Learning words through probabilistic inferences about speakers' communicative intentions*

Michael C. Frank
Department of Psychology, Stanford University

*Thanks to the editors of this volume for the opportunity to contribute, to Molly Lewis and Dan Yurovsky for helpful comments, and also to Noah Goodman, my collaborator in much of the work described here.

1 Introduction

The linguistic world of young children is likely an overwhelming place. Even if they are not assaulted with James' (1890) "blooming buzzing confusion," it must be perplexing for infants to be surrounded constantly by sounds whose only interpretable meaning at first comes from the tone of voice in which they are uttered. Perhaps another metaphor is more apt: the infant is a traveler trying to negotiate a task in a foreign language, with the help of a sympathetic interlocutor (a parent or caregiver). The child and caregiver may share goals, or at least an understanding of the other's goals. But without knowing any words, it is only the rare utterance that can be decoded from context; and the fewer words that are known, the less leverage the child has to infer the meanings of others. Even in the highly supportive contexts created by parents, the vast majority of language is likely to be incomprehensible to young infants.

How then do children begin to break into the vocabulary of their first language? Two broad proposals run throughout work on word learning from historical sources to contemporary models: *associative* and *intentional* proposals. In associative accounts from Locke (1690/1964) onward, infants are hypothesized to match elements of their linguistic environment with the world around them, identifying the consistent mappings between words and other stimuli. In intentional accounts from St. Augustine (397/1963) onwards, in contrast, the attention of learners is on the speakers who produce words. These words are then mapped to the speakers' intended meanings—often instantiated by the speakers' intended referent in the current context.[1] These distinct accounts have lead to the discovery of distinct phenomena and empirical paradigms, including "cross-situational word learning" (supportive of an associative account Yu & Ballard, 2007), to "fast-mapping" (supportive of an intentional account Carey, 1978). But the two views need not be in conflict.

I will argue here that these two accounts are compatible with one another and can be integrated in a single framework. This framework is fundamentally intentional in that it is oriented around representations of speakers' goals and intentions, but it also takes advantage of the graded, probabilistic nature of learning to aggregate information across multiple learning situations. The intentional aspects of the framework are congruent with research on children's social cognition (e.g. Onishi & Baillargeon, 2005; Csibra & Gergely, 2009; Vouloumanos, Onishi, & Pogue, 2012) and the framework supports graded probabilistic learning via recent developments in cognitive modeling (Chater & Oaksford, 2008; Tenenbaum, Kemp, Griffiths, & Goodman, 2011).

My goal in this chapter is to provide an overview of this framework for probabilistic, intentional models of word learning.[2] I refer to this set of ideas as a framework, rather than a model: To my mind a framework refers to a set of principles, while a model refers

[1] In this chapter, I will be talking about the ways that children learn the meanings of words. The problem of inferring meanings can be broken into two separate tasks: the task of mapping a word or string of words to a *referent* in the moment, and the task of identifying the *meaning* or *concept* corresponding to that referent. The distinction between these two problems is clear, but our terms for discussing them are often clumsy. The issue is often confused further by the use of simple noun learning as a case study, since solving the word-object mapping problem in that case comes close to giving away the concept-learning problem (modulo a bias for basic-level categories; Markman, 1991). I'll be focusing initially on word-object mapping but I will note when the equivalence between mapping and concept learning is broken (Section 4).

[2] For a more general treatment of social contributions to learning, see Shafto, Goodman, and Frank (2012)

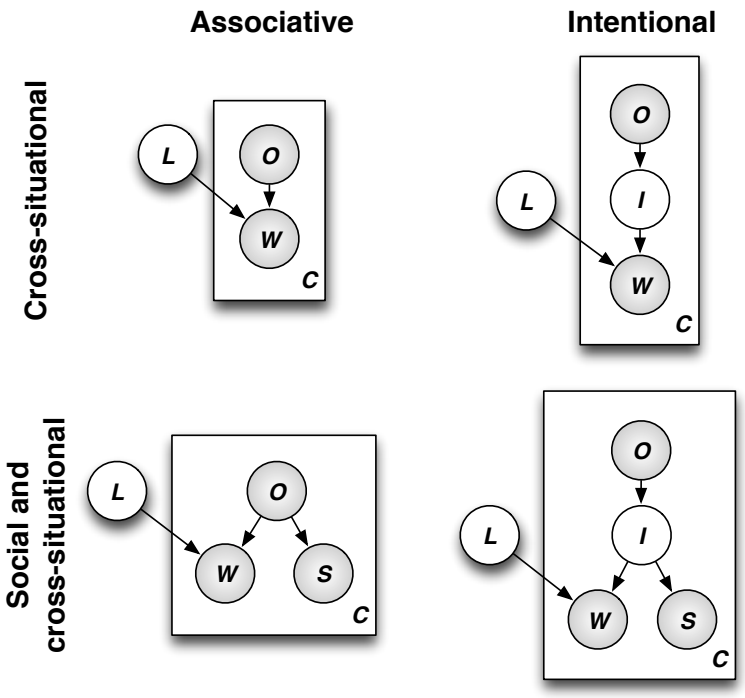

Figure 1: A progression of possible models of the basic challenge faced by early word learners. L is the child's lexicon, C refers to the contexts in which utterances containing words W are observed, accompanied by object referents O and social cues S, as well as (unobserved) communicative intentions I.

to a particular instantiation of those principles within a working, implemented system from which concrete, quantitative predictions can be made. Although we have made a number of such models, all are specific to particular applications or kinds of data. Nevertheless, I believe that there are some more general conclusions that can be drawn from these individual systems when they are examined together.

I will try to lay out some of the theoretical distinctions in more depth than is allowed in a strictly empirical report. In Section 2, I'll describe a taxonomy of computational models of word learning. Section 3 will specify formal details corresponding to this taxonomy. The goal of this section will be to distinguish between those cross-situational models that merely use social information and those that make an intentional assumption about the nature of the learning situation. Finally, Section 4 will focus on the links between our intentional model of word learning and our work on modeling pragmatic inference.

2 A taxonomy of models of word learning

The strategy of learning words via co-occurrence has been labeled "cross-situational" word learning. Recent empirical work provides strong support for the idea that both adults and infants can learn word-object mappings by gathering consistent associations across multiple, ambiguous exposures (Yu & Ballard, 2007; Vouloumanos, 2008; Smith & Yu, 2008;

Vouloumanos & Werker, 2009). A variety of theorists have asked about the utility of cross-situational learning in acquiring words of different types (Pinker, 1984; Gleitman, 1990; Fisher, Hall, Rakowitz, & Gleitman, 1994; Gillette, Gleitman, Gleitman, & Lederer, 1999; Akhtar & Montague, 1999). Though theoretically all are learnable (Siskind, 1996), there is likely to be a continuum from those words most learnable by co-occurrence, e.g. nouns and some property terms, to those least likely to be inferred from context alone. This latter category likely includes verbs—which refer to events that have multiple construals (e.g., "chase"/"flee")—context-dependent adjectives, and function words. Because of the relative simplicity of learning basic-level object nouns from context, we begin by discussing word-object mapping models.

A taxonomy of models of word-object mapping is shown in Figure 1. All of these models are "cross-situational" in the sense that all consider evidence about the relationship between words and parts of the world across multiple observations. All of them also attempt to learn a *lexicon*: a set of consistent word-object mappings. They differ, however, in both the information sources they consider and the ways they use these information sources. Each of these models describes a set of variables, both observed (shaded) and unobserved (unshaded) as well as a set of causal dependencies between these variables. These dependencies define a *generative process*: a set of steps by which the learner assumes the observed data have been generated. Unobserved aspects of this generative process can then be estimated using inference techniques.[3]

In our taxonomy, we differentiate models on two dimensions. The first—represented by the vertical axis in Figure 1—is the information sources considered by the model. "Pure" cross-situational models consider only the co-occurrence between words and objects in establishing links between words and objects in the lexicon. Social and cross-situational models consider also a set of social cues (envisioned here as signals like a point, a gesture, or a gaze towards an object, with some temporal connection to a particular utterance).

The second dimension is the way models represent the relationship between words and the world—shown on the horizontal axis in Figure 1. Associative models describe a generative process in which words are assumed to be generated directly by the presence of their associated objects (without the presence of an intervening speaker). In contrast, intentional models, as we define them, assume that words are generated via the *communicative intention* of a speaker to refer to an object. Thus, in the sense used by theorists of social cognition, intentional models are fundamentally *triadic*: they define a relationship between the child, the speaker, and objects in the context (Baldwin, 1995; Carpenter, Nagell, & Tomasello, 1998).

The taxonomy described here is primarily concerned with the assumptions that learners make, rather than the mechanisms by which they learn under these assumptions. In the current analysis, models are *ideal observers*: realizations of the assumptions that learners make and information sources they use, rather than the mechanisms by which these assumptions and information sources are processed (Geisler, 2003). Another way of putting this is that our analyses are at the *computational*, rather than *algorithmic* level, describing models as they represent the task faced by the child rather than as the child solves them (Marr, 1982).

[3]For a very nice tutorial introduction to this general style of modeling, as applied to developmental questions, see Perfors, Tenenbaum, Griffiths, and Xu (2011).

There is a rich literature describing how such constraints should be implemented in word learning models (Fazly, Alishahi, & Stevenson, 2010; Yu & Smith, 2012), and my own work in other domains has investigated these question as well (Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Frank & Gibson, 2011). In addition, there is a fascinating and growing literature investigating the ways that resource-bounded decision-making can relate to normative models, see e.g. Sanborn, Griffiths, and Navarro (2010) for review and discussion. Yet I worry that implementational considerations often mask (rather than reveal) the learner's underlying assumptions about the learning situation, and I believe that there is value in considering the computational level independent from—and perhaps in parallel to—the algorithmic.[4] Therefore, in this chapter I will not discuss an important body of modeling work that investigates the consequences of algorithmic details of representation for modeling human performance in word learning (e.g. Li, Farkas, & MacWhinney, 2004; Regier, 2005; Yu & Smith, 2012).

In the next section, I describe computational details underlying the models in Figure 1. Providing these details allows for clarity about how an intentional assumption can be implemented, and additionally allows for comparison and integration with the pragmatic models described in Section 4.

3 A formal framework for cross-situational learning

Formalization allows us to consider what has previously been a somewhat slippery distinction: between social cross-situational models, which consider information generated by other people, and intentional models, which are based on a stronger assumption about the generating source of this information. Following the taxonomy in Shafto et al. (2012), I take the fundamental assumption underlying a communicative model to be that language is produced as a rational action to accomplish a goal. The "language as rational action" assumption allows for stronger inferences from data than those possible under a view that considers social data but does not consider the communicative intentions that lead to those data being generated. This section walks through the formal mechanics of these ideas: Section 3.1 describes the basics of cross-situational learning under the general family of models, Section 3.2 lays out the key differences between associative and intentional models, and Section 3.3 describes

[4]Related to this issue is a set of recent criticisms of the idea that cross-situational observation is a factor in learning word-object mappings (Medina, Snedeker, Trueswell, & Gleitman, 2011; Trueswell, Medina, Hafri, & Gleitman, in press). The key empirical question underlying these criticisms is whether individual word learners represent multiple hypotheses about the meanings of words. The proposed alternative is that individuals make noisy, stochastic choices of individual hypotheses that they then test against data (a "propose but verify" strategy). The average of many such stochastic choices across items and individuals would then produce the pattern of gradual learning that is sensitive to the degree of cross-situational ambiguity that was observed in previous studies. This criticism is rooted in a long-standing debate about the nature of learning more generally, and whether it is typically gradual and associative in character or discrete, and hypothesis-based (Gallistel, 1990; Gallistel, Fairhurst, & Balsam, 2004). But such concerns relate primarily to the mechanisms of learning applied by the learner in recovering the lexicon of their language, rather than to the underlying assumptions that guide this learning. Even if "propose but verify" learners do not retain previous data to test their hypotheses, they still must make the assumption that their hypotheses should in principle be consistent with previous data as well as future observations. In other words, even if learners are memoryless, their behavior is still consistent with an assumption of cross-situational statistical consistency.

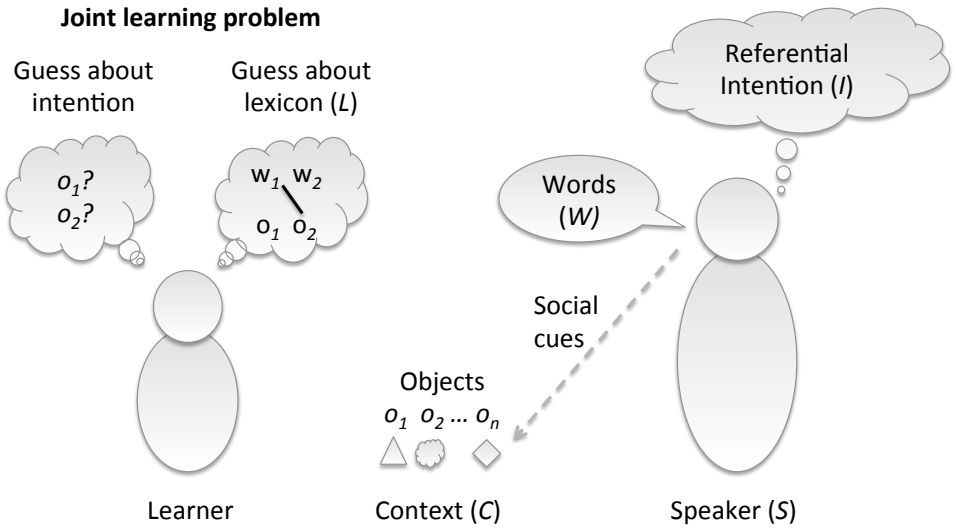

Figure 2: A schematic view of the intentional framework, linking learner and speaker via the two learning problems: guessing the speaker's intended referent and guessing the lexicon of the language.

how social information is used in associative and intentional models.

3.1 Basic cross-situational learning

Consider a schematic description of the child's learning problem, shown in Figure 2. The child finds herself in a set of contexts C. Each of these contexts contains possible referents $O_c = o_1...o_n$. In each of these contexts, a speaker has a communicative intention I_c, unobserved to the child. On the basis of this communicative intention, utters words $W_c = w_1...w_n$ and produces social gestures S_c. Having observed a set of these kinds of contexts, the learner's goal is to infer a set of correspondences between words and objects, which we denote L (the lexicon). This lexicon is assumed to be stable across contexts and individuals,[5] and can be modeled either as a set of discrete links or continuous, probabilistic associations.

The challenge for the child is that each individual context does not uniquely determine a set of lexical mappings. Which words go with which objects? A number of learning models that have been proposed in the broader literature can be considered as solutions to this problem, varying the information sources that are used and the nature of the additional

[5]Does the child know that her lexicon should be stable across time and identical across individuals? There are several such foundational assumptions that are necessary for *all* of the models described in this chapter. For example, we assume that words are linked to a particular level of description (objects or object concepts in the models we consider), and not to some others (e.g., motor actions, other sensory stimuli). The basic learning frameworks we describe could in principle be applied to a scenario with many more targets for words, no stability of word meaning across time, or vast individual differences in language use (whether due to bilingualism or even simply random variation), but there is no guarantee that they would be sufficient. In practice, whether these assumptions are inborn or discovered, it is likely that they are necessary for learning to proceed.

assumptions that are used. We can notate this problem of lexicon learning as a problem of Bayesian inference, that is, of inferring the most probable lexicon given the set of observed contexts:

$$P(L|C) \propto P(C|L)P(L). \tag{1}$$

The taxonomy in Figure 1 provides generative processes for four kinds of models that have been applied to this learning problem. The simplest approach to this problem is simply to estimate these probabilities, neglecting intentions or social cues. The model in the upper left corner, which assumes that words are generated via the observed objects and the unobserved lexicon, can be written

$$P(L|C) \propto P(W|O, L)P(L). \tag{2}$$

If we represent the words in a context as the set W_c and the objects as O_c, we can expand this expression to

$$P(L|C) \propto \prod_C P(W_c|O_c, L)P(L). \tag{3}$$

In other words, the probability of the lexicon under these models is the product across contexts of the probability of the words in the context, given the objects in the context and the lexicon. This "pure cross situational" approach is followed by a number of influential models (Yu & Ballard, 2007; Fazly et al., 2010).[6]

3.2 Differentiating associative and intentional models

In this section, we compare the assumptions made by associative and intentional models, differentiating between the upper left and upper right panels of Figure 1. Following the approach above, the next step in defining a word learning model is to define the likelihood of a word being uttered, given the presence of some object and its lexical entry (the term $P(W_c|O_c, L)$ above). There are two important sub-problems that arise in defining this term: first, defining the alignment between particular words and objects (assuming that there are multiples of each in each situation), and second, defining the probability of a word being used with a particular object (given that they are aligned). I'll discuss only the first of these here. [7]

[6]This modeling work has built directly from work from machine translation that attempted this optimization problem for aligned corpora (e.g. where words and objects are actually words in a target language and words in a source language; Brown, Pietra, Pietra, & Mercer, 1993).

[7]The key intentional assumption is orthogonal to whether a model assumes a discrete *lexicon*. Our initial work used a fairly discrete likelihood function to determine the probability that a particular word was used, given that the speaker had an intention to refer to some object:

$$P(w|I = o_c, L) = \begin{cases} \propto 1 & \text{if } L(w_c, o_c) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $L(w_c, o_c) = 1$ indicated that w_c and o_c are linked in the lexicon. But it would be equally possible to define a more clearly probabilistic function, using e.g. a multinomial distribution (which would then be

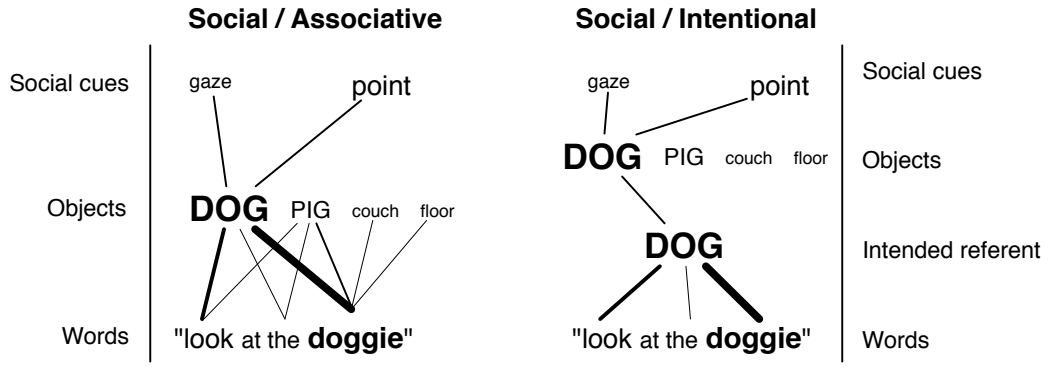

Figure 3: A schematic of a single situation for associative and communicative models that both use social and prosodic information. Gaze and pointing cues signal that a dog toy is more salient than a pig toy. Prosodic focus on the word "doggie" raises its salience. As a consequence, the strongest association is between the object DOG and the word "doggie" for both models. The communicative model includes a filtering step in which DOG is assumed to be the correct referent. Salience is shown by type weight and size, while associative weights are shown by line weight.

It is with respect to the problem of aligning words and objects in the context that the differences between associative and intentional models becomes clear. Associative models typically make minimal assumptions about alignment and assume that any word can be aligned with any object. In some sense, this is the basic tenet of an associative model: that all words and objects present in the context are associated with one another to some degree.

In an intentional model, in contrast, we assume that the relationship between words and objects is mediated by the speaker's intention to refer to some set of objects. This mediation relationship is shown graphically by the intervening node between O and W in the generative process for models on the right side of Figure 1. The concept of a referential intention—an intention by a speaker that mediates the relationship between the physical context and the words produced by that speaker—corresponds to the concept *goal* that plays a central role in work on social learning and rational action inference (Gergely, Bekkering, & Király, 2002; Baker, Saxe, & Tenenbaum, 2009; Shafto et al., 2012). This intuition is shown graphically in Figure 3, which shows the mediating relationship that the speaker's intention can play in learning from a single situation.

Formally, this mediation relationship results in a revision to Equation 5, where we notate this mediating intention I_C:

$$P(L|C) \propto \prod_C P(W_c|O_c, I_c, L)P(L). \tag{5}$$

The addition of this mediating variable affects the process of finding the alignment between

conjugate to a dirichlet prior). More generally, it is likely very difficult to differentiate between a continuous lexical representation and a posterior distribution representing uncertainty over a discrete lexicon. Although our work on this topic is occasionally cited as providing a "hypothesis testing" view of word learning, I see the discreteness of the lexicon in that particular model as an implementation choice rather than one that carries any particular theoretical weight.

words and objects: While associative models assume that all words are linked to all objects, the intentional models assume that there is an extra step that removes some of these associations from consideration.

In our work on cross-situational learning to date, the representation of communicative intention has been quite basic, representing the speaker's menu of possible intentions as the set of objects in the context.[8] In our initial model, we specified I_c as containing a subset of O_c, corresponding to the assumption that the speaker could talk about any subset of the objects in the context, including the empty set (Frank, Goodman, & Tenenbaum, 2009). Although this assumption allowed us to consider a wide variety of possibilities, if there were many objects present in a context it quickly became unwieldy, since it required considering the power set of the context (which grows at 2^n where n is the number of objects in c). Hence, in more recent work we have begun using the simplifying assumption that I_c is a single object in O_C (Johnson, Demuth, & Frank, 2012), congruent with our empirical observation that most of the caregivers' utterances in a corpus referred to at most one object.

Regardless of how the intended referent is chosen, under an intentional assumption, we can define the likelihood of a word as the product of two terms: the probability of the words given the intention, and the probability of a particular intention. Thus,

$$P(W_c|O_c, I_c, L) = \sum_{I_c \in O_c} P(W_c|I_c, L)P(I_c|O_c). \tag{6}$$

Thus, intentional models describes a two-step process of uttering a word: first decide which object to refer to, then decide the words to use to refer to it.

The key difference between associative and intentional models on our account is this two-step process, separating the choice of what to talk about from the choice of how to refer. Both types of models allow for information to "weight" the learner's estimate of which objects are most salient, for example via social information. But in intentional models, this weighting influences the learner's guess about which object(s) are being referred to. In contrast, in associative models, all words are associated with all objects: implicitly, this assumption is tantamount to assuming that all objects are being referred to and all words have referential status (just to greater or lesser degrees). In contrast, the definitional assumption of intentional models is that speakers have a discrete referential intention for each utterance, even if the learner is uncertain of what this intention is.

The intentional assumption—that there is a discrete choice of intended referent by a speaker that mediates between the physical context and the language that the learner hears—implies that one major part of word learning is in-the-moment interpretation. If a learner knows what object is being talked about (the referential intention), there is no need to compute associations between the words that are heard and the other objects that are present. In the language of causal models, the intention "screens off" the physical context from the words: knowing the speaker's intended referent is enough for learning. In this respect, the intentional framework model is deeply related to recent work by McMurray, Horst, and

[8]Nevertheless, I believe that this construct has the potential to be far more flexible and powerful than the use to which we have so far put it. Provided that this distribution over potential intended referents—or even intended meanings—is limited by context, discourse, and other pragmatic factors, a much wider variety of possible interpretations could be considered.

Samuelson (2012), who emphasized the role of learners' interpretations of reference in the moment in longer-term word learning.

In Frank, Goodman, and Tenenbaum (2009), we reported results based on running associative and intentional models on a small, hand-annotated corpus of infant-directed speech. The intentional model outperformed other models in the lexicon it learned. This success was due at least in part to its ability to "filter out" spurious associations between function words and objects. Under the intentional model, evidence that a word was mapped to an object in one situation could help constrain hypotheses about which object was being referred to in another situation. This mutual constraint meant that irrelevant co-occurrences could be explained away, rather than providing noise (as in the associative models).

3.3 Adding social information

It is relatively rare for a speaker to talk about a physically present referent without giving some signal—at least at some point in the conversation—that the referent is indeed the one being talked about. Speakers gaze to conversational referents during language production both as an explicit social signal (H. Clark, 1996) and as a consequence of processes underlying language production (Griffin & Bock, 2000). In addition, speakers often signal reference by pointing."Social cues" like eye-gaze or points to conversational referents are often cited as an important source of evidence for learning word-object mappings (St. Augustine, 397/1963; Baldwin, 1993; Hollich, Hirsh-Pasek, & Golinkoff, 2000; Bloom, 2002).

Social cues can be represented in an associative model as cues to which objects are most salient in a particular situation (Yu & Ballard, 2007). In other words, the presence of a point or gaze on a particular object endows it with some additional salience, which should in turn strengthen its associations with words. The left side of Figure 3 shows a caricature of what the data for a single situation might look under such a model. The associative weights between words and objects are determined by the social cues and perceptual salience of the objects in the scene (as well as the prosodic salience of the words, which we do not discuss here). The result is that the same associative computation is performed, but over a word/object set whose salience is no longer uniform.

In the generative processes of Figure 1, the lower left-hand panel shows a model in which social cues are a reflection of the underlying salience of individual objects. If social cues are present, object salience is inferred to be higher. In contrast, in the intentional in the lower right, social information in particular informs the process of interpretation (deciding which object, if any, is being referred to). Social cues are generated by the speaker's intention to refer and hence are signals to the underlying intention.

This interpretive use of social cues can be implemented in a number of ways. In our early work on this topic, we assumed that each cue could be the consequence of a relevant intentional action or could be the result of baseline looking without an underlying intention, and estimated these probabilities via a "noisy-or" model (Frank, Goodman, & Tenenbaum, 2007). This formulation allowed the model to learn that, for example, even though speakers' gaze was a frequent cue, its overall reliability was low.

Our more recent work embeds these social cue probabilities within a probabilistic, grammar-based formalism. In Johnson et al. (2012), we used an *adaptor grammar*, a probabilistic con-

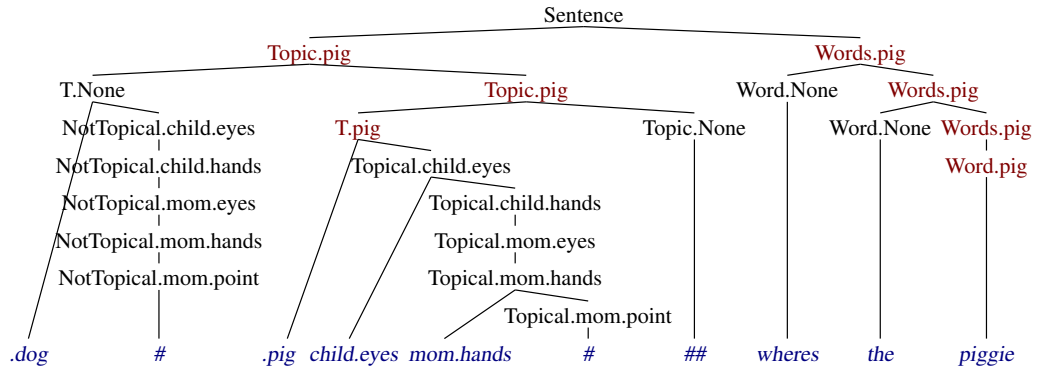

Figure 4: A parse tree for an entire situation, including a sentence along with its referential context and social cues. The sentence generates topical (referential) objects, the social cues that mark these objects, and the words that refer to them. Topic-specific words are marked in red and observed data are in blue. In this case, the referent is "pig" and referential words are propagated throughout the tree. Figure reprinted from Johnson et al. (2012).

text free grammar formalism that allows common structures to be reused efficiently (Johnson, Griffiths, & Goldwater, 2007). Consider the example shown in Figure 4. The observed input representation of the situation (shown in blue) specifies that the words "where's the piggie" are observed along with dog and pig toys, and the pig toy is marked by two social cues, the child's eyes and the mom's hands.[9] The tree above it shows a possible parse of the situation. The extraneous dog toy is parsed as "non-topical" (e.g. not an intended referent, marked as *T.none*) and the pig toy and its accompanying social cues are generated on the basis of the topic. The words are also generated by the same topic, with several "non-topical" words (*Word.none*) followed by a word generated from the topical lexicon (marked as *Words.pig*).

This grammatical formalism, although distinct from our previous work in some of its details, still encodes the same two-stage computation we have associated with intentional models. The first decision in the grammar is what the referent (topic, in the language of this model) of a sentence is. This decision affects all other aspects of the sentence including the probabilities of both the social cues and the individual words, whose choice together constitutes the second decision: how to refer. When we evaluated the grammatical model on a corpus tagged with social information, we found significant gains in the accuracy of both guessing the referents of utterances and the words associated with particular objects on the basis of adding the social cues. In addition, our analyses showed that the child's own gaze on an object was the most predictive cue, suggesting that our corpora encoded significant follow-in labeling (and replicating descriptive results from Frank et al., in press).

To summarize: both associative and intentional models allow for the inclusion of social information, but associative models allow for social information to make particular referents more salient and bias the computation of associations between words and objects. In contrast, intentional models go beyond this interpretation of social cues as signals of salience and allow the social information to bias the computation of reference. Yet all of the models de-

[9]We leave aside here the issue of whether the child's own eyes can be considered a "social cue"—this issue is discussed at length in Frank, Tenenbaum, and Fernald (in press).

scribed here still treat only the mapping problem, implicitly equating referent identification with meaning learning. In this next section, we broaden the set of possible word meanings we consider, beginning the process of differentiating word meaning from reference.

4 Adding pragmatic inference to intentional models

In the Quinian (1960) framing of the word learning problem, even if a single word is heard alongside a single object, there are still an infinite number of possible interpretations for the word. Quine distinguished between interpretations for which co-occurrence or pointing—the information sources relied on in the models above—provide no traction, and those for which they may. This first class includes, for example, the ambiguity between "rabbit" and "undetached rabbit parts"; Quine argues that these meanings may be impossible to distinguish without using language itself to do the distinguishing. On the other hand, the second class contains many word meanings that may be empirically distinguishable but are likely to be confounded in any given context. To take a small set of the many possible modes of reference, a particular rabbit might be talked about as "rabbit," but also as "white," (when the contrast is a brown rabbit),"animal" (when pointing out something in the bushes), or "small" (when the contrast is a larger rabbit).

Since basic-level object names are common in speech to children (Callanan, 1985), it should not be difficult to learn a word like "rabbit" from co-occurrence. The prospects for noticing the co-occurrence between animacy and "animal" seem somewhat lower; they are likely lower still for a color like "white," and close to nil for a gradable adjective like "small" (whose meaning changes from context to context). Information about the context of reference might provide a far more straightforward path to learning such terms (perhaps along with syntactic information, in the case of adjectives; Waxman & Booth, 2001). In this situation, a pragmatic word learner has an important advantage over any of the cross-situational learners described above: She can consider the context of use and the goal of the speaker in uttering a particular phrase, and crucially, she can consider why a term contrasting with the conventional descriptor is used (E. Clark, 1988). This is the intuition that I will follow in this last section.

On the standard cross-situational view, pragmatic contrast inferences belong to an entirely different class from the associative inference that leads the child to consider that "rabbit" = RABBIT. The intentional view provides a way to integrate these two inferences, however. Because long-term learning is mediated by in-the-moment interpretation, learners can use pragmatic computations to inform their guesses about what words refer and what objects (and even aspects of these objects) are being referred to. In this section I will show an example of our recent work modeling pragmatic inference in context, and then demonstrate that this kind of model of pragmatic inference can be integrated with the intentional word learning model above. There is much work yet to do to realize the promise of learning the meanings of, for example, context-dependent adjectives. But I believe this framework provides the beginnings of the tools necessary to extend the cross-situational paradigm beyond word-object mapping to a broader class of meanings.

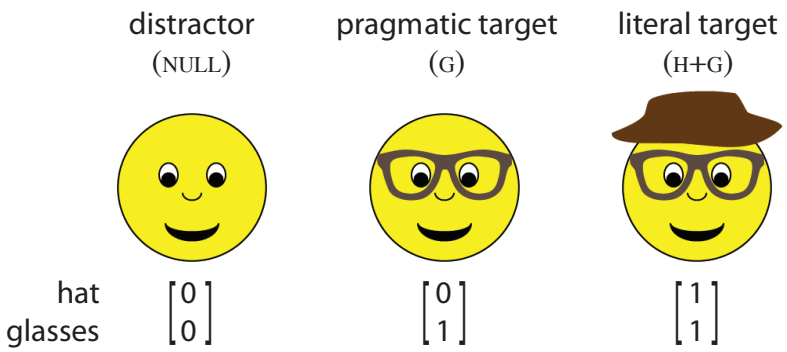

Figure 5: An example stimulus from our pragmatic inference experiments. Participants would be asked to identify the referent of a phrase containing "glasses" as the discriptor; given this message, the middle face (who has glasses but no hat) is the pragmatic implicature target, whereas the right-hand face (who has a hat and hence a better potential descriptor) is the literal target.

4.1 Modeling pragmatic disambiguation of reference

Grice (1975) proposed a set of maxims for normative communication: Speakers should be truthful, relevant, clear, and informative. On Grice's account, listeners should interpret utterances as though these maxims are being followed, allowing them to go beyond the truth-functional meanings of the words in the sentences to derive richer meanings in context that he called "implicatures." In our modeling work we have explored the idea that some of these implicatures can be captured in formal models. We focus here on the maxim of informativeness and model listeners as doing statistical inferences about what a speaker's intended referent is, given a presumption of informativeness.

Our model assumes that in a context C, the listener is attempting to infer the speaker's intended referent I from a set of possible referents.[10] The listener considers two factors: first, the relative informativeness of the speaker's utterance with respect to each of the referents, and second, the "contextual salience" of the referents. Contextual salience here refers to the relative likelihood, given the conversational context, the shared knowledge between communicators, and other factors that jointly determine that a particular referent will be the object of the speaker's expression. Although this quantity might in principle be derived from *a priori* considerations, in our current work we have measured it empirically.

This two-part model can be notated as follows:

$$P(I|w, C) \propto P(w|I, C)P(I) \tag{7}$$

with the two terms on the right corresponding to the two factors being considered. In this work, we investigate the idea that the informativeness of a word in context is inversely proportional to its specificity, so that

[10]Note that in Frank et al. (in press) we use r_S rather than I–we use I here for consistency with Section 3.

$$P(w|I,C) \propto \frac{1}{|w|} \tag{8}$$

where $|w|$ notates the number of objects in the context that can be referred to using w. (This inverse proportionality corresponds to the "size principle" of Tenenbaum & Griffiths, 2001).[11]

An example is useful in clarifying how this model setup naturally leads to Gricean pragmatic implicatures. Consider the situation pictured in Figure 5 (originally from Stiller, Goodman, & Frank, 2011). There are three possible referents shown, referred to below as NULL, G, and H+G. Each has different features, which lead to different possible expressions that can be used to refer to it. For the purpose of this game, we assume that this set is limited to the expressions "hat" and "glasses." A speaker utters the word "glasses"; the job of the listener is to decide which face is being referred to. The intuition, confirmed experimentally Stiller et al. (2011), is that it is G, the face without a hat.

This simple pragmatic inference has many of the elements of scalar implicature (the often-studied inference that "some" typically is strengthened pragmatically to mean "some but not all"), so it is a useful case study of how our model can be applied. In order to simplify the computation, we assume that the contextual salience of the three faces is even, and that the NULL referent is not considered. We can compute the strength of the pragmatic inference that "glasses" refers to G using Equation 7 and expanding the proportionality by normalizing over all possible referents:

$$\begin{aligned} P(\text{G}|\text{“glasses”}, C) &= \frac{P(\text{“glasses”}|\text{G}, C)}{\sum_{r' \in C} P(\text{”glasses”}|r', C)} \\ &= \frac{P(\text{“glasses”}|\text{G}, C)}{P(\text{“glasses”}|\text{G}, C) + P(\text{“glasses”}|\text{H+G}, C)} \end{aligned}$$

Now, we can expand, using Equation 8 and notating the set of vocabulary items that can be used to describe e.g. item G as $w \in \text{G}$:

[11] That this set of definitions implicitly assumes that words are no longer names of objects: instead, they are functions that can be applied to a context and that return true or false for each object in the context. This truth-functional model of word meaning can easily be used to capture predicates like "white" or "furry" and is in principle extensible to context-dependent adjectives like "small" (Schmidt, Goodman, Barner, & Tenenbaum, 2009).

$$
\begin{aligned}
P(\text{G}|\text{“glasses”}, C) &= \frac{\frac{\frac{1}{|\text{“glasses”}|}}{\sum_{w' \in \text{G}} \frac{1}{|w'|}}}{\frac{\frac{1}{|\text{“glasses”}|}}{\sum_{w' \in \text{G}} \frac{1}{|w'|}} + \frac{\frac{1}{|\text{“glasses”}|}}{\sum_{w' \in \text{H+G}} \frac{1}{|w'|}}} \\
&= \frac{\frac{\frac{1}{|\text{“glasses”}|}}{\frac{1}{|\text{“glasses”}|}}}{\frac{\frac{1}{|\text{“glasses”}|}}{\frac{1}{|\text{“glasses”}|}} + \frac{\frac{1}{|\text{“glasses”}|}}{\frac{1}{|\text{“glasses”}|} + \frac{1}{|\text{“hat”}|}}} \\
&= \frac{\frac{1/2}{1/2}}{\frac{1/2}{1/2} + \frac{1/2}{1/2+1}} \\
&= \frac{1}{1 + 1/3} = .75.
\end{aligned}
$$

In other words, the probability of G, the face with glasses and no hat, given the discriptor “glasses,” is predicted to be .75 (and hence the probability of H+G is .25). This computation encodes the intuition that, had the speaker been talking about H+G, he or she would have chosen the more specific descriptor “hat.” This prediction of an implicature in favor of G corresponds well with the judgments of both adults and preschoolers (Stiller et al., 2011).

As illustrated above, our pragmatics model provides a framework for quantifying the Gricean maxim “be informative” from the perspective of both a speaker and a listener. It and its extensions can provide an account for a wide variety of pragmatic phenomena (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2012), especially when combined with empirical measurements of contextual salience. The pragmatic implicature under this model can be computed for any situation in which the set of contextual and vocabulary alternatives is known, although the problem of identifying relevant alternative is still an open research challenge. In the next section, we illustrate how this pragmatic framework can be integrated with the intentional approach described above.

4.2 Using informativeness to learn words

Our pragmatics model is deeply related to the intentional communication model described in Section 3. Recall that in Equation 6 of the cross-situational learning model, we defined the probability of a particular word being uttered, given some referential intention and context as a product of two terms: the probability of the word given the intention, and the probability of the intended referent given the objects in the context. Schematically, this relation is stated $P(W|O, I, L) = P(W|I)P(I|O)$. Intuitively, these two terms govern the probability of choosing a particular referent and choosing the proper referring expression. Note now that the pragmatic model described above uses the same breakdown of the process of inferring reference. The “contextual salience” term we described above maps directly onto the referent choice term $P(I|O)$, and the Gricean informativeness term in Equation 8 maps onto the term $P(W|I)$.

With this equivalence in hand, we can reverse the pragmatics model and derive a word-learning version (a version of this derivation is given in Frank, Goodman, Lai, & Tenenbaum,

2009). In this version, we infer alternative meanings for a particular lexical item, given that a particular referent is known to be uttered. Consider the display in Figure 5. Imagine that a speaker pointed now to the literal target H+G, fixing the referent, but uttered a novel label, e.g. "fedora." In this case, the referent is known, but the meaning of a novel element in the lexicon L is unknown. We notate the possibility that a particular word in L refers to a feature (e.g. having a hat) as $w = f$. Using the same formulation from the intentional model above, we can write

$$\begin{aligned} P(L|I, w, C) &\propto P(I|L, w, C)P(L) \\ &\propto P(I|w = f, w, C)P(w = f), \end{aligned}$$

and if we assume that there is no prior reason to prefer one meaning for w over another ($p(w = f) \propto 1$) and substitute from Equation 8, then we have

$$P(L|I, w, C) \propto \frac{1}{|f|}. \tag{9}$$

In other words, all else being equal, a word is likely to have the meaning that would be informative in this context. So "fedora" would be more likely to refer to the hat (the feature that is most informative in this display) than the glasses. Thus, the pragmatic model described above can be used to capture inferences about the likely meanings of words in context.

In Frank, Goodman, Lai, and Tenenbaum (2009), we presented data that this relation in fact fit adult participants' judgments about novel adjective meanings with high accuracy. We used schematic displays of shapes that reproduced the same kind of game as shown in Figure 5 and varied the number of shapes with each property (in Figure 5 this would be the number of faces with hats vs. the number with glasses). As the relative extensions changed, participants altered their guesses about novel adjective meanings, suggesting that (at least as a group) they were sensitive to the relative informativeness of different possible word meanings. Although this work is still ongoing, we now have preliminary data from young children that they are able to make such judgments as well.

5 Conclusions

From a remarkably early age, children are not just passive absorbers of linguistic input. Instead, they are participants in conversation. At first this is a role we impose on them, scaffolding opportunities for turn taking and exchange, but soon enough, they become independent communicators. This communication provides their primary input for word learning. I've argued here that the task of understanding language in the moment—and ascertaining reference in particular—interacts powerfully with the language learning task.

The interaction between interpretation and learning leads to a class of models that I have referred to as communicative or intentional models. In the taxonomy described above, these models are dissociated from associative models not because of the information they include—both model classes can take advantage of social information—but because of the

way they break down the learning task. While associative models use social information to bias associative learning, intentional models use social information (as well as pragmatic inference) to inform a guess about what speakers are trying to say (their "referential intention"). This guess mediates between the physical context of the conversation and the words that are uttered.

I have argued here that intentional models are a powerful framework for using social and pragmatic information in the service of learning the meanings of words. The evidence is strong that by the age of 18 months, children take advantage of this information and make inferences that go beyond the association of words and objects (Baldwin, 1993; Hollich et al., 2000). Yet it is debated whether they begin the language learning process this way or whether infants begin with an associative model of language use (for opposed views on the subject, see e.g. Hollich et al., 2000; Vouloumanos et al., 2012). Since recent evidence indicates the possibility of word knowledge in even younger infants than had previously been suspected (Bergelson & Swingley, 2012), infants' assumptions during this early learning remain an important open question for both empirical and computational investigation.

Independent of whether infants begin life as probabilistic intentional learners, it seems likely that they converge to this position as their vocabulary expands to encompass terms that cannot be learned via contextual associations. For a learner who only need acquire basic level descriptors in a word of repeated exposures, the consequences of intentional learning are relatively modest. But for a learner of a language that contains a wide variety of complex, context-dependent predicates, it is essential to understand the contribution of the speaker's communicative intentions to the words they utter.

References

Akhtar, N., & Montague, L. (1999). Early lexical acquisition: The role of cross-situational learning. *First Language*, *19*, 347–358.

Baker, C., Saxe, R., & Tenenbaum, J. (2009). Action understanding as inverse planning. *Cognition*, *113*, 329-349.

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

Baldwin, D. (1995). Understanding the link between joint attention and language. In C. Moore & P. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 131–158). Lawrence Erlbaum Associates, Inc.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*, 3253–3258.

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

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

Callanan, M. (1985). How parents label objects for young children: The role of input in the acquisition of category hierarchies. *Child Development*, 508–523.

Carey, S. (1978). The child as word learner. In *Linguistic theory and psychological reality* (pp. 264–293). Cambridge, MA: MIT Press.

Carpenter, M., Nagell, K., & Tomasello, M. (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*.

Chater, N., & Oaksford, M. (2008). *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford University Press, USA.

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

Clark, H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*, 148–153.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*, 1017–1063.

Fisher, C., Hall, D., Rakowitz, S., & Gleitman, L. (1994). When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. *Lingua*, *92*, 333–375.

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

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

Frank, M. C., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Frank, M. C., Goodman, N., & Tenenbaum, J. (2007). A Bayesian framework for cross-situational word learning. *Advances in Neural Information Processing Systems*, *20*.

Frank, M. C., Goodman, N. D., Lai, P., & Tenenbaum, J. B. (2009). Informative communication in word production and word learning. In *Proceedings of the 31st 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., Tenenbaum, J. B., & Fernald, A. (in press). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language, Learning, and Development*.

Gallistel, C. (1990). *The organization of learning.* The MIT Press.

Gallistel, C., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences*, *101*, 13124–13131.

Geisler, W. (2003). Ideal observer analysis. In *The visual neurosciences* (pp. 825–837). Cambridge, MA: MIT Press.

Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature*, *415*, 755.

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

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

Goodman, N., & Stuhlmüller, A. (2012). Knowledge and implicature: Modeling language understanding as social cognition. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

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.

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*.

James, W. (1890). *The principles of psychology, vol i.* New York, NY: Henry Holt and Company.

Johnson, M., Demuth, K., & Frank, M. (2012). Exploiting social information in grounded language learning via grammatical reductions. In *Proceedings of the Association for Computational Linguistics.*

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

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks*, *17*, 1345–1362.

Locke, J. (1690/1964). *An essay concerning human understanding.* Cleveland, OH: Meridian Books.

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

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

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological*

Review, *119*, 831 – 877.

Medina, T., Snedeker, J., Trueswell, J., & Gleitman, L. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*, 9014–9019.

Onishi, K., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*, 255–258.

Perfors, A., Tenenbaum, J., Griffiths, T., & Xu, F. (2011). A tutorial introduction to bayesian models of cognitive development. *Cognition*, *120*, 302–321.

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

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

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

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

Schmidt, L., Goodman, N., Barner, D., & Tenenbaum, J. (2009). How tall is tall? Compositionality, statistics, and gradable adjectives. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2759–2764).

Shafto, P., Goodman, N., & Frank, M. (2012). Learning from others the consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*, 341–351.

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

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

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

Stiller, A., Goodman, N., & Frank, M. (2011). Ad-hoc scalar implicature in adults and children. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*.

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

Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279–1285.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (in press). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*.

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

Vouloumanos, A., Onishi, K., & Pogue, A. (2012). Twelve-month-old infants recognize that speech can communicate unobservable intentions. *Proceedings of the National Academy of Sciences*, *109*, 12933–12937.

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

Waxman, S., & Booth, A. (2001). Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives. *Cognitive Psychology*, *43*, 217–242.

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

Yu, C., & Smith, L. (2012). Modeling cross-situational word–referent learning: Prior questions. *Psychological Review*, *119*, 21.