clearly indicated reference. We introduced a corpus of videos of child-directed speech across a range of ages, which we annotated with information about the objects visible to the child, the speakers' referential intentions, and the various social interactions of the child and caregiver with the objects. We found that, with the exception of pointing, social cues like eye-gaze and hand position were at best noisy indicators of referential intention, and that no individual cue revealed the speaker's referent more than a small portion of the time. In contrast, discourse continuity (assuming that the speaker was talking about the same thing as in their previous utterance)—when available—provided more reliable information about what was being talked about. A final set of simulations with a supervised classifier suggested that, despite their overlap, aggregating information across these information sources together provided the best estimate of referential intentions.

The current study is motivated by and provides support for the project of modeling early word learning as a process of statistical inference about speakers' referential intentions (Frank, Goodman, & Tenenbaum, 2009). Since no individual cue would consistently allow an observer of our corpus to infer what the speaker was talking about, an efficient learner would combine social information sources and aggregate this information over time using discourse cues. It remains an open question whether children conform to this normative prediction, but our results suggest that this strategy would be considerably more efficient than considering cues or utterances independently.

The current study has a number of limitations in both the scope of the dataset and the coding scheme. We address these in turn. First, our dataset was collected in a single type of experimental situation, in which infants and caregivers interacted over pairs of objects that were exchanged after short periods of time. This specificity may limit the strength of the generalizations that we can make from our analysis. Nevertheless, our measurements take a first step towards establishing norms which can be compared to measurements in more naturalistic studies.

Second, in order to make the coding task tractable across the relatively large corpus we used, it was important to break down the data at a relatively coarse temporal granularity. As a consequence, although we attempted to capture any look made to an object, our coding necessarily neglected some of the quick temporal dynamics of caregivers' and children's eye-movements. We hope that future work will use technical advances such as head cameras and eye-tracking to make more direct estimates of children's visual environment and the availability of social information from observed eye-gaze (R. Aslin, 2009; Yoshida & Smith, 2008).

Third, we have spoken throughout our analyses as though complex physical gestures can be individuated into discrete "cues" which can easily be associated with a particular utterance. Again, this approximation will almost certainly miss nuances of gestural communication (for example, anecdotally, caregivers in our sample often moved the object they were holding and talking about more than one that they were not talking about), but this approximation was necessary to code the volume of data reported here. Technical advances such as motion capture or motion recognition from computer vision may provide some traction on these questions (L. Smith, Yu, & Pereira, in press).

Finally, we have equated an eye-movement by the caregiver (which may or may not be visible to the child) with an eye-movement by the child (which controls what is visible to him or her). From the perspective of the child, this equivalence is not valid: the child's own eye-movements control what is being looked at, while the adult's eye-movements constitute an ephemeral signal to another person's attention. Nevertheless, in order to understand the relative validity of the child's own attention compared with external social information, we believe it is important to include these cues on the same footing as caregiver's cues.

An interesting possibility raised by our data is that the relative informativeness of what the child looks at and touches with respect to what the caregiver is talking about could provide support for the belief that words refer to the child's own interests rather than signaling the speaker's referential intentions (an *egocentric* theory of reference) (Baron-Cohen, Baldwin, & Crowson, 1997; Hollich et al., 2000). On this kind of account, a gradual shift from egocentric beliefs to speaker-centric beliefs might be due to gradual changes in the data that children receive, such that only over time do they disentangle their own attention from the following-in behavior of their caregiver. This disentangling would be neither purely a consequence of internal changes in the child nor external changes in caregiver input. Instead, the shift would be due to a gradual change in the style of interaction between the two. While younger, more passive infants often find their attention more closely aligned with that of their caregiver—due at last in part to the following-in behavior we observed—more mobile, more independent toddlers are likely to dissociate their own attention from that of the caregiver more frequently, in turn creating the data that they need in order to support a shift from egocentric to speaker-centric. One important prediction of this account is that, since there is significant variation with culture and socioeconomic status in the amount and kind of input that children receive from caregivers (Ochs, 1988; Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991; Hart & Risley, 1995; Hurtado, Marchman, & Fernald, 2008), this variation should affect children's perspective towards reference (as well as their vocabulary acquisition).

We believe that the results reported here constitute an important first step in quantifying the relative contributions of social and discourse information to the determination of reference. An account of the factors underlying the determination of reference will be crucial in understanding the emergence of communicative language. Understanding how children know what is being talked about is crucial to understanding how children learn the meanings of words.

Chapter 3

Learning words by assuming that speakers are informative[1]

Communicators do not always use language to code facts about the world directly. Instead, speakers and listeners rely on shared assumptions to allow them to communicate more efficiently than if every assumption were made explicit. If word learners take these implicit, shared assumptions into account, they should be able to make better guesses about what words mean than if they simply assume that language codes true facts about the world. In this chapter, we formalize the Gricean assumption that speakers choose their words in order to be informative about a target meaning, given some referential context. We show that this assumption leads to a derivation of the *size principle*—a statistical principle that, when applied to the current situation, implies that descriptions should be maximally unique relative to the context. Experiments with adults and children demonstrate that the word learning inferences are well fit by the assumption of informative communication. This work takes a first step towards formalizing the pragmatic assumptions necessary for effective communication in under-constrained, real-world situations.

[1] Joint work with Noah D. Goodman, Avril Kenney, and Joshua B. Tenenbaum. Parts of this chapter were published as Frank, M.C., Goodman, N. D., Lai, P., and Tenenbaum, J. B. (2009). Informative communication in word production and word learning. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society.*

3.1 Introduction

Imagine you are meeting with an interior decorator to decide on the details of a renovation. He pulls out several swatches of carpet which differ only in their texture. You indicate that you like one of them, and he says "oh, yes, that one is daxy." It seems instantly apparent that the novel word "daxy" refers to the specific texture of that particular swatch of carpet. *A priori*, "daxy" could have referred to any other aspect as well (including its size, color, or shape). But the only feature that would have been informative or useful for the decorator to name—given the contrast with the other swatches—was the carpet's texture.

In order to explain inferences of this type, philosophers and linguistics have suggested that language relies on shared assumptions about the nature of the communicative task. Grice (Grice, 1975) proposed that speakers follow (and are assumed by comprehenders to follow) a set of conversational maxims. In turn, if listeners assume that speakers are acting in accordance with these maxims, that gives them extra information to make inferences about speakers' intended meanings. For instance, we assume in this example that the decorator is following the Maxim of Quantity—"Be informative"—and therefore is using a descriptive adjective that appropriately picks the particular swatch out from its context. This adjective must then describe the carpet's texture, so we are justified in assuming that is the meaning of "daxy."

Other theories of communication also provide related tools for explaining this type of inference. For example, Sperber and Wilson (Sperber & Wilson, 1986) have suggested that there is a shared "Principle of Relevance" which underlies communication. On their account, the key part of this interaction is the shared knowledge between decorator and client that texture is the most relevant feature of the context; otherwise the inference is largely the same. In what follows we use the original Gricean language because it is simplest and best known, but our ideas do not depend specifically on Grice's formulation.

Positing shared assumptions between communicators allows for stronger inferences than are possible from pure physical causality. If our carpet sample (but not the others) had been exposed to sunlight and we observed that it had changed color, we would not be justified in concluding that this happened because of its texture unless others had also been exposed but had not changed. Although being called "daxy" and changing color are equivalent in some sense—both involve an event happening to one item in a set and not the others—the inferences licensed by a communicator *choosing* to say a particular word (and not others) are much stronger than those licensed by a physical event happening to it by chance without any rational agent causing it.

Could children be making the same kinds of inferences when they are learning the meanings of words as we made in learning the meaning of "daxy"? If so, this sort of inference is one tool by which children could eliminate some of the referential uncertainty inherent in learning a new word, whether the word is a property term or the name for an action or object (Quine, 1960; Gleitman, 1990). Many general theories of language acquisition assume that children bring some knowledge of the pragmatics of human communication to bear on the task of word learning (Bloom, 2002; E. Clark, 2003; Tomasello, 2003), but evidence on children's use of Gricean maxims is mixed.

Most accounts suggest that children younger than around five years have difficulty reasoning about the beliefs, knowledge, and perspective of others in communicative contexts (Glucksberg, Krauss, & Weisberg, 1966), though after this age they appear to be able to do so (Nadig & Sedivy, 2002). Gricean reasoning also has not been observed for children younger than four years, and only inconsistently before the age of six. For example, Conti and Camras (Conti & Camras, 1984) tested children on whether they could identify a maxim-violating ending to a story, and found that while four-year-olds could not do so, six- and eight-year-olds were able to succeed in this task. However, recent evidence from Eskritt et al. (Eskritt, Whalen, & Lee, 2008) showed that 4-year-olds could choose a puppet who followed the Maxim of Quantity over one who did not but 3-year-olds were not able to. In the same vein, children do not seem to be able to compute scalar implicatures (one possible example of a Gricean implicature, though cf. (Chierchia, Crain, Guasti, Gualmini, & Meroni, 2001; Gualmini, Crain, Meroni, Chierchia, & Guasti, 2001; Guasti et al., 2005)) until quite late (Noveck, 2001), although accounts differ on the age at which they first succeed (Papafragou & Musolino, 2003; Guasti et al., 2005). Thus, evidence largely suggests that children younger than four—who are likely to need sophisticated pragmatic inferences to infer word meanings—are the ones who are least able to make them.

Although evidence about sophisticated Gricean reasoning is mixed, infants still show impressive pragmatic abilities in the service of word learning. To take several influential examples, Akhtar and Tomasello (Akhtar et al., 1996) showed that two-year-olds could use the fact that an object was new to an experimenter to infer the meaning of a novel word that experimenter used. Baldwin (Baldwin, 1993) found that 18-month-olds were able to map a novel word to a referent that was hidden but signaled by the caregiver's attention to its location. In the most surprising recent demonstration of such abilities, Southgate, Chevallier, and Csibra (Southgate, Chevallier, & Csibra, 2010) showed that 17-month-olds were able to use knowledge about a speaker's false belief to map a novel name to an object, based on the speakers' naming of the location where she thought it was, not the location where it actually was.

In addition, recent evidence suggests that infants and young children have sophisticated statistical and inferential abilities at their disposal. These abilities are manifest in both word segmentation (Saffran, Aslin, & Newport, 1996; R. N. Aslin et al., 1998) (often referred to as "statistical learning") and word learning (Xu & Tenenbaum, 2007b, 2007a; L. Smith & Yu, 2008), as well as in more general reasoning about the world (Xu & Garcia, 2008; Xu & Denison, 2009). In one example of this kind of statistical inference, Xu and Tenenbaum (Xu & Tenenbaum, 2007b) showed that three- and four-year-old children were able to use the "suspicious coincidence" of seeing several examples of a word that were all from the same subordinate-level of a category (e.g. three Dalmatians) to infer that the novel word referred not just to dogs but to Dalmatians in particular. They argued that a Bayesian formulation of this idea of a coincidence would allow learners to choose the appropriate level of categorization for a new word with only a few positive examples. The particular mechanism for this inference was *strong sampling* (Tenenbaum & Griffiths, 2001): the idea that examples of a particular category are drawn at random from the members of that category. The assumption of strong sampling led to the *size principle*, that the probability of a particular example given a category was inversely proportional to the generality (size) of that category. In other words, although seeing three Dalmatians as examples of a word "dax" would be logically consistent with the hypothesis that "dax" picked out all dogs, those three examples would be far more likely if the category were Dalmatians.

Returning to the example we began with, it is still unknown whether young word learners are able to use the pragmatics of a particular context to learn the meanings of words, inferring for example that "daxy" refers to a texture because of the contrast between the texture of the named object and that of other objects. The goal of this chapter is to investigate this question. We do so by introducing a computational framework that describes this inferential situation in terms of Gricean inferences about the speakers' intentions. Though this framework could be applied to a number of situations in language production and comprehension, we introduce it in the context of word learning. The key assumption of our framework is that both listeners and speakers assume that communications are informative given the context.

Although the basis of our framework is general, making predictions within it requires a model of the space of possible meanings and how they map to natural language expressions. Thus, in order to make a first test of our framework, we study simple games that are similar to the "language games" proposed by Wittgenstein (Wittgenstein, 1953). In the games we study, the shared task of communicators is to identify an object from a set using one or a few words. This very restricted task allows us to define the possible meanings that communicators entertain, in turn allowing us to define an intuitive mapping between words and meanings: that a word stands for the subset of the context it picks out (its extension). Although these simplifications bring our tasks further away from natural language use, they also allow us to derive strong quantitative predictions from our framework.

We first show how, in these simple language games, our framework derives the

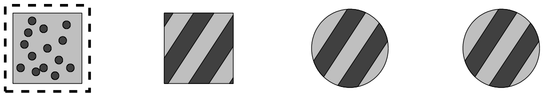

Figure 3-1: An example context: a polka-dot square, a striped square, and two striped circles. The dotted line shows the object the speaker intends to talk about.

size principle (Xu & Tenenbaum, 2007b; Tenenbaum & Griffiths, 2001). Then in Experiment 1, we test whether adults make inferences that are quantitatively congruent with the size principle when they are presented with contextual word-learning inferences like the interior-decorator example above. In Experiment 2, we show that three-year-olds also succeed in making such inferences in a version of the same task.

3.2 Modeling Informative Communication

Consider the context in Figure 3-1, representing the context in a language game. Imagine an English speaker in this game who is told to use a single word to point out the polka-dot square. Intuitively, she is likely to refer to the polka-dots since it would not be clear which object she was talking about if she talked about the shape of the object. Working from this intuition, a language learner—who knew that the speaker was pointing out the polka-dot square (perhaps because of some non-linguistic marker of intention, like a point or an eye-movement)—could make a very informed guess about the meaning of a novel word that she used to refer to that object. (This example is exactly parallel to the interior decorator example with which we began).

The intuition that a speaker would be likely to use the descriptor "polka-dot" to talk about the polka-dot square can be stated in terms of the assumption that speakers are choosing their words to best inform listeners of their intentions. If speakers did not try to communicate informatively, they would not necessarily choose labels that distinguished their intended referent from other possible objects. For example, in our game, an uninformative speaker—who nevertheless still respected the truth conditions of the language—might just as well have chosen to talk about the shape of the object as its pattern; correspondingly, a learner who did not assume an informative speaker would not be able to infer what the word they used meant.

We formalize these intuitions through an inferential model of language within this restricted world. We model the speaker as selecting speech acts in order to be informative, and derive predictions both for speakers and for learners who assume this about speakers.[2]

We assume that there is a fixed context C, consisting of some set of objects $o_1 ... o_m$. Both the speaker's intended meaning and the learner's guess about the speaker's intended meaning are probability distributions over C—that is, a meaning assigns a probability to each object in C. In the context of the simple language games we explored in Experiments 1 and 2, meanings simply carry information about which object in the physical context is the intended referent; however, the setup of the model does not change in the case that the context is non-physical (e.g., a set of possible propositions that could be communicated by the speaker). For us, intended meanings will be delta distributions picking out the speaker's intended referent precisely, but the same model will naturally handle vague intended meanings (which might arise, for instance, from incomplete knowledge).

We then assume there is a vocabulary $V = \{w_1, ..., w_p\}$ that is known to the speaker. Each word in V is a Boolean function over objects, indicating whether the word applies to that object. Thus the extension of word w in context C is the set of objects $\{o \in C | w(o) = 1\}$ that the word applies to. We denote by $|w|$ the size of a word's extension. We define the extensional meaning of w in context C to be the distribution:

$$\tilde{w}_C(o) = \begin{cases} \frac{1}{|w|} & \text{if } w(o) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

In this framework, a word is a function that can be applied to the objects in the

[2] One case which we do not treat here is the case of a teacher who is searching for the *best* example of a word to show a learner. This case is discussed in detail in (Shafto & Goodman, 2008), and we believe the current framework is compatible with their analysis.

context; a word's extension is the subset of objects in the context for which the function returns *true*.

3.2.1 Rational speaker

We assume that speakers act rationally according to Bayesian decision theory by choosing a word to (soft-)maximize their utility:

$$P(w|M_S, C) \propto e^{\alpha U(w;M_S,C)} \tag{3.2}$$

where M_S is their intended meaning and U is a utility function. For the purpose of the present games, we assume that only a single word is uttered; however, this framework can be modified to evaluate conjunctions of words as well. The decision noise parameter α measures the speaker's deviation from optimal. For all computations in the current chapter we set $\alpha = 1$, which recovers the standard Luce choice rule (Luce, 1963).

The speaker's goal is to choose the word which is most informative about M_S. The Kullback-Leibler divergence between two distributions X and Y, written $D_{KL}(X||Y)$, is the expected amount of information about X that is not contained in Y (Cover & Thomas, 2006). We formalize the goal of "informativeness" by assuming that the speaker's utility increases as the KL divergence between M_S and the literal meaning of the chosen word decreases:

$$U(w; M_S, C) = -D_{KL}(M_S||\tilde{w}_C) + F \tag{3.3}$$

where F represents other factors (such as utterance complexity) which affect the speaker's utility. For all simulations in the current paper, we set $F = 0$.

Combining Equations 3.2 and 3.3 and dropping F and α, we have

$$P(w|M_S, C) = \frac{e^{-D_{KL}(M_S||\tilde{w}_C)}}{\sum\limits_{w' \in V} e^{-D_{KL}(M_S||\tilde{w}'_C)}}, \tag{3.4}$$

which simplifies greatly in the language games we treat here. The speaker's intended meaning is a single object o_S (the value of M_S is 1 for o_S, 0 for all other objects). Thus:

$$
\begin{aligned}
D_{KL}(M_S||\tilde{w}_C) &= \sum_{o \in C} M_S(o) \log \frac{M_S(o)}{\tilde{w_C}(o)} \\
&= -\log(\tilde{w}_C(o_S)).
\end{aligned}
\tag{3.5}
$$

Substituting Equation 3.5 into Equation 3.4 gives:

$$
P(w|M_S, C) = \frac{\tilde{w}_C(o_S)}{\sum\limits_{w' \in V} \tilde{w}'_C(o_S)}. \tag{3.6}
$$

By Equation 3.1:

$$
P(w|M_S, C) \propto \begin{cases} \frac{1}{|w|} & \text{if } w(o) = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}
$$

Thus, given the set of simplifying assumptions we have made, the very abstract goal of "being informative" reduces to a simple formulation: choose words which pick out relatively smaller sections of the context. This recovers the "size principle" of Tenenbaum and Griffiths (Tenenbaum & Griffiths, 2001). This principle has more recently been rederived by Navarro and Perfors (Navarro & Perfors, 2009). Our work here can be thought of as a third derivation of the size principle—based on premises about the communicative task, rather than about the structure of generalization—that licenses its application to the kinds of cases that we have treated here.

3.2.2 Rational word learner

A Bayesian learner equipped with the theory of informative speakers captured in Equation 3.7 can use the knowledge that speakers are informative to learn words in an unknown language. A real language learner often has uncertainty about both the speaker's meaning, M_S, and the lexicon, L, the mappings between words and meanings (Frank, Goodman, & Tenenbaum, 2009). Although our framework can be extended to this case, we focus here on the case that we treat in our experiments: the learner knows M_S and the possible set of meanings that words could map to and has only to infer facts about L.

By Bayes' rule:

$$P(L|w, M_S, C) \propto P(w|L, M_S, C)P(L) \tag{3.8}$$

For simplicity, let us assume that the object has two truth-functional features f_1 and f_2, that there are two words in the language w_1 and w_2, that there are only two possible lexicons $L_1 = \{w_1{=}f_1, w_2{=}f_2\}$ and $L_2 = \{w_1{=}f_2, w_2{=}f_1\}$, and that there is a uniform prior on vocabularies. Then:

$$\begin{aligned} P(L_1|w_1, M_S, C) &= \frac{P(w_1|L_1, M_S, C)}{P(w_1|L_1, M_S, C) + P(w_1|L_2, M_S, C)} \\ &= \frac{|f_1|^{-1}}{|f_1|^{-1} + |f_2|^{-1}}, \end{aligned} \tag{3.9}$$

where $|f|$ indicates the number of objects with feature f (substituting Equation 3.7 for the second step).

Returning to the example in Figure 3-1, imagine that the speaker points to the polka-dot square and says "feppy" (a novel word w that you have never heard before). We used Equation 3.9 to calculate the probability that learners judge that w means `polka-dot` as opposed to `square`:

$$\begin{aligned} P(w = f_1|M_S, C) &= \frac{|\texttt{polka-dot}|^{-1}}{|\texttt{polka-dot}|^{-1} + |\texttt{square}|^{-1}} \\ &= \frac{\frac{1}{1}}{\frac{1}{1} + \frac{1}{2}} = \frac{2}{3} \end{aligned}$$

Thus, our prediction is that learners should be around 67% confident that "feppy" means polka-dot.

3.3 Experiment 1

Our first experiment tested the hypothesis that word learners' inferences conform to the informativeness framework described above. We asked adults for quantitative judgments about the meanings of novel words in situations like Figure 3-1. In order to gather data across a large number of conditions (including many different predicates and many different arrangements of objects and features), we made use of Amazon Mechanical Turk (`www.mturk.com`), an online crowd-sourcing tool. Mechanical Turk allows users to post small jobs to be performed quickly and anonymously by workers (users around the world) for a small amount of compensation. In order to elicit a continuous, quantitative judgment, we posted many displays of the same type as Figure 3-1, and for each one asked participants to bet on whether a novel adjective referred to one or the other property of the object with the box around it. (This betting measure gives us an estimate of speakers' subjective probability, rather than a purely qualitative judgment.)

3.3.1 Materials and Methods

Participants

We posted 1500 separate trials. We excluded individual trials on the basis of a manipulation check (described below) as well as if their answers did not add up to 100 or if they failed to complete the questionnaire fully. The final sample consisted of 1271 trials, contributed by 221 unique individuals. The majority of workers (173) performed only a single trial, but a small number of workers performed many more.

Stimuli and Procedure

Each Mechanical Turk job consisted of a single web page displaying a set of objects (as in Figure 3-1). For each set, each object was assigned a color (red/blue/green), a shape (circle/square/cloud), and a texture (solid/polka-dot/striped) feature. Two feature dimensions of interest were chosen to vary across the set, while the third was held constant. Objects were chosen such that the target object with the box around

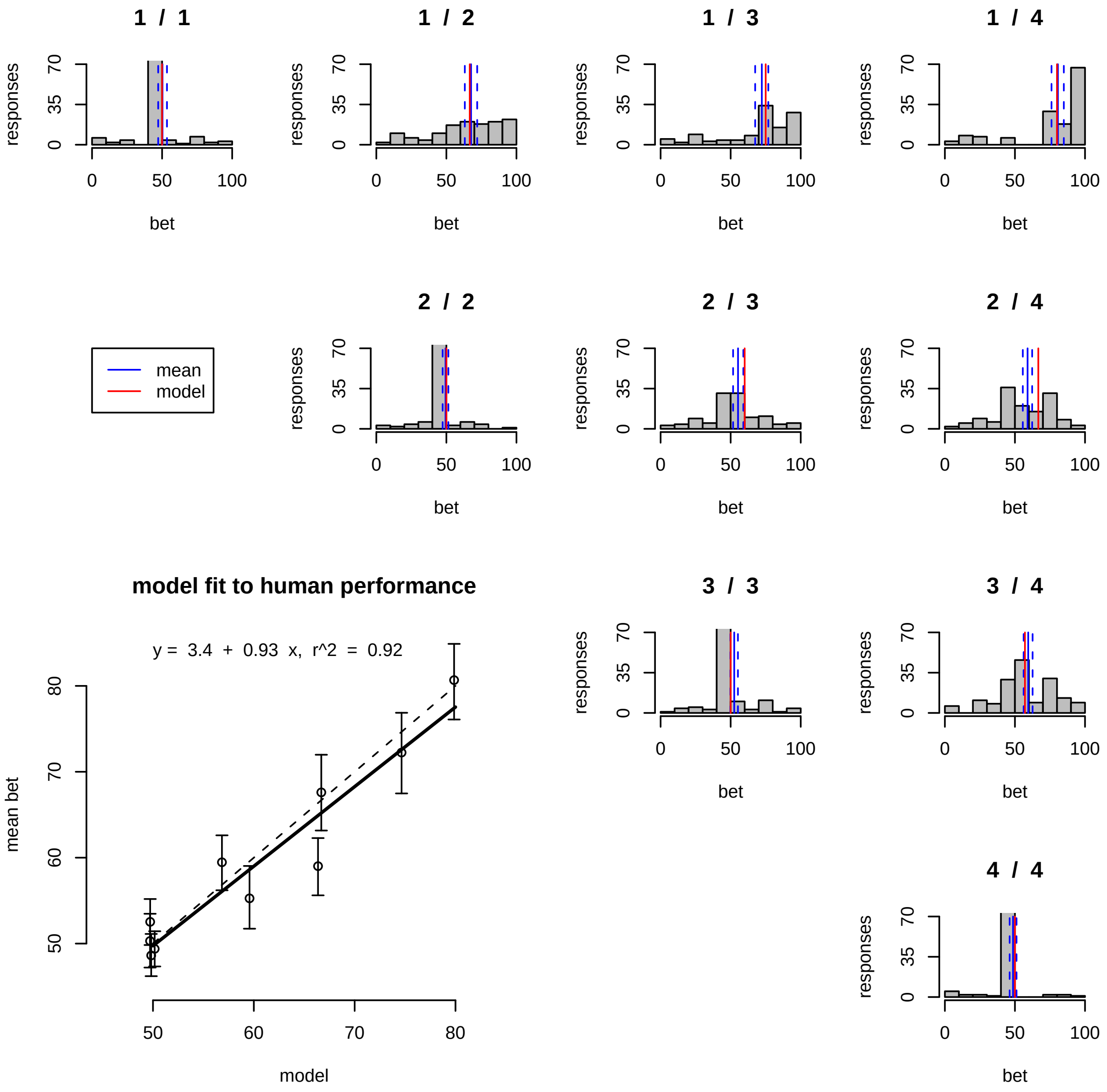

Figure 3-2: Each subplot shows the histogram of participants' bets on the more informative feature (e.g., polka-dot), along with the mean and 95% confidence intervals (blue) and the model predictions (red). Confidence intervals are computed via non-parametric bootstrap. The inscribed plot (lower-left) shows mean bets for each condition plotted by model predictions with 95% confidence intervals. X positions are jittered slightly to minimize overplotting. Line of best fit is shown in red.

it was assigned to share its feature along each of the varying dimensions with 0, 1, 2, or 3 of the other objects.

The manipulation of interest was the number of objects that shared features with the target object. For instance, in Figure 3-1, only the target object is polka-dot while the target object and one other are square. We denote this trial type as a 1/2 trial. Trials were chosen so that they would be uniformly distributed over the 10 possible combinations of shared features (1/1; 1/2; 1/3; 1/4; 2/2; 2/3; 2/4; 3/3; 3/4; 4/4). All other aspects of trials (including position and features of all objects) were randomized. (Note that e.g. 2/1 trials are impossible because the feature that applies to fewer objects is always the preferred one).

To make sure that participants were inspecting the objects (rather than simply responding at random), we included a manipulation check (Oppenheimer, Meyvis, & Davidenko, 2009): we asked them to report the number of objects with each of the two features of the target in the dimension of interest (e.g., How many objects are polka-dot? How many objects are square?).

Participants were then instructed that someone speaking a foreign language used a word to refer to the object with the box around it and were asked to guess whether the word referred to one feature or the other of the target object (e.g., square or polka-dot). They were told that their guess would take the form of a bet and that they should allocate $100 between the two possible meanings for the word.

3.3.2 Results and Discussion

Since we randomized the dimensions used in each trial, we averaged across this aspect of the data and considered only the distribution of bets relative to the number of objects with each of the two possible features. We plot the histogram of responses for each condition in Figure 3-2, with mean performance for each condition plotted against model predictions in the inset figure.

When there were equal numbers of objects with each feature (e.g., two polka-dot and two square objects)—represented by the diagonal in Figure 3-2—mean bets were very close to $50, reflecting equal probability. In contrast, in the case shown in Figure 3-1, there is only one object in the more informative category (polka-dot) but there are two in the less informative category (square). Our informativeness model predicted an average bet of \$67 in this condition, very close to the average bet of \$65.9. More generally, the relationship between the values predicted by the informative communication model and the experimentally determined means was high ($r^2 = .92$, $p < 10^{-6}$). Thus, in their inferences about the meanings of novel words, participants' mean judgments conformed quite precisely to the predictions of our model.

These results are not unique to participants on Mechanical Turk: we replicated this study with two different populations with entirely different stimuli. The first replication was with a sample of 700 MIT undergraduates who were asked to complete three trials via email, reported in (Frank, Goodman, Lai, & Tenenbaum, 2009); stimuli in that study again varied on shape, texture, and color. The second study included a sample of 100 MIT undergraduates and community members. They were asked to complete a paper-and-pencil study, also including three trials, while visiting the lab for unrelated research; stimuli in that study were a set of cartoon insects that varied on properties like what kind of tail they had or what shape their head was. The fit between the model and participants' average performance across conditions in these two studies was $r^2 = .86$ and $r^2 = .79$, respectively. These converging results suggest that results were not obtained as a consequence of the population being tested, the fact that some individuals completed many trials, or the method of online presentation.

3.4 Experiment 2

In our second experiment, we asked whether children would also be able to make use of the informativeness of features to learn the meanings of novel adjectives. Because of known biases in children's word learning (Landau, Smith, & Jones, 1988; Bartlett, 1978), we chose to vary stimuli on two dimensions: texture and pattern. We created sets of real objects that had a novel texture and a novel pattern (an example trial is

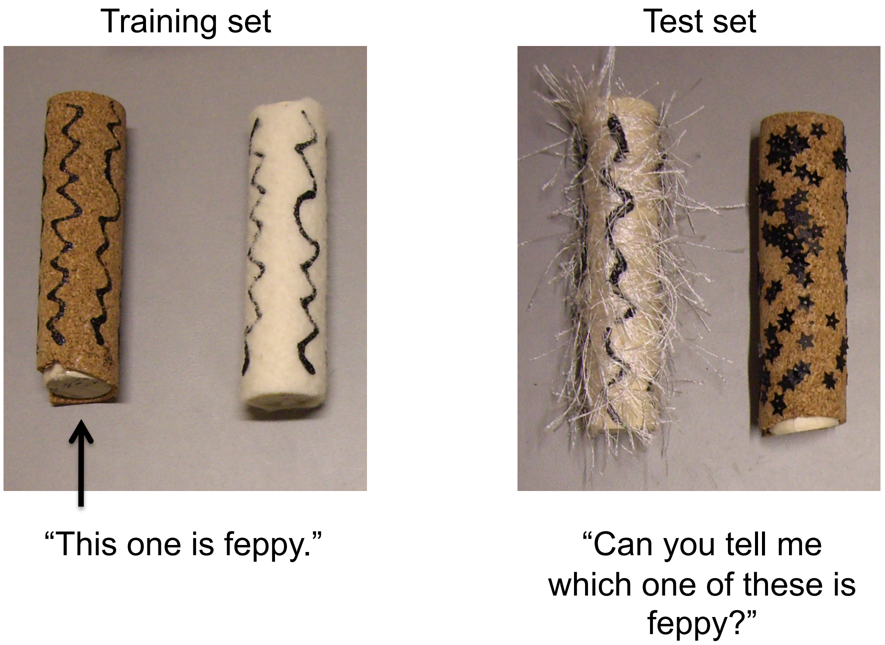

Figure 3-3: Stimuli for a trial in Experiment 2. The predicted answer is the test item on the right, because it shares a texture (cork) with only the indicated training example, while the left test item shares the squiggle feature with both training examples.

shown in Figure 3-3). We showed the children two objects that shared one feature but differed on the other and, indicating one of the objects, told the child that it was "feppy." To determine which meaning had been inferred for the novel word, we then showed the child a new pair of objects, one of which had the feature that had been unique to the "feppy" object, and the other of which had the feature that the "feppy" object had shared with the other object" and asked them which one of these two new objects was also "feppy." If children were able to use information from the context to identify which property was being talked about, this would provide evidence for the claim that they assume speakers are informative in choosing which label to apply to an object. Further, this evidence would support the use of the assumption of informativeness to learn which property a novel adjective maps to.

3.4.1 Methods

Participants

Sixteen children between three and four years old (mean age = 3;7.6) took part in the informativeness condition and a separate group of 16 3–4 year olds (mean age = 3;4.23) took part in the baseline condition. All participants were recruited during a visit the Boston Children's museum.

Stimuli

Pilot testing revealed that children's judgments were very sensitive to the particular texture and pattern that were being used, so we created a large set of stimuli and used Amazon's Mechanical Turk online service to collect norming data about which property of an object was more likely to be referred to by a novel adjective. We then constructed the stimuli for use with children by choosing two base object shape and assigning each a pair of novel properties that were relatively close in the proportion of the time they were chosen as the meaning of a novel adjective. The first base object was popsicle sticks and the second was wooden dowels (example stimuli are pictured in Figure 3-3).

Procedure

Each child completed two trials. In the Informativeness condition, the experimenter presented the child with a set of two objects on a cardboard tray in each trial. The experimenter then picked up one object and told the child that "This object is feppy [/ziffy]." The experimenter then hid the first two objects and put out two more on the tray and asked the child which one they thought was feppy/zipfy. All trials were videotaped for offline coding. Across trials we counterbalanced which object was named with which label, which property was most informative, which side the named object was on, and which side the correct answer was on. The Baseline condition was identical to the Informativeness condition except that only one object (the target object) was presented in the first set of objects.

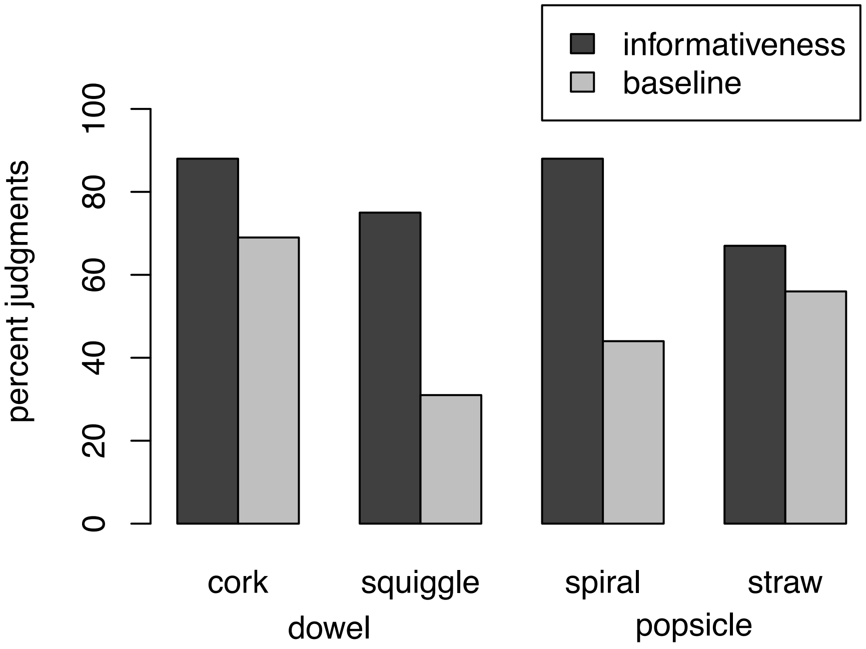

Figure 3-4: Percentage of children's judgments in Experiment 2, plotted by base object and feature. Light gray bars show proportion of baseline judgments that a novel name applied to a particular feature; dark gray bars show judgments that the name applied to the same feature after the object was named in a context where that feature was informative.

3.4.2 Results and Discussion

Figure 3-4 shows the percentage of children that picked each feature, both when it was most informative and in the Baseline condition. For both objects, both features were chosen more often by children in the Informativeness condition than in the Baseline condition. The baseline across all four items was necessarily .5, thus we used a binomial test to test whether children picked the more informative feature more often than chance (25/32 trials, $p = .0001$).

3.5 General Discussion

A model of language as a code for facts does not account for the rich interpretations that language users are able to extract from limited data. Instead, research on language use (Grice, 1975; Sperber & Wilson, 1986) posits that speakers and listeners share assumptions about the nature of the communicative task that allow meanings to be inferred even in the presence of ambiguous or limited data. In our work here, we investigated the implications of this research for language acquisition. We began by asking whether children are able to use the assumption that speakers act rationally to convey their intended meaning—equivalent to Grice's maximum of informativeness—to learn the meanings of novel words in ambiguous contexts. We showed that, in restricted referential worlds, the assumption of informativeness derives the "size principle," a generalization which describes children's behavior in word learning (Xu & Tenenbaum, 2007b). We then showed empirical evidence that adults' quantitative judgments matched the predictions of this framework and that three-year-olds were able to succeed in learning a novel adjective in a comparable situation. These results provide support for the suggestion that word learners may make inferences about the meanings of words based on the underlying assumption that speakers are acting rationally.

The work we presented here is related to formal work on human communication in a number of different traditions. Early work by Rosenberg and Cohen (Rosenberg & Cohen, 1964, 1966) described a similar formalism, which they used for disambiguating synonym pairs, though they did not apply their framework to word learning. Our work is also related to game-theoretic approaches to pragmatics from linguistics (Benz, Jäger, & Van Rooij, 2005), though it differs from these approaches in that it does not rely on complex, recursive computations but instead on a simple formulation that can be computed whenever the space of meanings and mappings to linguistic expressions is known. Finally, in natural language processing, researchers in the field of referring expression generation (Reiter & Dale, 1997; Dale & Reiter, 1995) have described a similar approach to deriving Gricean maxims from assumptions about speakers' rationality. The convergence of interest in this general topic across widely-varying fields suggests a need for a formal framework for human communication that goes beyond information theory to incorporate insights from the study of pragmatics.

Our work here takes a first step towards this goal. Rather than providing a well-worked out model of a broad range of tasks, we view our contribution instead as describing a framework for adding pragmatic inferences to systems of meanings. In order to calculate inferences in this framework, it is necessary to have a model of the space of possible meanings and their mappings to linguistic forms. While we used very simple examples of each of these in our experiments, in principle our framework could be applied to a propositional model of semantics in combination with a grammar for creating sentences that express those propositions. More generally, with definitions of a meaning space and linguistic mapping, our framework can be used to make predictions in a wide variety of situations, ranging from simple non-linguistic communication experiments to complex cases like scalar implicature and anaphora resolution. Our hope is that future work will make use of this framework to address a broad range of questions in language use and language learning.

Chapter 4

Using speakers' referential intentions to model early cross-situational word learning[1]

Word learning presents a "chicken-and-egg" inference challenge. If a child could understand speakers' utterances, it would be easy to learn the meanings of individual words; and once a child knows what many words mean, it is easy to infer speakers' intended meanings. To the beginning learner, however, both individual word meanings and speakers' intentions are mysterious. We describe a Bayesian model that solves these two inference problems in parallel, rather than learning exclusively from the inferred meanings of utterances or relying only on cross-situational word-meaning associations. Our model infers pairings between words and object concepts from CHILDES data with high precision. Our model also explains a variety of behavioral phenomena from the word learning literature, as the result of making probabilistic inferences about speakers' intentions. These phenomena include mutual exclusivity, one-trial learning, cross-situational learning, the role of words in object individuation, and the use of inferred intentions to disambiguate reference.

[1] Published as Frank, M. C., Goodman, N. D., and Tenenbaum, J. B. (2009). *Psychological Science*, 20, 578–585.

4.1 Introduction

When children learn their first words, they face a challenging joint inference problem: they are both trying to infer what meaning a speaker is attempting to communicate at the moment a sentence is uttered and trying to learn the more stable mappings between words and referents that constitute the lexicon of their language. With either of these pieces of information, their task becomes considerably easier. Knowing the meanings of some words, a child can often figure out what a speaker is talking about; on the other hand, inferring the meaning of the speaker's utterance allows the child to work backwards and learn basic-level object names with relative ease. However, for a learner without either of these pieces of information, word learning is a hard computational problem. Following Quine's metaphor, a word learner is climbing the inside of a chimney, "supporting himself against each side by pressure against the others" (Quine, 1960).

Many accounts of word learning focus primarily on one aspect of this problem. Social theories suggest that learners rely on a rich understanding of the goals and intentions of speakers and assume that—at least in the case of object nouns—once the child understands what is being talked about, the mappings between words and referents are relatively easy to learn (St. Augustine, 397/1963; Baldwin, 1993; Bloom, 2002; Tomasello, 2003). These theories must assume some mechanism for making mappings, but this mechanism is often taken to be deterministic and its details are rarely specified. In contrast, cross-situational accounts of word learning take advantage of the fact that words often refer to the immediate environment of the speaker, allowing learners to build a lexicon based on consistent associations between words and their referents (Locke, 1847; Siskind, 1996; L. Smith, 2000; Yu & Ballard, 2007).

Computational models of word learning have primarily followed the second, cross-situational strategy. Models using connectionist (Plunkett, Sinha, Møller, & Strandsby, 1992), deductive (Siskind, 1996), competition-based (Regier, 2005), and probabilistic methods (Yu & Smith, 2007) have had significant successes in accounting for many phenomena in word-learning. However, speakers often talk about objects that are not visible and actions that are not in progress at the moment of speech (Gleitman, 1990), adding noise to the correlations between words and objects. Thus, cross-situational and associative theories often appeal to external social cues like eye-gaze (L. Smith, 2000; Yu & Smith, 2007), but they are used as markers of salience (the "warm glow" of attention), rather than as evidence about internal states of the speaker as in social theories.

More generally, cross-situational theories address only one part of the learners' task—they are able to learn words, but they do not use the words that speakers utter to infer the speakers' intended meanings. By focusing only on the long-term mappings between items in the lexicon and referents in the world, purely cross-situational models treat the complex and variable communicative intentions of speakers as noise to be averaged out via repeated observations or minimized via the use of attentional cues, rather than as an important aspect of communication to be used in the learning task.

Here we present a Bayesian model that captures both aspects of the word learning task: it simultaneously infers what speakers are attempting to communicate and learns a lexicon. We first present the structure of the model and show that it obtains competitive results in learning from corpus data. We then show how the probabilistic structure of the model allows it to predict experimental results such as mutual exclusivity, one-trial word learning, and rapid cross-situational learning, while its explicit representation of intention allows it to predict results on object individuation and the use of intentional cues.

4.2 Model design

Our model consists of a set of variables representing the word learning task and a set of probabilistic dependencies linking these variables in accordance with our assumptions about the task (Figure 4-1). The variables represent the lexicon of the language being learned, the referential intentions of the speaker, the words she utters, and the learner's physical context at the time of the utterance. We define the relationships

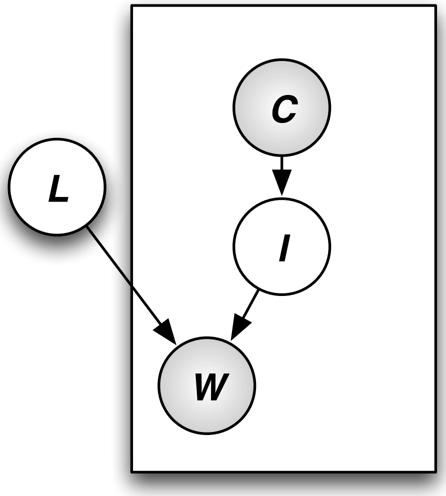

Figure 4-1: The graphical model representing dependence relations in our model. C, I, and W represent the objects present in the context C (we note this variable as O in the rest of the chapter to emphasize that it only represents objects), the objects that the speaker intends to refer to, and the words that the speaker utters, respectively. These variables are related within each situation s (shown by the plate under these variables), and the words that the speaker utters are additionally determined by the lexicon of their language, L, which does not change from situation to situation (and hence lies outside of the plate).

between these variables via two assumptions. First, what speakers intend to say is a function of the physical world around them. Second, the words that speakers utter is a function of what they intend to say and how those intentions can be translated into the language they are speaking. With these assumptions and an observed corpus of situations—utterances and their physical context—our model can work backwards using Bayesian inference to find the most likely lexicon.

Though the speaker's intentions could in principle be very complex, we limit ourselves here to the task of learning names for objects. Thus, we represent the physical context of an utterance as the set of objects present during the utterance, the speaker's referential intention as the object or objects she intends to refer to, and the lexicon as a set of mappings between words and objects. We also assume that objects are identified as instances of basic-level object categories, putting aside the challenge of identifying the particular aspect of an object being named (Xu & Tenenbaum, 2007b).

Formally, our model defines a probability distribution over unobserved lexicons L and the observed corpus C of situations. Our goal is to infer the lexicon with the highest posterior probability. We find this posterior probability using Bayes' rule:

$$P(L|C) \propto P(C|L)P(L) \tag{4.1}$$

Bayes' rule factors the posterior probability of a lexicon given the corpus into two terms, the likelihood of the corpus given the lexicon and the prior over lexicons. We chose a parsimony prior, making lexicons exponentially less probable as they include more word-object pairings: $P(L) \propto e^{-\alpha|L|}$. The choice of this simple prior puts most of the work of the model in the likelihood term $P(C|L)$, which captures the learner's assumptions about the structure of the learning task.

This term can be written as a product over situations of the probability of the components of the corpus (the words W_s, objects O_s, and speaker's intentions I_s for each situation s), given the lexicon:

$$P(C|L) = \prod_{s \in C} P(W_s, O_s, I_s|L). \tag{4.2}$$

We can now use our assumptions about the structure of the task to factor Equation 2. First, W and O are conditionally independent given I (as shown in Figure 4-1). Thus we can rewrite the right-hand side as a product of $P(W_s|I_s, L)$, the probability of the words given the speaker's referential intentions and the lexicon, and $P(I_s|O_s)$, the probability of the speaker's intentions given the physical context. Second, since we cannot directly observe the speaker's referential intention, we sum over all possible values of I_s under the constraint that $I_S \subseteq O_s$, that is, that the relevant subset of possible intentions are those that refer to a subset of the objects in the physical context. Since speakers often refer to objects outside of the physical context, I_s can be empty (c.f. (Gleitman, 1990)). We rewrite Equation 2 as

$$P(C|L) = \prod_{s \in C} \sum_{I_s \subseteq O_s} P(W_s|I_s, L) \cdot P(I_s|O_s) \tag{4.3}$$

For simplicity (and since we have no information about I_s other than the words that are uttered) we set $P(I_s|O_s) \propto 1$ so that all possible intentions are equally likely.

To complete our definition of the model, we define the term $P(W_s|I_s, L)$ by assuming that the words $\{w_1 ... w_n\}$ in W_s are generated independently (ignoring any syntax), and that there can be two possible causes for uttering a word. A word is either uttered *referentially*—in order to refer to an object in the speaker's intention set—or *non-referentially.* The probability of a word being uttered if it is used referentially (P_R) is the probability that it will be chosen from the lexicon to refer to any of the intended referents. The probability of a word being uttered if it is used non-referentially (P_{NR}) is just the probability that it will be picked from the lexicon at random, independent of the speakers' referential intention. Verbs, adjectives, and function words are generated non-referentially as well as object nouns for which the relevant object is not currently present. The parameter γ represents the probability that a word is used referentially in any given context. Thus, we have

$$P(W_s|I_s, L) = \prod_{w \in I_s} [\gamma \cdot \sum_{o \in I_s} \frac{1}{|I_s|} P_R(w|o, L) + (1 - \gamma) \cdot P_{NR}(w|L)] \tag{4.4}$$

The probability of a word being used referentially for an object, $P_R(w|o, L)$, is the probability that the word is chosen uniformly from the set of words linked to that object in the lexicon. If there are, for example, two words linked to an object in the lexicon, each word has a probability of 0.5 to be used to refer to that object; if a word is not linked to an object, its referential probability for that object is 0. The non-referential probability of a word being used is the probability of a word being picked from the full set of words observed in the corpus with probability proportional to 1 if it is not in the lexicon and to κ otherwise. Thus, if $\kappa < 1$, it is less likely (but not impossible) that a word which is already in the lexicon will be uttered in a context in which the speaker does not use it referentially.

In the simulations below, we employ stochastic search methods using simulated tempering (Marinari & Parisi, 1992) to find the lexicon with the maximum a posteriori probability given an observed corpus. In our online supplementary materials we provide code, corpora, and details of our search methods and simulations.

4.3 Corpus evaluation

4.3.1 Corpus

We coded two video files of 10 minutes each from the Rollins section of CHILDES (me06 and di03) (MacWhinney, 2000) in which two preverbal infants and their mothers played with a set of toys. Each line of the transcripts was annotated with a list of all midsize objects judged visible to the infant.[2]

[2] These videos are the same as those used by Yu and Ballard (2007); the annotations are our own.

4.3.2 Alternative models

For comparison, we implemented several other models of cross-situational word learning using co-occurrence frequency, conditional probability, and point-wise mutual information. We also implemented IBM Machine Translation Model I (Brown, Pietra, Pietra, & Mercer, 1993), the statistical machine translation model used by (Yu & Smith, 2007). We used the translation model to compute association probabilities both for objects given words (as in (Yu & Smith, 2007)) and words given objects.

4.3.3 Evaluation

We evaluated all models both on the accuracy of the lexicons they learned and on their inferences regarding the speakers' intent. Each of the comparison models produced a single summary statistic linking words and objects (e.g., association probability). We thresholded this statistic to find the lexicon maximizing the model's possible F-score.[3] We then used each model to make guesses about the speaker's intended referents for each utterance. For our model, we chose the intention with the highest posterior probability given the best lexicon; for the comparison models, we assumed that the intended referents were those objects for which the matching words in the best lexicon had been uttered. We computed scores relative to a gold-standard lexicon and set of intents, both created by a human coder. The gold-standard lexicon incorporated all standard word-object pairings (for a lamb toy, "lamb"), plurals ("lambs"), and babytalk ("lambie"); the gold-standard intent contained the human coder's best judgment of which objects in the visual context were being referred to in each of the speaker's utterance.

4.3.4 Results

The Bayesian model substantially outperformed the comparison models, especially with respect to the lexicons the model learned (Tables 4.1, 4.2 and 4.3). This advan-

[3] F-score is the harmonic mean of precision (proportion of the lexicon that was correct) and recall (proportion of total correct pairings included in the lexicon).

Table 4.1: Precision, recall, and F-score of the best lexicon found by each model when run on the annotated data from CHILDES.

Model	Precision	Recall	F-score
Association frequency	.06	.26	.10
Conditional probability (object|word)	.07	.21	.10
Conditional probability (word|object)	.07	.32	.11
Mutual information	.06	.47	.11
Translation model (object|word)	.07	.32	.12
Translation model (word|object)	.15	.38	.22
Intentional model	.67	.47	.55
Intentional model (one parameter)	.57	.38	.46

Table 4.2: Precision, recall, and F-score for the referential intentions found by each model, using the lexicons scored in Table 4.1.

Model	Precision	Recall	F-score
Association frequency	.27	.81	.40
Conditional probability (object|word)	.59	.36	.45
Conditional probability (word|object)	.32	.79	.46
Mutual information	.36	.37	.37
Translation model (object|word)	.57	.41	.48
Translation model (word|object)	.40	.57	.47
Intentional model	.83	.45	.58
Intentional model (one parameter)	.77	.36	.50

Table 4.3: The best lexicon found by the intentional model.

Word	Object
bear	bear
bigbird	bird
bird	duck
birdie	duck
book	book
bottle	bear
bunnies	bunny
bunnyrabbit	bunny
hand	hand
hat	hat
hiphop	mirror
kittycat	kitty
lamb	lamb
laugh	cow
meow	baby
mhmm	hand
mirror	mirror
moocow	cow
oink	pig
on	ring
pig	pig
put	ring
ring	ring
sheep	sheep

tage was robust across systematic variation of the model's three free parameters (α, the strength of the prior on lexicon size; γ, the proportion of words that are used referentially; and κ, the probability of a word in the lexicon being used non-referentially). In addition, the advantage remained when κ and γ were set to their maximum a posteriori values (the empirical Bayes estimate, see (Carlin & Louis, 1997)), reducing the number of free parameters to one—the same number as the baseline models.

Both the simple statistical models and the translation model found a large number of spurious lexical items; the best lexicons found by these models were considerably larger than the best lexicon found by our model.[4] The high precision of the lexicon found by our model was likely due to two factors. First, the distinction between referential and non-referential words allowed our model to exclude words from the lexicon which were used without a consistent referent. Second, the ability of the model to infer an empty intention allowed it to discount utterances in which no object in the immediate context was being talked about.

4.4 Prediction of Experimental Results

As a consequence of its structure, our model exhibits a graded preference for certain kinds of lexicons and utterance interpretations. First, lexicons should be sparse because the simplicity prior we impose biases the model against adding word-object mappings that do not increase the likelihood of the data. Second, one-to-one lexicons tend to be preferred, if they are consistent with the observed data, because having multiple words that can refer to an object reduces the probability of any single word being used consistently to refer to that object. Finally, the model prefers that people have intentions to talk about the objects that are present, since words that are generated referentially from an intention to talk about an object have higher likelihood than words that are generated non-referentially at random from the entire vocabu-

[4] The performance we report for the translation model is considerably lower than that reported by (Yu & Smith, 2007). Several factors may have contributed to this difference: the speech transcripts used in our study were taken from CHILDES, while those in the Yu and Ballard study were automatically extracted; our corpus coding may have included different objects in each situation; and our gold-standard lexicon differed from Yu and Ballard's.

lary of the language. These three preferences allow the model predict a number of empirical results in early word learning.

4.4.1 Rapid cross-situational word learning

Recent work has provided strong evidence that both adults and children are able to learn associations between words and objects even in the absence of individually unambiguous trials (L. Smith & Yu, 2008; Vouloumanos, 2008; Vouloumanos & Werker, 2009; Yu & Ballard, 2007). Because the statistics in these experiments so sharply favor the correct lexicon, our model and all of the comparison models successfully found the correct word-object pairings with perfect precision and recall when presented with the artificial lexicons from (Yu & Ballard, 2007). Thus, these experiments are not strongly diagnostic among competing models.

4.4.2 Mutual exclusivity

In classic demonstrations of mutual exclusivity, a child is presented with two objects, one familiar and one novel. The experimenter asks "Can you hand me the dax?" and the child hands over the novel object, indicating that she has correctly inferred that the novel name refers to it (Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992; E. M. Markman & Wachtel, 1988). Markman and colleagues (E. Markman, 1991; E. M. Markman & Wachtel, 1988; E. Markman, Wasow, & Hansen, 2003) have suggested that children possess a principle of mutual exclusivity which leads them to prefer lexicons with only one label for each object. Other researchers have suggested alternate explanations, including more limited principles that are learned with experience (Golinkoff, Mervis, & Hirsh-Pasek, 1994; Mervis & Bertrand, 1994) or more general pragmatic principles (E. Clark, 1988, 2003).

Without building in an explicit assumption of mutual-exclusivity, our model shows a soft preference for one-to-one mappings. We tested our model in the classic mutual exclusivity paradigm (E. M. Markman & Wachtel, 1988) and found it correctly inferred that the novel word mapped to the novel object. We scored four possible

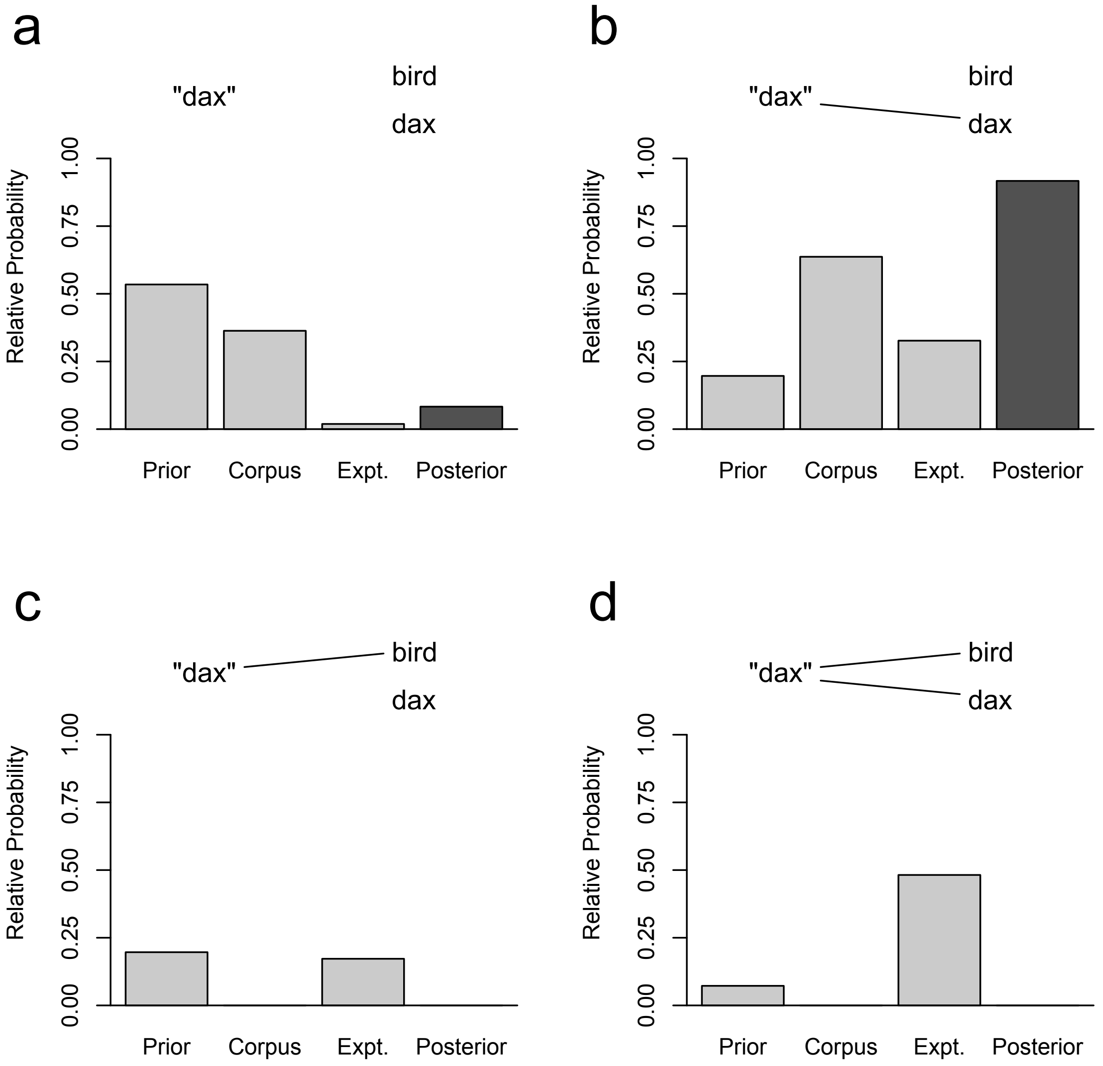

Figure 4-2: Schematic depiction of possible hypotheses in a mutual exclusivity experiment. If the experimenter utters the novel word "dax" in the presence of a novel object (a DAX) and a known object (a BIRD), the learner can decide the word refers to both, one or the other, or neither. Each panel represents one of these options, with a line between a word and an object signifying that the link is represented in the lexicon. The corpus likelihood, the likelihood of the experimental situation, the prior probability of the lexicon, and the posterior (total) probability, normalized across the four lexicons, are shown for each hypothesis.

lexicons (Figure 4-2) on our original CHILDES corpus extended with a mutual exclusivity scenario (the word "dax" is uttered in the presence of a BIRD toy and a novel object DAX). Lexicons where "dax" was mapped to the familiar object BIRD (Figure 4-2, part C, part D) were unlikely with respect to the original corpus because each sentence where the word "bird" was uttered became less likely due to the unrealized possibility of hearing "dax" as well. Learning no new words (part A) was favored by the prior because it involved no growth in the size of the lexicon, but received low likelihood on the experimental scenario (since the word "dax" was not in the lexicon). Overall, our model preferred the correct lexicon (part B).

This result is not unique to our model: the basic finding of mutual exclusivity is captured by many of the baseline models we tested. In the example above, the conditional probability of the word "dax" given seeing a BIRD is quite low, while the probability of the word "bird" given the object BIRD is still very high. Combined with the demonstration that adults and infants are able to use some sort of statistical information in cross-situational learning tasks (L. Smith & Yu, 2008; Vouloumanos, 2008; Vouloumanos & Werker, 2009; Yu & Ballard, 2007), the success of our model and others suggests that it is not necessary to posit domain-specific principles to account for findings of mutual exclusivity.

4.4.3 One-trial learning

Another classic result in the literature on word learning is the ability of children to learn a new word from only one or a small number of incidental exposures (Carey, 1978; Markson & Bloom, 1997). Our model and the comparison models predict that there are some situations which—in conjunction with the learner's previous experiences—can provide sufficient evidence for a word's referents to be inferred after a single exposure; in fact, the mutual exclusivity experiment described above provides one such situation. We next turn to a set of experiments which to the best of our knowledge cannot be captured by the comparison models.

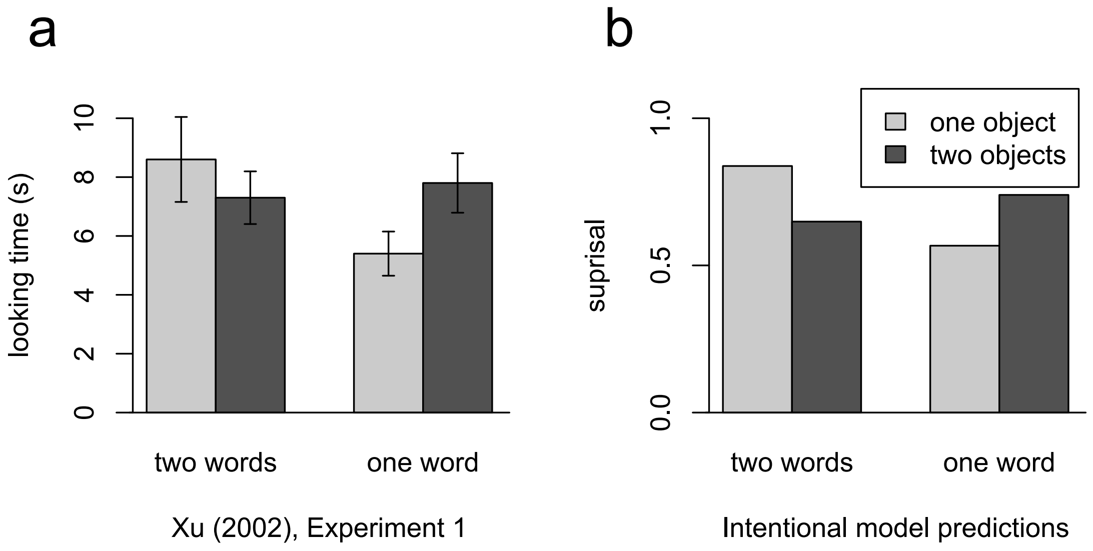

Figure 4-3: Infant data from Xu (2002)'s Experiment 1 on the use of labels to individuate objects along with model surprisal (negative log probability) in the four conditions of Xu's experiment. Our model predicts the results of infants in these studies, mirroring the interaction in looking times between the number of objects seen and number of labels heard.

4.4.4 Object individuation

Even before their first birthday, infants are able to use the presence of words to help individuate objects (Booth & Waxman, 2002; Waxman & Booth, 2003; Waxman & Markow, 1995; Xu, 2002). In one experiment (Xu, 2002), infants saw first a duck and then a ball emerge and then retreat behind a screen. Infants in the two-word condition heard "look, a duck" and then "look a ball" while infants in the one-word condition heard "look, a toy" twice. At test, the screen dropped, revealing either one or two objects. Infants in the one-word conditioned looked longer at two objects (indicating that they expected only one object), while infants in the two-word condition instead looked slightly longer at the single object (indicating that they expected two objects and were surprised that one had disappeared).

Why would hearing two different labels allow infants to make the inference that two different objects were behind the screen? Perhaps infants' assumptions about how words are normally used allows them to infer what state of the world (one object or two) would be most likely to make a speaker utter the labels they heard. Since our model prefers lexicons with more one-to-one mappings and lexicons which interpret the corpus as having more referential words, the best interpretation of Xu's two-word condition in our model is that each word is mapped to a different object and that both words are being used referentially. Under this interpretation there must be two different objects behind the screen, so that the two words can be used referentially to refer to each of them. Likewise, in the one-word condition, the most likely interpretation is that the one word refers to one object and that it is being used referentially for that object—thus, there is likely only one object behind the screen.

To simulate the Xu paradigm formally in our model, we created sets of situations corresponding to the stimuli from the two experimental conditions. For each set, we created two construals: one in which there were two objects (though they were seen one at a time) and one in which there was only one object. To simulate the infant's uncertainty about the meanings of the word or words in the experiment, we evaluated each construal for all possible lexicons. We then compared the surprisal of the model—a quantity that has been shown to map model probability to reaction time data (Levy, 2008)—for the two construals of each experimental condition (e.g., two words, one object vs. two words, two objects). This comparison can be interpreted as measuring, for a learner with no knowledge of what the words mean, how much more surprising it would be to find one object as opposed to two behind the screen. We found a cross-over interaction—higher surprisal when the number of words did not match the number of objects—mirroring the results found by (Xu, 2002) (shown in Figure 4-3). Thus, the model was able to use its assumptions about how words work to make inferences about the states of the world that caused a speaker to produce those words.

4.4.5 Intention-reading

(Baldwin, 1993) conducted an experiment in which 19-month-old toddlers were shown two opaque containers, each containing a different novel toy. The experimenter opened one container, named the toy inside without showing the child the contents of the container, gave the child the toy from the second container to play with, and then finally gave the child the first (labelled) object. Despite the greater temporal contiguity between the label and the second toy, the children showed evidence of learning that the label corresponded to the first toy. Baldwin interpreted these results as evidence that the children used the experimenter's referential intention as their preferred guide to the meaning of the novel label. Our model, built around inferring the speaker's intended referents, can capture this interpretation directly. To illustrate, we constructed a situation with two novel objects and a single novel word. While we previously treated the speaker's intention as a hidden variable, to model Baldwin's task we now gave the model additional information that the speaker intended to refer to the first novel object. The model then highly prefers the correct pairing.

This result should not be surprising, as we have directly incorporated the referential intention of the speaker into our simulation. But a model which does not incorporate a representation of referential intent will be unable to predict Baldwin's results. Models which rely directly on perceptual salience do not capture this result since the object which must be more salient for the correct mapping to occur is actually out of sight when it is being labeled.

4.5 General Discussion

We have presented a Bayesian model which unifies cross-situational statistical approaches and intentional approaches to word learning. The model performs well in learning words from a natural corpus and also predicts a variety of behavioral phenomena from the word-learning literature. Previous evaluations of word-learning models have focused on either their behavioral coverage (Regier, 2005) or their objective corpus performance (Yu & Smith, 2007), but to our knowledge this represents the first systematic attempt to evaluate models on both criteria.

Our model operates at the "computational theory" level of explanation (Marr, 1982). It describes explicitly the structure of a learner's assumptions in terms of relationships between observed and unobserved variables. Thus, in defining our model, we make no claims about the nature of the mechanisms that might instantiate these relationships in the human brain. This kind of ideal-observer analysis is only one part of a full account of early word learning, and many other computational models can provide insights into different aspects of this process (Colunga & Smith, 2005; K. Gold & Scassellati, 2007; Li et al., 2007; Regier, 2005).

The success of our model supports the hypothesis that specialized principles may not be necessary to explain many of the smart inferences that young children are able to make in learning words. Instead, in some cases, a representation of speakers' intentions may suffice. Our model is only a first step in this direction, but we hope that our work here will inspire future modelers to use intentional inference to unite the rich variety of information available to young word learners.

Chapter 5

Conclusions

This thesis proposed a general perspective on the role of statistical inference in language acquisition—that many aspects of language acquisition can be described as statistical inference within a generative model structured around the communicative use of language—and then attempted to formulate and test some aspects of this perspective in the domain of word learning. Chapter 2 presented a corpus study of child-directed speech suggesting that social signals like eye-gaze and pointing do not always cue reference correctly. These data lent support to a view in which learners aggregate information about word meanings over time by guessing speakers' intentions rather than by computing associative statistics. Chapter 3 described a model of learning words by assuming that speakers were following the Gricean maxim of informativeness and showed that this model predicted both adults' and children's judgments about novel word meanings. Chapter 4 described a more complex computational model that jointly learned words and made guesses about referential intentions. Taken together, these results provide evidence in favor of a view of early word learning as a process of "communicative inference": using facts about the way language is used in context to help infer the meanings of words.

Nevertheless, the studies described here are only the first steps in the work necessary to formulate such a view and provide strong evidence that it captures a meaningful amount of the phenomena in children's early word learning. A crucial challenge comes in demonstrating that this computational-level theory (Marr, 1982) can nevertheless describe not only what children *should* learn from the input (sufficiency) but also what they *do* learn (fidelity). Evidence for this claim will come from the application of our models and those that follow them to modeling a wide range of empirical results. We have already begun this work through a series of studies on the phenomenon of mutual exclusivity (Ichinco et al., 2009), but a proof of concept that our model is capable of the same performance as human learners should also be accompanied by evidence that our model fits the failures as well as the successes of children in these and other paradigms.

In addition, the model described in Chapter 4 is still a fragment of the full framework that we sketched in Chapter 1, even for learning words for basic-level object categories. Extending this model to take into account social cues and discourse continuity, as suggested in Chapter 2, is an important outstanding technical challenge. Once the scope of the model is extended beyond the basic-level, there are many other important challenges, including adding a more expressive space of possible conceptual mappings and using the machinery described in (Xu & Tenenbaum, 2007b) and Chapter 3 to narrow this space. Another set of possible extensions come from the composition of our word learning model with models of other aspects language learning, such as word segmentation (B. Jones et al., 2010) and syntactic category learning. The probabilistic framework within which our models are formulated provides an important benefit in that they can be composed with a wide variety of other models. Our hope is that future work by ourselves and others can formulate models of word learning that go beyond the level of learning names for objects and address some of the fundamental questions that remain in understanding children's ability to acquire a lexicon.

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.

Anderson, J. R., & Schooler, L. J. (1990). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396–408.

Arunachalam, S., & Waxman, S. (in press). Meaning from syntax: Evidence from 2-year-olds. *Cognition.*

Aslin, R. (2009). How infants view natural scenes gathered from a head-mounted camera. *Optometry & Vision Science*, *86*, 561.

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.

Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using r.* Cambridge, UK: Cambridge University Press.

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.

Baron-Cohen, S., Baldwin, D., & Crowson, M. (1997). Do children with autism use the speaker's direction of gaze strategy to crack the code of language? *Child Development*, 48–57.

Bartlett, E. (1978). The acquisition of the meaning of color terms: A study of lexical development. *Recent advances in the psychology of language: Language development and mother–child interaction*, 89–108.

Benz, A., Jäger, G., & Van Rooij, R. (2005). *Game theory and pragmatics.* Palgrave Macmillan.

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.

Booth, A., & Waxman, S. (2002). Object names and object functions serve as cues to categories for infants. *Developmental Psychology*, *38*(6), 948–957.

Borensztajn, G., Zuidema, W., & Bod, R. (2008). Children's grammars grow more abstract with ageevidence from an automatic procedure for identifying the productive units of language. *Proc. Cog-Sci 2008*, 47–51.

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

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

Brown, P., Pietra, V., Pietra, S., & Mercer, R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, *19*(2), 263–311.

Bruner, J. (1975). From communication to language: a psychological perspective. *Cognition*, *3*(3), 255–287.

Carey, S. (1978). The child as word learner. *Linguistic theory and psychological reality*, *264293*.

Carlin, B., & Louis, T. (1997). Bayes and empirical bayes methods for data analysis. *Statistics and Computing*, *7*(2), 153–154.

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the society for research in child development*, *63*(4).

Carpenter, P., Miyake, A., & Just, M. (1995). Language comprehension: Sentence and discourse processing. *Annual Review of Psychology*, *46*(1), 91–120.

Carroll, G., & Charniak, E. (1992). *Two experiments on learning probabilistic dependency grammars from corpora.*

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

Chierchia, G., Crain, S., Guasti, M., Gualmini, A., & Meroni, L. (2001). *The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures.*

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.

Clark, E. (2003). *First language acquisition.* Cambridge, UK: Cambridge University Press.

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.*

Conti, D., & Camras, L. (1984). Children's understanding of conversational principles* 1. *Journal of Experimental Child Psychology*, *38*(3), 456–463.

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.

Cover, T., & Thomas, J. (2006). *Elements of information theory.* New York: Wiley-Interscience.

Dale, R., & Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science: A Multidisciplinary Journal*, *19*(2), 233–263.

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

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., & 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.

Eskritt, M., Whalen, J., & Lee, K. (2008). Preschoolers can recognize violations of the gricean maxims. *British Journal of Developmental Psychology*, *26*(3), 435–443.

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

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

Fernald, A., & Morikawa, H. (1993). Common themes and cultural variations in japanese and american mothers' speech to infants. *Child Development*, *64*, 637–56.

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.

Frank, M. C., & Gibson, E. (under review). Overcoming memory limitations in rule learning.

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

Frank, M. C., Goodman, N. D., Lai, P., & Tenenbaum, J. B. (2009). Informative communication in word production and word learning. In *Proceedings of the annual meeting of the cognitive science society.*

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., Ichinco, D., & Tenenbaum, J. B. (2008). *Principles of generalization for learning sequential structure in language.*

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. B. (under review). Three ideal observer models of rule learning in simple languages.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models.* New York: Cambridge University 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.

Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules. *Psychological Science*, *17*(8), 684.

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.

Glucksberg, S., Krauss, R., & Weisberg, R. (1966). Referential communication in nursery school children: Method and some preliminary findings. *Journal of Experimental Child Psychology*, *3*(4), 333–342.

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

Gold, K., & Scassellati, B. (2007). *A robot that uses existing vocabulary to infer non-visual word meanings from observation.*

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

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.

Golinkoff, R., Hirsh-Pasek, K., Bailey, L., & Wenger, N. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology*, *28*(1), 99–108.

Golinkoff, R., Mervis, C., & Hirsh-Pasek, K. (1994). Early object labels: The case for a developmental lexical principles framework. *Growing points in child language*, *21*, 125–155.

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.

Graesser, A., Millis, K., & Zwaan, R. (1997). Discourse comprehension. *Annual Reviews in Psychology*, *48*, 163–189.

Grice, H. (1975). Logic and conversation. *Syntax and Semantics*, *3*, 41–58.

Griffin, Z., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, 274–279.

Gualmini, A., Crain, S., Meroni, L., Chierchia, G., & Guasti, M. (2001). *At the semantics/pragmatics interface in child language.*

Guasti, M., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes*, *20*(5), 667.

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

Hart, B., & Risley, T. (1995). *Meaningful differences in the everyday experience of young american children.* Baltimore, MD: Brookes Publishing Company.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: data mining, inference, and prediction.* New York, NY: Springer.

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.

Hoff-Ginsberg, E. (1986). Function and structure in maternal speech: Their relation to the child's development of syntax. *Developmental Psychology*, *22*(2), 155–163.

Hoff-Ginsberg, E. (1990). Maternal speech and the child's development of syntax: A further look. *Journal of Child Language*, *17*(01), 85–99.

Hollich, G., Hirsh-Pasek, K., & Golinkoff, R. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, *65*(3).

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

Horst, J., & Samuelson, L. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy*, *13*(2), 128–157.

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

Hurtado, N., Marchman, V., & Fernald, A. (2008). Does input influence uptake? links between maternal talk, processing speed and vocabulary size in spanish-learning children. *Developmental Science*, *11*(6), F31–F39.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*(2), 236–248.

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., 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., 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.

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.

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.

Kuntay, A., & Slobin, D. I. (1996). Listening to a turkish mother: some puzzles for acquisition. In *Social interaction, social context, and language* (p. 265). Listening to a Turkish mother: some puzzles for acquisition.

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

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

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

Li, P., & Farkas, I. (2002). A self-organizing connectionist model of bilingual processing. *Bilingual sentence processing*, 59–85.

Li, P., Zhao, X., & Whinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive Science: A Multidisciplinary Journal*, *31*(4), 581–612.

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.

Locke, J. (1847). *An essay concerning human understanding.* Troutman & Hayes.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology.* New York: Wiley.

MacWhinney, B. (1989). Competition and lexical categorization. *Linguistic categorization*, 195–242.

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. (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.

Marinari, E., & Parisi, G. (1992). Simulated tempering: a new monte carlo scheme. *EPL (Europhysics Letters)*, *19*, 451.

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

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

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

Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*, *385*, 813–815.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York, NY: Henry Holt and Co.

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.

Merin, N., Young, G., Ozonoff, S., & Rogers, S. (2007). Visual fixation patterns during reciprocal social interaction distinguish a subgroup of 6-month-old infants at-risk for autism from comparison infants. *Journal of Autism and Developmental Disorders*, *37*(1), 108–121.

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.

Nadig, A., & Sedivy, J. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, *13*(4), 329.

Navarro, D. J., & Perfors, A. F. (2009). Similarity, bayesian inference and the central limit theore. *Acta Psychologica*.

Noveck, I. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188.

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

Ochs, E. (1988). *Culture and language development: Language acquisition and language socialization in a Samoan village*. Cambridge, UK: Cambridge University Press.

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.

Onnis, L., Waterfall, H., & Edelman, S. (2008). Learn locally, act globally: Learning language from variation set cues. *Cognition*, *109*(3), 423–430.

Oppenheimer, D., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872.

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.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: Experiments at the semantics-pragmatics interface. *Cognition*, *86*(3), 253–282.

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

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

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. (2003). The self-organizing consciousness as an alternative model of the mind. *Behavioral and Brain Sciences*, *25*(03), 360–380.

Piantadosi, S., Goodman, N., Ellis, B., & Tenenbaum, J. (2008). A Bayesian model of the acquisition of compositional semantics. *Proceedings of the 30th Annual Conference of the Cognitive Science Society.*

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

Plunkett, K., Sinha, C., Møller, M., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net. *Connection Science*, *4*(3), 293–312.

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.

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and cognitive processes*, *8*(1), 1–56.

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

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.

Reeder, P., Newport, E., & Aslin, R. (2009). *The role of distributional information in linguistic category formation.*

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

Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, *3*(01), 57–87.

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

Rosenberg, S., & Cohen, B. (1964). Speakers' and listeners' processes in a word-communication task. *Science*, *145*(3637), 1201.

Rosenberg, S., & Cohen, B. (1966). Referential processes of speakers and listeners. *Psychological Review*, *73*(3), 208–231.

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

Rumelhart, D., & McClelland, J. (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., & group the PDP research. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT 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.

Shafto, P., & Goodman, N. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. *Proceedings of the 30th Annual Conference of the Cognitive Science Society.*

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

Smith, L., Yu, C., & Pereira, A. (in press). Not your mother's view: The dynamics of toddler visual experience. *Developmental Science*.

Snow, C. (1972). Mothers' speech to children learning language. *Child development*, *43*(2), 549–565.

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.

Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*, *9999*(9999).

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition.* Oxford, UK: Blackwell Publishers.

St. Augustine. (397/1963). *The Confessions of St. Augustine.* New York, NY: Clarendon Press.

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

Stern, D. N. (2002). *The first relationship: Infant and mother.* Cambridge, MA: Harvard University PressPr.

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.

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.

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.

Waterfall, H. (2006). *A little change is a good thing: Feature theory, language ac-*

quisition and variation sets. Unpublished doctoral dissertation, University of Chicago.

Waxman, S., & Booth, A. (2003). The origins and evolution of links between word learning and conceptual organization: New evidence from 11-month-olds. *Developmental Science*, *6*(2), 128–135.

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

Waxman, S., & Markow, D. (1995). Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302.

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.

Wittgenstein, L. (1953). *Philosophical Investigations.* Oxford, UK: Blackwell Publishers.

Wolf, F., & Gibson, E. (2006). *Coherence in natural language: data structures and applications.* Cambridge, MA: MIT Press.

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

Woodward, A., Markman, E., & Fitzsimmons, C. (1994). Rapid word learning in 13-and 18-month-olds. *Developmental Psychology*, *30*(4), 553–566.

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

Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition*, *112*(1), 97–104.

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

Xu, F., & Tenenbaum, J. (2007a). Sensitivity to sampling in bayesian word learning. *Developmental Science*, *10*(3), 288.

Xu, F., & Tenenbaum, J. (2007b). 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.

Yoshida, H., & Smith, L. (2008). What's in view for toddlers? using a head camera to study visual experience. *Infancy*, *13*, 229–248.

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., & 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.*