Inferring word meanings by assuming that speakers are informative

Michael C. Frank and Noah. D Goodman[1]

Department of Psychology, Stanford University

Abstract

Language comprehension is more than a process of decoding the literal meaning of a speaker's utterance. Instead, by making the assumption that speakers choose their words to be informative in context, listeners routinely make pragmatic inferences that go beyond the linguistic data. If language learners make these same assumptions, they should be able to infer word meanings in otherwise ambiguous situations. We use probabilistic tools to formalize these kinds of informativeness inferences—extending a model of pragmatic language comprehension to the acquisition setting—and present four experiments whose data suggest that preschool children can use informativeness to infer word meanings and that adult judgments track quantitatively with informativeness.

Keywords: Language acquisition, Pragmatics, Word learning, Bayesian models

1. Introduction

Children learn the meanings of words with remarkable speed. Their vocabulary increases in leaps and bounds relatively soon after the emergence of

[☆]Many thanks to Allison Kraus, Kathy Woo, Janelle Klaas, Andrew Weaver, and Stephanie Nicholson for assistance in stimulus design and data collection and to Susan Carey, Ted Gibson, Avril Kenney, Peter Lai, Rebecca Saxe, Jesse Snedeker, and Josh Tenenbaum for valuable discussion. Some ideas described in this paper were originally presented to the Cognitive Science Society in Frank et al. (2009a). We gratefully acknowledge the support of ONR grant N00014-13-1-0287.

[1]Please address correspondence to Michael C. Frank, Department of Psychology, Stanford University, 450 Serra Mall (Jordan Hall), Stanford, CA, 94305, tel: (650) 724-4003, email: mcfrank@stanford.edu

productive language (Fenson et al., 1994), and they often require only a small amount of exposure to begin the process of learning the meaning of an individual word when it is presented in a supportive context (Carey, 1978; Markson & Bloom, 1997). The ability to infer and retain a huge variety of word meanings is one of the signature achievements of human language learning, standing alongside the acquisition of discrete phonology and hierarchical syntactic and semantic structure (Pinker & Jackendoff, 2005).

Nevertheless, figuring out what an individual word means can be a surprisingly difficult puzzle. In Quine's 1960 classic example, he considers an anthropologist who observes a white rabbit running by. One of his subjects points and says "gavagai." Even assuming that he interprets the pointing gesture as a signal of reference (Wittgenstein, 1953; Tomasello, 2008), the anthropologist must still infer *which* property of the rabbit the word refers to. Some properties may be logically impossible to distinguish from one another—think "rabbit" and "undetached mass of rabbit parts." But beyond these philosophical edge cases, even useful properties can be strikingly difficult to distinguish: how can he decide between "rabbit," "animal," "white," "running," or even "dinner"? We can think of this as an easy—but perhaps more common—version of the Quinian puzzle: for any known referent (the rabbit), there are many conceptually natural referring expressions that include the referent in their extension (Gleitman & Gleitman, 1992).[2] Our argument here is that many of these can be ruled out on pragmatic grounds, by considering the communicative context and the goals of the speaker.

Language learners have many tools at their disposal to try and narrow down the possibilities, including patterns of consistent co-occurrence (Yu & Smith, 2007), the contrasting meanings of other words they have learned (Clark, 1988;

[2]This easy puzzle is of course distinct from the harder version, the "true" Quinian puzzle: that there are *infinitely* many conceptually *possible* referring expressions that include the referent in their extension, and some of these (think "rabbit" and "undetached mass of rabbit parts in the shape of a rabbit") are extensionally identical.

Markman & Wachtel, 1988), and the syntactic structure in which the word appears (Gleitman, 1990). In the current work, however, we consider cases where these strategies are ineffective, yet learners can nevertheless infer word meanings by considering the speaker's communicative goal.[3] These are cases where the pragmatics of the situation—roughly speaking, the fact that a particular communicator is trying to achieve a particular goal in this context, and that he or she is following a rational strategy to do so—help in inferring word meaning. In our Quinian example, the intuition we are pursuing is that the anthropologist may consider information necessary in the context in assigning a tentative meaning to "gavagai." If the white rabbit is tailed by a brown one, perhaps "gavagai" means WHITE, while in the absence of such a context, a basic level object label might be more appropriate.

Consider the analogous—though simplified—case in Figure 1. If a speaker describes the dinosaur on the right (marked by the arrow) as "a dinosaur with a dax," the novel word could mean HEADBAND or BANDANNA, or even in principle TAIL or FOOT. All of these meanings for "dax" would make the speaker's statement truthful. Nevertheless, several of these would be quite odd things to say: although that dinosaur has one foot, it's also true that he has two (and for that matter, so does the other dinosaur as well). On the other hand, if "dax" meant HEADBAND, then it would be quite an apt description in the current context. Hence, this example might provide evidence to a pragmatically-savvy learner that "dax" has the meaning HEADBAND.

Importantly, there is no cross-situational information present in this single

[3]We use the term "inference" to distinguish between the process of figuring out what a word means and the later retention of that meaning. Retention is a necessary component of learning (and there may be cases, for example ostensive naming, where retention is the *only* component of learning). Nevertheless, we are interested here in the process of inference in ambiguous situations. We also note that the use of the term "inference" does not connote to us that the psychological computation is necessarily symbolic or logical. Statistical inferences of the type described below can be instantiated in probabilistic logics, neural networks, or just about any other formalism (MacKay, 2003).

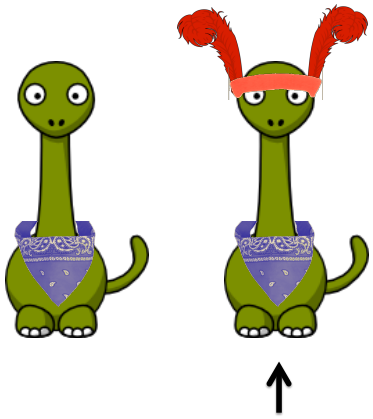

Figure 1: An example stimulus item for our experiments. The arrow represents a point or some gesture that signals that the dinosaur on the right is being talked about, but does not give away which aspect of it is being referred to. In our experiments, the goal of the learner is to infer whether a novel word (e.g. "dax") means BANDANNA or the HEADBAND.

scenario, and neither the learner's previous vocabulary nor the syntax of the sentence reveal the word's meaning. Yet the intuition is still quite clear that HEADBAND is a more likely candidate (and our experiments reported below confirm this intuition). Although not accounted for by the classic set of acquisition strategies, inferences like this one fit well with theories of pragmatic reasoning in language comprehension.

Philosophers and linguists have long suggested that language relies on shared assumptions about the nature of the communicative task that allow comprehenders to go beyond the truth-functional semantics of speakers' utterances. Most canonically, 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. Other theories of pragmatic communication also provide related tools for explaining this type of inference. For example, Sperber & Wilson (1986) have suggested that there is a shared "Principle of Relevance" which underlies com-

munication. On their account, the key part of this interaction is the shared knowledge between speaker and listener that the headband is the most relevant feature of the dinosaur in this context; otherwise the inference is largely the same. Many additional neo-Gricean formulations have also been proposed (e.g. Clark, 1996; Levinson, 2000). Here we use the original Gricean language because it is best known, but our ideas do not depend specifically on Grice's formulation.

Returning to the example in Figure 1, if the speaker is trying to pick out the dinosaur on the right, then using a word that referred to the HEADBAND would be a good choice. This choice would typically be motivated with reference to Grice's Maxim of Quantity, which impels speakers to "be informative" (though we return below to the question of how to provide an operational definition for "informativeness"). The inference that "dax" means HEADBAND goes beyond the simple application of Gricean reasoning, however.

To infer that "dax" means HEADBAND, the learner must presuppose that the speaker is being informative and then use this assumption, working backwards, to infer the meaning of a word (rather than the intended meaning of the speaker's utterance, as is more typical in Gricean situations). This inference has a counterfactual flavor: If the speaker were being informative, they would have said something that referred to the HEADBAND; they said "dax," whose meaning I don't know; therefore perhaps "dax" means HEADBAND. Can children make this kind of inference in the course of language acquisition? If so, such inferences could be an important tool for eliminating some of the referential uncertainty inherent in learning a new word. We next consider related evidence on children's pragmatic abilities.

While many 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; Clark, 2003; Tomasello, 2003), evidence on children's use of Gricean maxims specifically is mixed. On the one hand, an influential body of work suggests that young children can use pragmatic inferences to learn the meanings of words. For example, 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 (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. And in a surprising recent demonstration of such abilities, Southgate et al. (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. Thus, by their second birthday, children appear to be able to make relatively sophisticated inferences about speakers' knowledge and intentions in word learning situations.

On the other hand, another body of work suggests that much older children still struggle to make pragmatic inferences in language production and comprehension—or at least that what inferences they do make can often be explained in other ways. Even five-year-old children have trouble understanding what information is available to communicative partners (Glucksberg et al., 1966), though more recent evidence has shown some sensitivity to speaker knowledge in online measures (Nadig & Sedivy, 2002). In addition, Gricean reasoning has not been observed for children younger than four years, and is seen only inconsistently before the age of six. For example, 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 (but cf. Eskritt et al., 2008). 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 et al., 2001; Gualmini et al., 2001; Guasti et al., 2005 for alternative accounts) until quite late (Noveck, 2001).

Nevertheless, accounts differ considerably on the age at which children first succeed in making implicatures (Papafragou & Musolino, 2003; Guasti et al., 2005) and on the factors that prevent them from succeeding (Barner & Bachrach, 2010; Barner et al., 2011; Katsos & Bishop, 2011). It may be that the specifics of scalar implicature are difficult for young children. Much younger children

are also sensitive to the informativeness of their own and others' communication (O'Neill & Topolevec, 2001; Liszkowski et al., 2008; Matthews et al., 2007, 2012). And some evidence indicates that preschoolers can make other kinds of pragmatic inferences slightly earlier (Kurumada, 2013; Stiller et al., 2014), though none as early as the "pragmatic word learning" findings summarized above.

To summarize, the evidence on children's pragmatic abilities is mixed. Children are sensitive to aspects of speakers' goals and beliefs in word learning, and certainly they make substantial use of social cues like eye-gaze and gesture. But it is still unknown how well they are able to use Gricean reasoning to infer word meanings. We have suggested that the Gricean maxim of quantity ("be informative") may help learners infer word meanings in otherwise ambiguous situations, but whether children—or even adults—are in fact able to make these inferences remains an open question. The current work investigates this issue.

A key challenge in providing a Gricean account for word learning is defining "informativeness," a concept that is often left frustratingly vague. Without a clear account of what makes a particular term or utterance informative in context, we are left with a theory that fails to make concrete and easily-tested predictions (Pea, 1979). For this reason, our work here uses a computational formulation of the idea that speakers are informative, using tools from information theory to make quantitative predictions in simple situations like Figure 1. This framework builds on our recent work modeling adults' pragmatic judgments as a process of probabilistic inference (Frank & Goodman, 2012). Its value here is that it allows us to make quantitative predictions about behavior in a range of cases where previous theories have made at best directional predictions. The next section describes this framework and its application to word learning.

Our experiments then test predictions derived from this framework. In Experiment 1, we make a quantitative test with adults and find that there is high correspondence between adults' aggregate judgements about the meanings of novel words and the predictions of our model. Experiments 2 and 3 then test whether preschool children are able to make similar inferences in simplified

cases. Experiment 4 replicates the finding of Experiment 3 and rules out a number of alternative explanations. Together our results suggest that adults and children are sensitive to the relative informativeness of labels and can use this information to make inferences about the meanings of novel words in ambiguous situations.

2. Modeling pragmatic inference in word learning

In order to motivate its use in word learning, we begin by giving a brief exposition of the probabilistic model of pragmatic inference introduced in Frank & Goodman (2012). We then show how this model can be adapted to make predictions from the perspective of a language learner who has uncertainty about what individual terms mean. The probabilistic modeling framework provides a convenient tool for formalizing this set of ideas. Related predictions can be derived in a game-theoretic framework for pragmatics (Benz et al., 2005; Franke, 2009; Jäger, 2010; Franke, 2013), though to our knowledge such a framework has not been used to model language learning.

The model introduced in Frank & Goodman (2012) describes normative inferences in simple reference games under the assumption that listeners view speakers as having chosen their words *informatively*—that is, relative to the information that they would transfer to a naive listener (see also Goodman & Stuhlmüller, 2013). The heart of our model is the idea that a rational[4] listener will attempt to make inferences about the speaker's intended referent r_s, given the word w they uttered, the lexicon of their language L, and the context C. This inference can be described using Bayes' rule:

[4]Note that the use of the term "rational" here does not imply a claim of *human* rationality, much less optimality (Frank, 2013). Our current experiments test that the predictions of such a model are satisfied by the aggregate judgments of many human observers; these data leave open the question of the psychological mechanisms that produce the observed patterns of human performance.

$$P(r_S|w, L, C) = \frac{P(w|r_S, L, C)P(r_S)}{\sum_{r' \in C} P(w|r', L, C)P(r')}. \quad (1)$$

In other words, the posterior probability of some referent is proportional to the product of two terms: the likelihood $P(w|r_S, C)$ that some word is used to describe a referent, and the prior probability $P(r)$ that this referent will be the subject of discourse. Because the situations we treat here all assume that the speaker knows the intended referent r_S, we do not discuss the prior term further (for more details see Frank & Goodman, 2012).

We defined the likelihood of a word being used to describe some referent as proportional to a formal measure of the information transfered by an utterance (its surprisal given the base context distribution). This information-theoretic definition of what it means to be "informative" leads to

$$P(w|r_S, L, C) = \frac{|w|_L^{-1}}{\sum_{w' \in W} |w|_L^{-1}}, \quad (2)$$

where $|w|_L$ refers to the number of objects in a particular context to which w can truthfully be applied, given the known meaning of w in L. In other words, "be informative" translates to "say words that apply to your referent and few others," which seems to approximate the general Gricean intuition.

A Bayesian learner can use the assumption that speakers are informative to learn the meaning of unknown words. A language learner often has uncertainty about *both* the speaker's intended referent and the lexicon mapping words to their meanings, which we notate L (a simple version of this case is treated in our work on cross-situational learning in Frank et al., 2009b). But although our framework can be extended to this case of joint uncertainty about meaning and reference, we focus here on the case where the referent is known and we must infer only word meanings. (This is the "easy" Quinian case described above, where the rabbit is indicated but the meaning of "gavagai" is unknown).

In the case where we know the speaker's intended referent, we can now reverse the inference and write the probability of a lexicon L, given the observation

of a word w used to refer to some object r_S:

$$P(L|w, r_S, C) \propto P(w|L, r_S, C)P(L) \tag{3}$$

We next walk through the case shown in Figure 1. We assume that the speaker's intended referent (r_S) has two truth-functional features f_1 and f_2 (HEADBAND and BANDANNA), and that there are two words in the language w_1 and w_2. We further assume that each word has exactly one meaning linked to it.[5] Hence there are only two possible lexicons: $L_1 = \{w_1{=}f_1, w_2{=}f_2\}$ and $L_2 = \{w_1{=}f_2, w_2{=}f_1\}$, which are equally probable.

Under these assumptions,

$$\begin{aligned} P(L_1|w_1, r_S, C) &= \frac{P(w_1|L_1, r_S, C)}{P(w_1|L_1, r_S, C) + P(w_1|L_2, r_S, C)} \\ &= \frac{\frac{|f_1|^{-1}}{|f_1|^{-1}+|f_2|^{-1}}}{\frac{|f_1|^{-1}}{|f_1|^{-1}+|f_2|^{-1}} + \frac{|f_2|^{-1}}{|f_2|^{-1}+|f_1|^{-1}}} \\ &= \frac{|f_1|^{-1}}{|f_1|^{-1} + |f_2|^{-1}}, \end{aligned} \tag{4}$$

where $|f|$ indicates the number of objects with feature f (substituting Equation 2 for the second step by noting that word w would be used informatively depending on the extension of the relevant feature). Note that, as in Frank & Goodman (2012), this computation requires no parameter values to be set by hand.

Returning now to the example in Figure 1, we can use Equation 4 to calculate the probability that learners judge that w ("dax") means HEADBAND (f_1) as opposed to BANDANNA (f_2):

[5] Relaxing this assumption has interesting consequences with respect to "mutual exclusivity" inferences, which are treated in more depth in Frank et al. (2009b) and Lewis & Frank (2013).

$$
\begin{aligned}
P(w = f_1 | M_S, C) &= \frac{|\text{HEADBAND}|^{-1}}{|\text{HEADBAND}|^{-1} + |\text{BANDANNA}|^{-1}} \\
&= \frac{\frac{1}{1}}{\frac{1}{1} + \frac{1}{2}} = \frac{2}{3}
\end{aligned}
$$

Thus, our prediction—all else being equal—is that learners should be around 67% confident that "dax" means HEADBAND, because the feature HEADBAND has the smaller extension in context.

Of course, there are many other aspects of the situation that might alter this prediction. For example, we assume that there are no alternative competitor meanings for "dax" that are considered in participants' judgments; indeed our experiments use a two-alternative forced choice for this reason. If we were to allow participants to consider other competitor meanings (such as LONG NECK or ON THE LEFT), the denominator in Equation 4 would grow, causing the overall prediction for HEADBAND to go down. If such competitors were included, a natural next step would be to attempt to measure learners' prior expectations about the types of features that are typically named (rather than leaving this prior uniform as we have here). In these initial experiments, however, we test the general form of the model rather than how it would be extended to larger feature sets.

To summarize, 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. We recover the "size principle" of Tenenbaum & Griffiths (2001) (see also (Xu & Tenenbaum, 2007)). This principle originated with Shepard's 1987 work on generalization behavior in psychological spaces and has more recently been rederived by Navarro & Perfors (2009). Our work 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. In the following

experiments we test whether adults and preschoolers are sensitive to contextual informativeness in their inferences about word meanings.

3. Experiment 1

Our first experiment investigated whether adult word learners could make inferences about word meaning on the basis of the relative informativeness of a word in context. We were additionally interested in whether these judgments conformed quantitatively to the framework described above. To test these hypotheses, we asked adults for quantitative judgments about the meanings of novel words in situations like Figure 2, left. We used these slightly more complex displays to allow for the controlled manipulation of the relative extensions of the two candidate features.

3.1. Methods

3.1.1. Participants

We recruited 201 unique individuals on 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 United States, in the case of our experiments) for a small amount of compensation (Buhrmester et al., 2011; Crump et al., 2013).

3.1.2. Materials and Methods

Each participant completed a short survey that included 4 questions about what words meant. Each question showed a stimulus picture containing three objects (dinosaurs, rockets, bears, or robots), with one target indicated by a box around it. Each object had two features (e.g. bandanna, headband). Participants were told that someone had used a word in a foreign language (e.g. "daxy") to refer to the object with the box around it and asked to make bets on which feature the word referred to. An example stimulus is shown in Figure 2, left. The assignment of object to condition, the position of the target object, and target feature were all counterbalanced between subjects.

Trials were arranged into one of the four conditions (1/1, 1/2, 1/3, and 2/3). Conditions refer to the arrangement of features among the three objects: the numerator refers to the number of objects with the first feature. The denominator refers to the number of objects with the second feature. Consider the example in Figure 2, left: the target dinosaur (with the box around it) has two features. The first—by convention, the one with a smaller extension, in this case the headdress—is unique to that object, so the numerator is 1. The second, the bandanna, is shared with another dinosaur. Thus, this trial is a 1/2 trial.

Following this convention, a 1/1 trial was a trial in which a target object with two features, each of which was unique to that object. A 1/2 trial was a trial in which one of the target object's features was unique and the other was shared with one other object (as in our example). A 1/3 trial had a target with a single unique feature and a second feature shared with all three objects. Finally, a 2/3 trial target had no unique features, but had one feature shared with a single other object and one feature shared with both other objects.

In each trial in the survey, the participant was asked to make one judgment, in the form of a "bet" of $100 dollars on whether a novel adjective referred to one or the other property of the object with the box around it, spreading the money between the two alternatives by entering two numerical values (the two alternatives were denoted by a picture next to each text box). This betting measure gives us an estimate of speakers' subjective probability, rather than a purely qualitative judgment (Frank & Goodman, 2012). For each trial, we also included two manipulation check questions, in which we asked participants to write how many objects had each of the two target features (Oppenheimer et al., 2009; Crump et al., 2013).

3.2. Results and Discussion

In our analysis, we excluded trials on which participants' bets did not sum to 100 (2.5% of trials) and on which they failed to answer the check questions correctly (2.9%). These exclusions did not change the qualitative or quantitative pattern of results. We also verified that there were no effects of object type or

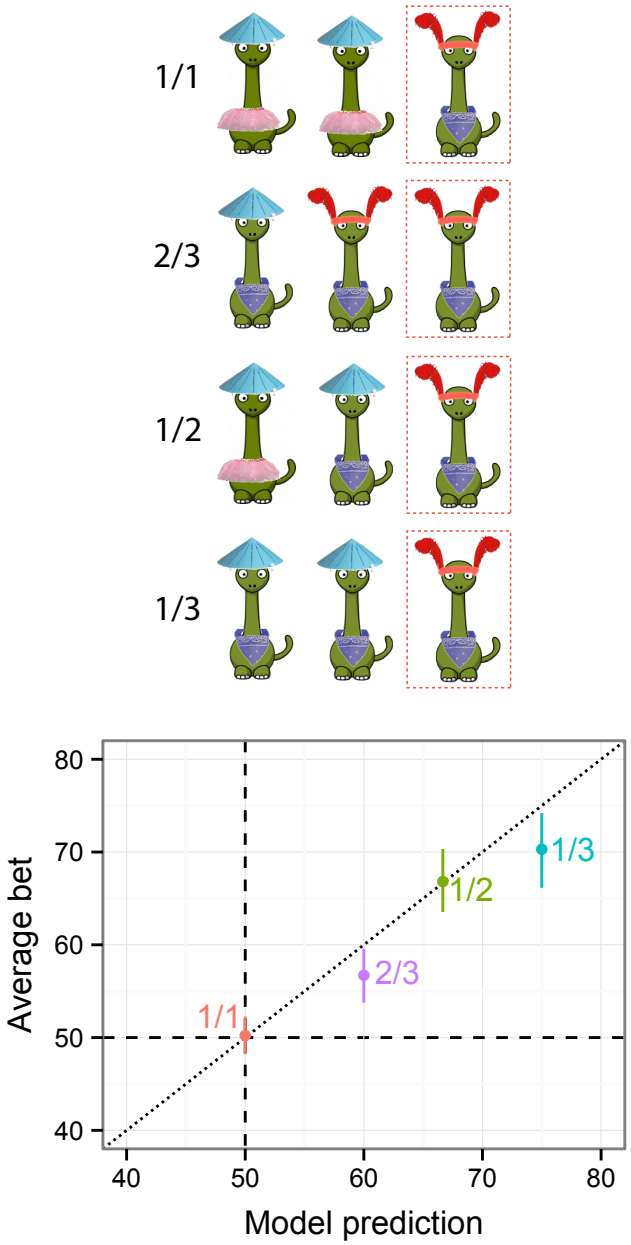

Figure 2: Top: Stimuli for Experiment 1 in one version of the four trial types (see text for description of condition labels). Bottom: Data from Experiment 1. Points show participants' mean bet with 95% confidence intervals (computed via non-parametric bootstrap), plotted by the predictions of the informative communication model. The dashed lines show chance responding for human responders and model; the dotted line shows the diagonal, indicating perfect correspondence between model and human data).

Trial	Model prediction	M	SD	t	df	p
1/1	50.0	50.2	14.1	0.21	193	.83
1/2	66.7	66.8	23.1	10.01	189	$< .0001$
1/3	75.0	70.3	27.7	10.19	193	$< .0001$
2/3	60.0	56.7	19.8	4.65	188	$< .0001$

Table 1: Summary statistics and two-tailed one-sample t-tests against chance performance ($50) for each condition in Experiment 1. M, and SD denote the mean and standard deviation of bets on the target feature. Degrees of freedom vary from condition to condition due to exclusions (see text for more details).

target position in a simple linear regression predicting participants' bets.[6] Thus, we averaged across these aspects of the data and analyzed bets on the target feature by condition. The target feature was designated as the feature that constituted the numerator in the condition name, e.g. the unique feature in 1/2 trials (HEADBAND in our running example).

Participants' mean bets on the target feature are plotted by model predictions in Figure 2, right, and all data are reported in Table 1. The primary prediction in our experiment was that participants' bets would favor the features that were more informative (had smaller extensions). We found that this prediction was satisfied: In the 1/2, 1/3, and 2/3 conditions (all the conditions where there was a difference in extension between features), participants picked the feature with the smaller extension significantly more than chance (t-tests are reported in Table 1). In addition, all three of these conditions differed significantly from the 1/1 baseline condition (all ps $< .001$). Thus, participants

[6]Due both to a desire to maintain comparability with the development experiments (Experiments 2 and 3) and due to concerns that participants would pick up on consistencies in the type of inferences being studied, we limited the number of trials that any given participant completed. Thus, the amount of data we collected for each participant did not allow us to make accurate estimates of participant-level effects in a mixed-effects model (Gelman & Hill, 2006; Jaeger et al., 2011).

in our experiment reliably assigned the meaning of the novel word to the more informative feature of the target object in the context.

In addition, participants' bets scaled with the relative informativeness of the two features. We found a tight quantitative correspondence between (parameter-free) model predictions and human behavior. When there were equal numbers of objects with each feature, mean bets were very close to \$50, reflecting equal probability. In contrast, in the 1/2 case shown in Figure 2, our informativeness model predicted a bet of \$67 in this condition, nearly identical to the participants' average bet of \$67. Although there were only four conditions with distinct model predictions, the correlation between mean bets and model predictions was quite high ($r = .98$, $p = .02$).[7]

Despite the high correlation between model and data, participants' bets were overall slightly lower than predicted by the model (the 2/3 and 1/3 conditions were below the diagonal in Figure 2). This trend is consistent with the idea that human judgments includes some "lapses"—cases where participants make errors or pick at chance—and hence behavioral measures are biased towards chance relative to ideal observer model predictions (Wichmann & Hill, 2001; see Frank et al., 2010 for a recent exposition of this issue in the probabilistic language modeling literature).

Thus, our data suggest that adults' judgments show a quantitative correspondence between the relative informativeness of a property in context and inferences about word meaning. Our next experiments test whether preschool children also show evidence of such sensitivity to informativeness.

[7]This finding additionally replicates results of an adult experiment reported in Frank et al. (2009a) with a distinct population and stimulus set. In that experiment, which used arrays of six geometric shapes, there were a total of 21 conditions ranging from 1/6 to 5/6, and the overall correlation between model predictions and participants' mean bet was $r = .93$. We conducted this simplified version of the experiment in order to use stimuli more comparable to those used with children in Experiments 2 and 3.

4. Experiment 2

We next asked whether preschool children would also be able to make use of the informativeness of features to learn the meanings of novel adjectives. For this paradigm, we used a simplified version of the 1/2 condition of Experiment 1 that used only two objects and two features, as in our original example in Figure 1.

4.1. Methods

4.1.1. Participants

Participants were 24 children from an on-campus preschool, recruited from their classrooms by an experimenter who had previously spent time in their classroom to establish rapport. Children were recruited to fulfill a planned sample of 3 – 4 year-olds (N=12, mean age = 3;7) and 4 – 5 year-olds (N=12, mean age = 4;6).

4.1.2. Materials and Methods

Children completed eight total trials, distributed into two conditions: *filler* and *inference*. Inference trials contained two objects: the target object (indicated by a point) had two features, while the distractor object had only one of these (as in the running example shown in Figure 1). Filler trials were identical but the target had only one feature, which was not shared with the distractor. For example, a filler version of Figure 1 would be identical but the target dinosaur would appear without a bandanna, so that the label would unambiguously refer to the headband (because this was the only salient accessory the dinosaur had). Trials were interleaved by condition, with a filler trial always appearing first.

At the beginning of the paradigm, children were introduced to a stuffed animal named Felix who they were told was visiting a toy store and who they were to help in identifying some new toys. Experimental materials were presented via printed pictures shown in a binder, with training and testing phases shown on subsequent pages. In the training portion of each trial, the experimenter

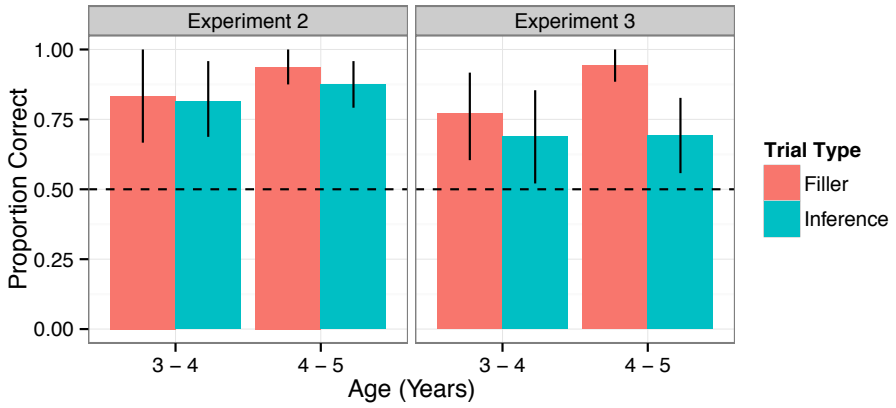

Figure 3: Data from Experiments 2 and 3. Mean proportion correct is plotted by age group for both filler and inference trials. The dashed line shows chance performance. Error bars show 95% confidence intervals, computed via a non-parametric bootstrap over participant means.

pointed to the target object and said e.g. "This is a dinosaur with a dax! How neat! A dinosaur with a dax." This frame ensured that the target word ("dax") was spoken twice. The first part of the naming phrase was always "this is a," while the exclamation varied from item to item to provide variety. In the test portion of the trial, children saw two additional images in which one object had each each feature (e.g. a dinosaur with a bandanna only and a dinosaur with a headband only; identical to the filler trials). They were asked "Here are some more dinosaurs. Which of these dinosaurs has a dax?" and responded by pointing.

Materials for the inference trials were identical to those used in Experiment 1; filler trials used monkeys, dogs, cell phones, and cats as the objects. Novel words were "tupe," "sep," "zef," "gabo," "dax," "fid," "keet," and "toma." We counterbalanced trial order, target position in both training and test trials (crossed), and which feature was the target. Features were chosen to be equally salient based on pilot studies using the same paradigm.

4.2. Results and Discussion

If children were able to make use of the relative informativeness of the two possible word meanings, they should choose the more informative word meaning significantly more often than chance. Congruent with this hypothesis, we found that in inference trials, children chose the unique feature (the one that would have been more informative to name in this context) the majority of time (3–4 year olds: M=81%, SD=39% and 4–5 year olds: M=88%, SD=33%) and nearly as often as they chose the correct feature in filler trials (3–4 year olds: M=83%, SD=38% and 4–5 year olds: M=94%, SD=24%). Results are shown in Figure 4, left. These data suggest that children in our task were sensitive to the contextual distribution of features, even though the literal meaning of the utterance did not strictly rule out the non-unique feature.

To quantify the reliability of this pattern, we fit a logistic mixed effects model (Gelman & Hill, 2006; Jaeger, 2008) to children's responses, with age group and condition as fixed effects, and with random effects of condition fit for each participant and each target item (a "maximal" random effect structure Barr et al., 2013). The resulting coefficient estimates suggested that three-year-olds (the reference level) were above chance in their responding on inference trials ($\beta = 1.74$, $z = 3.70$, $p = .0002$). There was also a significant coefficient indicating higher performance on filler trials ($\beta = 4.66$, $z = 1.92$, $p = .02$). In this study there was no significant effect of age group ($\beta = .47$, $z = .67$, $p = .51$). A model with an interaction term did not provide better fit ($\chi^2(1) = .16$, $p = .69$), though under this model the coefficient estimate for filler trials was slightly lower and only trended towards significance ($\beta = 3.91$, $p = .09$); the reliability of other results did not change.

Evidence from this study suggests that children successfully mapped words to features that would have been more likely to be named by an informative speaker. The mean proportion of informativeness-congruent judgements by children in both groups was actually higher than the strict probability assigned by our model (67%) and higher than that assigned by adults in the betting task in Experiment 1. There are several reasons to be cautious about this kind of

quantitative interpretation, however. The context of Experiment 2 was far less stripped down than that of Experiment 1, and the linguistic frame for the novel label encouraged a contrastive reading (something we investigate in Experiment 3). In addition, the two-alternative forced choice measure might have led children to *maximize*, more consistently choosing the highest-probability of the two alternatives (Hudson-Kam & Newport, 2005). Thus, although the evidence strongly points in favor of informativeness, we do not believe a quantitative interpretation is warranted.

5. Experiment 3

As mentioned above, one question about the findings of Experiment 2 comes from the use of the contrastive sentence frame "This is a dinosaur with a dax." The deictic "this" is, in the terminology of Clark & Wong (2002), a "direct offer"—the use of a deictic term for the exclusive purpose of providing a label. This exclusive purpose may have given participants a greater sense that the utterance should be chosen with maximal informativeness. While the goal of the label is ambiguous—either to teach the label itself or to distinguish one dinosaur from the other—both would lead to a strong presumption of informativeness. In addition, the deictic "this" is easy to stress contrastively, implying to listeners that "this [and not that other one] is a dinosaur with a dax."

In Experiment 3, we replicated the methods of Experiment 2 exactly but used the frame "here is a" instead. By virtue of its focus on location, rather than identity, "here is a" provides an alternative goal for the utterance: establishing in the common ground the location of a particular dinosaur (Clark, 1996). In addition, in Experiment 3, we avoided the strong prosodic phrase boundary between "here" and "is" that would be necessary to imply contrastive stress in this condition (e.g. "here... is a dinosaur with a dax"). A mapping of "dax" to the unique feature in this study would imply that the results of Experiment 2 are not specific to a single construction type.

5.1. Methods

5.1.1. Participants

Participants were 25 children from the same on-campus preschool as Experiment 2. Children were recruited to fulfill a planned sample of 3 – 4 year-olds (N=12, mean age = 3;8) and 4 – 5 year-olds (N=13, mean age = 4;3).

5.2. Materials and Methods

Materials and methods for Experiment 3 were identical to those in Experiment 2 except that we replaced the naming phrase "This is a" with the phrase "Here is a."

5.3. Results and Discussion

Results are shown in Figure 4, right. Overall, performance in the inference trials was lower than in Experiment 2, but was still above chance (3–4 year olds: M=69%, SD=47% and 4–5 year olds: M=69%, SD=47%). Filler trial performance remained quite high (3–4 year olds: M=77%, SD=42% and 4–5 year olds: M=94%, SD=24%).

We again applied logistic mixed effects regression, though in this case we retained the interaction between condition and age because it increased model fit. We found that three-year-olds in the inference condition were significantly above chance ($\beta = .93$, $z = 2.17$, $p = .03$), and there was no main effect of age group ($\beta = .04$, $z = .06$, $p = .95$). Performance on filler trials was higher than on inference trials, though not significantly so ($\beta = 1.04$, $z = 1.42$, $p = .15$), but there was a marginally significant interaction of trial type and age group ($\beta = 2.02$, $z = 1.64$, $p = .10$). This interaction suggests that the age-related increase in filler trial performance was not seen reliably in the inference trials—both 3–4 and 4–5 year olds were above chance, but they were not even numerically different from one another.

Although children's performance was numerically lower in the inference trials in Experiment 3 than it was in Experiment 2, we nevertheless replicated the use of informativeness to make inferences about word meaning. Either the "this

is a" construction or the stress with which it was marked likely contributed to the somewhat higher level of inferences in Experiment 2, and in naturalistic situations, these information sources both likely scaffold children's performance in making similar inferences. But even in their absence, children still appeared to notice the differential informativeness of the unique feature and treat that property as the extension of the novel word.

Nevertheless, two alternative explanations remain possible. Because the unique feature could have been more salient to children, they might either have noticed that feature and then simply selected it at test (a completely non-linguistic explanation) or they could have noticed it and assumed it was being talked about, but failed to encode its link with any particular word. Both of these explanations would bear against our hypothesis about the role of informativeness inferences for word learning. Experiment 4 provides control conditions that rule these explanations out.

6. Experiment 4

In Experiment 4, we replicated Experiment 3 with a slightly larger sample of children (24, rather than 12, per cell). In addition, we added two comparison conditions. The first, the *Disambiguation* condition, was intended to test that children made a connection between the word they heard and the feature they indicated at test. To test this hypothesis, we exploited the finding that children will reliably choose an unnamed or novel object when asked about the referent of a new word (Markman & Wachtel, 1988; Mervis & Bertrand, 1994). We taught a first novel word using the same paradigm as in Experiment 3, but then asked for a second, distinct novel word at test. If they had made a connection between the first word and the informative feature, then children should choose the uninformative feature in this test.

The second comparison condition, the *Non-Linguistic Salience* condition, simply asked children to "find another one" at test. This condition tested the hypothesis that children would choose features that matched the unique feature

at test irrespective of the presence of a label at all. We hypothesized that if children were relying on a linguistic inference (as proposed above), they would select features at chance in this condition.

6.1. Methods

6.1.1. Participants

Participants were 144 children recruited from the San Jose Children's Discovery Museum. Children were recruited to fulfill a planned sample of 3 – 4 year-olds (N=24 per condition, M_{rep} = 3;6, $M_{disambig}$ = 3;6, $M_{salience}$ = 3;8) and 4 – 5 year-olds (N=24 per condition, M_{rep} = 4;5, $M_{disambig}$ = 4;5, $M_{salience}$ = 4;7). In the replication condition, an additional 4 children completed the task but were excluded from the final sample (3 for falling under a pre-specified criterion of 75% parent-reported English exposure, 1 for experimenter error); in the Disambiguation condition, 3 children were excluded (2 for language, 1 for non-compliance); and in the Non-Linguistic Salience condition, 2 were excluded (both for language).

6.2. Materials and Methods

Materials and methods for Experiment 4 were identical to those in Experiment 3, except as noted. In the inference trials of the *Disambiguation* condition, we used a different novel name at test than was used in training. For example, if the child was taught about a "dax" in training, he or she might be asked about a "toma" at test. In both the inference and filler trials of the *Non-Linguistic Salience* condition, we did not name the feature in training ("Here is a dinosaur. How neat! Look at this dinosaur.") and we asked "Can you find another one?" at test.

6.3. Results and Discussion

We replicated the findings of Experiment 3: Children chose the more informative word meaning more often than chance (3–4 year olds, M=63%, SD=29%; 4–5 year olds, M=69%, SD=27%). On the other hand, in the disambiguation

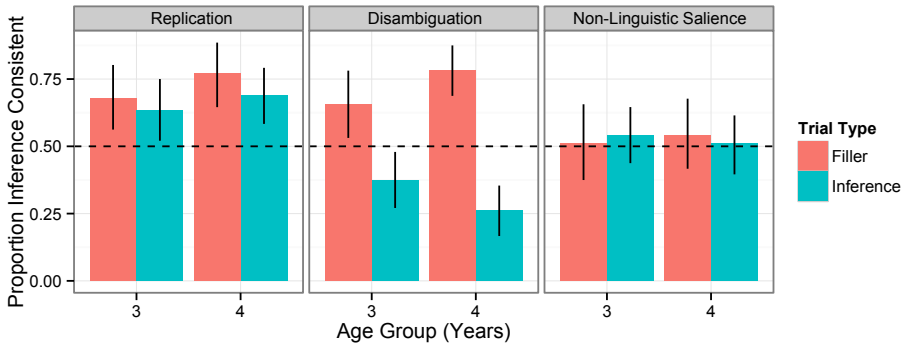

Figure 4: Data from Experiment 4. Mean proportion implicature consistent responding (and mean proportion correct for fillers) is plotted by age groups. The three panels show data from the replication of Experiment 3, the disambiguation condition, and the salience condition. The dashed line shows chance performance. Error bars show 95% confidence intervals, computed via a non-parametric bootstrap over participant means.

condition, participants' performance on inference trials flipped (3–4 year olds, M=37%, SD=26%; 4–5 year olds, M=27%, SD=25%), indicating that they were choosing the feature that had *not* been informative at training, and hence that they had encoded the link between the feature and the novel label used during the initial naming event. In the Non-Linguistic Salience condition, participants chose at chance (3–4 year olds, M=54%, SD=25%; 4–5 year olds, M=51%, SD=28%), indicating that the salience of the unique feature alone was insufficient to drive children's choices.

We fit a mixed effects logistic regression across all three conditions. For ease of interpretation, we fit the model for only inference trials. A model that included condition by age interactions did not significantly increase fit ($\chi^2(2) = 2.18$, $p = .34$), so we did not include an interaction in the final model. We set the three-year-olds' performance in the Replication condition as the reference level—performance in this condition was significantly above chance ($\beta = .78$, $z = 3.81$, $p = .0001$). Performance was reliably different from the Replication condition (and reliably below chance when set to the reference level) in the

Disambiguation condition ($\beta = -1.54$, $z = -6.24$, $p < .0001$). Performance in the Non-Linguistic Salience condition was reliably different from Replication condition performance ($\beta = -.61$, $z = -2.54$, $p = .01$). Finally, there were no reliable age effects, as would be expected averaging across conditions ($\beta = -.12$, $z = 0.19$, $p = .53$); the model including the interaction of age and condition also did not yield any reliable age effects. The results of Experiment 4 thus support the hypothesis that children noticed the differential informativeness of the unique feature and linked this feature to the particular word they heard.

7. General Discussion

We began by revisiting a fundamental question in language acquisition: How do children infer the meanings of words in ambiguous situations? Although a variety of partial answers to this problem have been identified by prior research, a large class of situations (including some construals of Quine's famous problem) are not addressed by these. We have argued here that in some of these cases, word meaning can be disambiguated by the combination of knowledge of speakers' communicative goals and the assumption that they are using language informatively to achieve those goals (Grice, 1975). Our contributions here are then to formalize this inference using a model of pragmatic reasoning and to show that adults and children are able to use contextual informativeness in simple situations to infer word meanings.

In our prior work, we described a framework for pragmatic inference that can be used to make predictions about the behavior of speakers and listeners in simple reference games (Frank & Goodman, 2012). As we showed here, this framework can also be extended straightforwardly to make predictions about what novel words should mean, given that they are uttered by an informative speaker. We then tested predictions from this framework. In Experiment 1, we showed that the aggregate judgments of adult learners conformed quantitatively to the predictions of the pragmatic computation we described. In Experiments 2 and 3, we provided evidence that preschoolers' also made use of contextual

informativeness in word learning inferences. In Experiment 4, we ruled out alternative explanations of these findings in terms of general salience. Together these data suggest that adults and children can use Gricean considerations to infer word meaning in otherwise ambiguous situations.

Our work builds on a long tradition of considering pragmatic reasoning in early language learning (Clark, 2003; Bloom, 2002; Tomasello, 2003). Some of its closest antecedents come from early work on the role of pragmatics in young children's language, where Greenfield, Bates, and their colleagues developed the notion of informative use in context (Bates, 1976; Greenfield & Smith, 1976; Greenfield, 1978). These authors were interested in how children chose which aspects of the world to label using their early language. They posited that children chose the most informative element of a situation and encoded it in speech.

Yet many of the ideas in this work did not see extensive further development. In a critique of this work, Pea (1979) noted that

> the term 'informativeness' is defined in loose *pragmatic* terms... yet no pragmatic theory of information, with the intricacies which would be required in incorporating the belief-states of [speakers] A and B and their changes over time, has ever been developed. ... So the allusion to a formal pragmatic information theory is based on an illusion. (p. 406–407)

Pea's comment highlights a key weakness of these early approaches: they had no formal framework in which to ground observations about pragmatics. Our work here revisits the same set of questions posed in this earlier work (although from the perspective here of language learning as well as language use): how can we formalize powerful Gricean notions of informativeness in context such that it can be applied to make quantitative predictions? We believe that the use of formal models points the way forward for further investigations of children's pragmatic abilities in early learning.

Our data leave open the question of the psychological mechanisms by which

adults and children compute the informativeness of a word in context. We note two particular issues here. First, we cannot differentiate between the case in which each participants' judgments are slightly affected by aspects of the contexts and the case in which some participants notice the informativeness of a feature and others do not. This is a general issue in translating computational-level models of human cognition to the psychological process level (Frank, 2013).

Second, and more specific to the particular domain at hand, it may be that the relative infrequency (uniqueness) of the most informative feature draws attention to it. We have not attempted to differentiate this psychological explanation experimentally because, to a first approximation, unique features, objects, and events are in fact more likely to be referred to. Thus, the same mechanisms that draw our attention to the unexpected, surprising, and rare may be those that help us decide what is informative to talk about. Experiment 4 rules out the explanation that non-linguistic salience or attention alone could explain children's choices in our target condition. It is nevertheless consistent with this result that referential salience and joint attention could be major factors in everyday communication, as these factors can be combined with pragmatic inferences to yield good predictions about reference judgments (Frank & Goodman, 2012).

More generally, the degree to which children (or adults) take others' perspective in judging the novelty of a stimulus is an open question. Experiments on discourse novelty suggest some degree of perspective-taking (Akhtar et al., 1996), but it is controversial even for adults the degree to which others' perspectives are considered in language comprehension (Keysar et al., 2003; Nadig & Sedivy, 2002; Brown-Schmidt et al., 2008). We found no non-linguistic coordination effects in this particular paradigm (indeed, our Non-Linguistic Salience condition was set up to avoid a possible framing that would encourage non-linguistic coordination). But more generally, non-linguistic coordination is an important phenomenon that suggests that "salience" as discussed here is a construct requiring much more carefully investigation (Schelling, 1980; Clark et al., 1983) Thus, this topic will be a fruitful direction for future work.

We proposed a simple model of word learning through informativeness here. A major strength of this model is that it provides a precise, parameter-free fit to adult judgments. But extending this model to more complex stimuli and scenarios will require substantial work. Smith et al. (2013) presented an extension of the cross-situational word learning model of Frank et al. (2009b) that included a pragmatic computation of the type described here. That model was able to make a variety of pragmatically-motivated inferences in single- and multi-trial word learning scenarios, suggesting a possible unification between single-trial "pragmatic" inferences (described here) and multi-trial "cross situational" inferences (Yu & Smith, 2007; Frank et al., 2009b).

Nevertheless, the encoding of context in that model remains as schematic as the one presented here, leaving richer representations of context as another challenge for future work. In more complex environments we expect that performance (especially children's performance) would suffer. Toward the goal of understanding this prediction, Vogel et al. (2014) presented a model that relied on a neural-network representation of pragmatic reasoning and showed more graded generalization across contexts. The focus of that work was on understanding the depth of participants' pragmatic reasoning (how deeply they reason about others' beliefs), but the same model might provide a platform for understanding how pragmatic computations would scale to larger or more naturalistic scenarios.

Children can make many partial solutions to the Quinian 1960 puzzle of ambiguity, employing strategies from cross-situational observation to disambiguation with prior linguistic knowledge, and such strategies can be very helpful. Yet there are still many examples where they fail, including the cases studied here. We have argued that cases where other strategies fail may still be disambiguated by considering the speaker's pragmatic goals. In fact, as we have argued elsewhere (Frank et al., 2009b), this consideration of the speaker's communicative goals may form a broader strategy for language acquisition, accounting for other phenomena as a byproduct of statistical inference over social representations.

References

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

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

Barner, D., & Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive Psychology*, *60*, 40–62.

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in childrens pragmatic inference. *Cognition*, *118*, 84.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Bates, E. A. (1976). *Language and context*. New York, NY: Academic Press.

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

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

Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, *107*, 1122–1134.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5.

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

Chierchia, G., Crain, S., Guasti, M., Gualmini, A., & Meroni, L. (2001). The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. In *Proceedings of the Boston University Conference on Language Development* (pp. 157–168).

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.

Clark, E., & Wong, A. D. W. (2002). Pragmatic directions about language use: Offers of words and relations. *Language in Society*, *31*, 181–212.

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

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground at the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, *22*, 245–258.

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

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE*, *8*, e57410.

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

Fenson, L., Dale, P., Reznick, J., Bates, E., Thal, D., Pethick, S., Tomasello, M., Mervis, C., & Stiles, J. (1994). Variability in early communicative development. *Monographs of the society for research in child development*, *59*.

Frank, M. C. (2013). Throwing out the bayesian baby with the optimal bathwater: Response to endress (2013). *Cognition*, *128*, 417–423.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*, 107–125.

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

Frank, M. C., Goodman, N. D., Lai, P., & Tenenbaum, J. B. (2009a). Informative communication in word production and word learning. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

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

Franke, M. (2009). *Signal to act: Game theory in pragmatics*. Ph.D. thesis Universiteit van Amsterdam.

Franke, M. (2013). Game theoretic pragmatics. *Philosophy Compass*, *8*, 269–284.

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*, 3–55.

Gleitman, L. R., & Gleitman, H. (1992). A picture is worth a thousand words, but that's the problem: The role of syntax in vocabulary acquisition. *Current Directions in Psychological Science*, *1*, 31–35.

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

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*, 173–184.

Greenfield, P. M. (1978). Informativeness, presupposition, and semantic choice in single-word utterances. In N. Waterson, & C. Snow (Eds.), *Development of Communication*. London, UK: Wiley.

Greenfield, P. M., & Smith, J. H. (1976). *The structure of communication in early language development*. New York, NY: Academic Press.

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

Gualmini, A., Crain, S., Meroni, L., Chierchia, G., & Guasti, M. (2001). At the semantics/pragmatics interface in child language. In *Proceedings of SALT XI* (pp. 231–247). Ithaca, NY: Cornell University Press.

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*, 667.

Hudson-Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*, 151–195.

Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Jaeger, T. F., Graff, P., Croft, W., & Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology*, *15*, 281–320.

Jäger, G. (2010). Game-theoretical pragmatics. In J. van Benthem, & A. ter Meulen (Eds.), *Handbook of Logic and Language* (pp. 467–491). Amsterdam, Netherlands: Elsevier. (2nd ed.).

Katsos, N., & Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, *120*, 67–81.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*, 25 – 41.

Kurumada, C. (2013). Contextual inferences over speakers pragmatic intentions: Preschoolers comprehension of contrastive prosody. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press.

Lewis, M., & Frank, M. C. (2013). Modeling disambiguation in word learning via multiple probabilistic constraints. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*.

Liszkowski, U., Carpenter, M., & Tomasello, M. (2008). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition*, *108*, 732 – 739.

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

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

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

Matthews, D., Butcher, J., Lieven, E., & Tomasello, M. (2012). Two- and four-year-olds learn to adapt referring expressions to context: Effects of distracters

and feedback on referential communication. *Topics in Cognitive Science*, *4*, 184–210.

Matthews, D., Lieven, E., & Tomasello, M. (2007). How toddlers and preschoolers learn to uniquely identify referents for others: A training study. *Child Development*, *78*, 1744–1759.

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

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

Navarro, D. J., & Perfors, A. F. (2009). Similarity, Bayesian inference and the central limit theorem. *Acta Psychologica*, *133*, 256–268.

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

O'Neill, D. K., & Topolevec, J. C. (2001). Two-year-old children's sensitivity to the referential (in) efficacy of their own pointing gestures. *Journal of Child Language*, *28*, 1.

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

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

Pea, R. D. (1979). Can information theory explain early word choice. *Journal of Child Language*, *6*, 397–410.

Pinker, S., & Jackendoff, R. (2005). The faculty of language: what's special about it? *Cognition*, *95*, 201–236.

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

Schelling, T. C. (1980). *The strategy of conflict*. Cambridge, MA: Harvard University Press.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Smith, N. J., Goodman, N., & Frank, M. (2013). Learning and using language via recursive pragmatic reasoning about other agents. In *Advances in Neural Information Processing Systems* (pp. 3039–3047). volume 26.

Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*, *16*, 907–912.

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

Stiller, A., Goodman, N. D., & Frank, M. C. (2014). Ad-hoc implicature in preschool children. *Language Learning and Development*, .

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

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Tomasello, M. (2008). *Origins of human communication*. Cambridge, MA: MIT press.

Vogel, A., Emilsson, A. G., Frank, M. C., Jurafsky, D., & Potts, C. (2014). Learning to reason pragmatically with cognitive limitations. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & psychophysics*, *63*, 1293–1313.

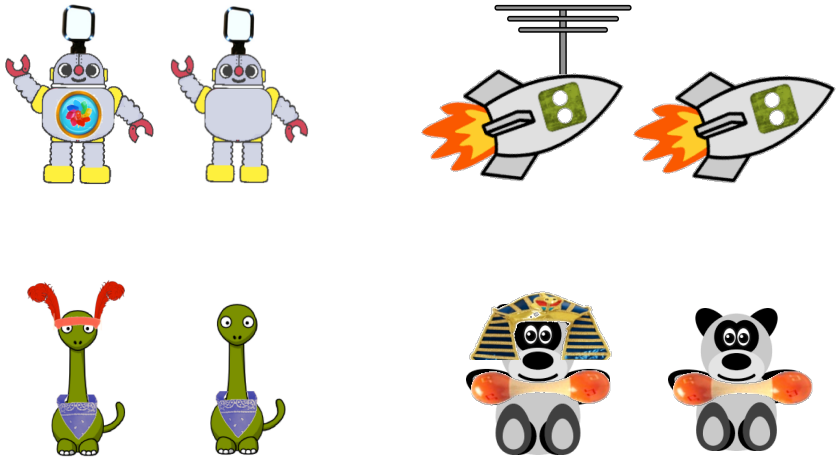

Figure A.1: One version of all four stimuli for target trials in Experiments 2 and 3.

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

Xu, F., & Tenenbaum, J. (2007). Word Learning as Bayesian Inference. *Psychological Review*, *114*, 245.

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

Appendix A. Materials

Stimulus items for all four target items in each of the experiments are shown in Figure A.1. For each stimulus item (robots, dinosaurs, rockets, and bears), there were two possible features: robots had an antenna and a screen, dinosaurs had a bandanna and a headband, rockets had an antenna and a window, and bears had a club and a headdress.