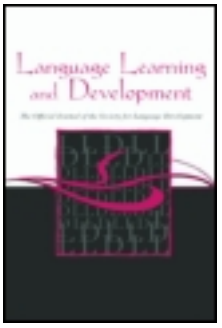

Language Learning and Development
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hlld20

Social and Discourse Contributions to the Determination of Reference in Cross-Situational Word Learning

Michael C. Frank [a], Joshua B. Tenenbaum [b] & Anne Fernald [a]
[a] Department of Psychology, Stanford University
[b] Department of Brain and Cognitive Sciences, MIT

Version of record first published: 21 Nov 2012.

To cite this article: Michael C. Frank , Joshua B. Tenenbaum & Anne Fernald (2012): Social and Discourse Contributions to the Determination of Reference in Cross-Situational Word Learning, Language Learning and Development, DOI:10.1080/15475441.2012.707101

To link to this article: http://dx.doi.org/10.1080/15475441.2012.707101

Language Learning and Development, 00: 1–24, 2012

ISSN: 1547-5441 print / 1547-3341 online
DOI: 10.1080/15475441.2012.707101

Psychology Press
Taylor & Francis Group

Social and Discourse Contributions to the Determination of Reference in Cross-Situational Word Learning

Michael C. Frank
Department of Psychology, Stanford University

Joshua B. Tenenbaum
Department of Brain and Cognitive Sciences, MIT

Anne Fernald
Department of Psychology, Stanford University

How do children infer the meanings of their first words? Even in infant-directed speech, object nouns are often used in complex contexts with many possible referents and in sentences with many other words. Previous work has argued that children can learn word meanings via cross-situational observation of correlations between words and their referents. While cross-situational associations can sometimes be informative, social cues to what a speaker is talking about can provide a powerful shortcut to word meaning. The current study takes steps toward quantifying the informativeness of cues that signal speakers' chosen referent, including their eye-gaze, the position of their hands, and the referents of their previous utterances. We present results based on a hand-annotated corpus of 24 videos of child-caregiver play sessions with children from 6 to 18 months old, which we make available to researchers interested in similar issues. Our analyses suggest that although they can be more useful than cross-situational information in some contexts, social and discourse information must also be combined probabilistically to be effective in determining reference.

INTRODUCTION

Imagine attending a dinner party where you do not speak the language. Most of the time you will likely have trouble understanding any aspect of the conversation. Of course, if you do not understand what is being talked about, you will also have a hard time guessing the meanings of new words. In the flood of new sounds, many of which will be functors like "of," or "it" or even bound morphemes with no individual meaning of their own, picking out consistent associations

Thanks to Maeve Cullinane and Allison Kraus for their work in corpus annotation. This work was supported by a Jacob Javits Graduate Fellowship to the first author and NSF DDRIG #0746251. An earlier version of portions of this work was presented to the Cognitive Science Society in Frank, Goodman, Tenenbaum, and Fernald (2009).

Correspondence should be addressed to Michael C. Frank, Department of Psychology, Stanford University, 450 Serra Mall (Jordan Hall), Stanford, CA, 94305, tel: (650) 724-4003. E-mail: mcfrank@stanford.edu

between sound sequences and their meanings—abstract topics like "the upcoming elections in Germany"—will be difficult at best.

There may be some opportunities when you can guess the topic of conversation, however. If a guest indicates her dinner plate as she makes a comment to you, you might infer that the topic is the food at the party, or perhaps even that one of the words she used means "trout" (which you are both currently eating). Her prosody may even give away her enthusiasm; combined with your knowledge of the etiquette of dinner parties, you may be able to infer that she is giving a compliment. This physically grounded utterance, when paired with its clear prosodic structure and the direct social cue to its referent, now presents an important learning opportunity. If this opportunity is supported by a consistent pattern of co-occurrence between the word and its referent, you may be able to map "trout" to trout and retain this mapping for future use.

While there are many differences between first- and second-language learning, there are nevertheless important parallels between this example and the problem of word learning for young children. If children are engaged in a joint activity or even a moment of joint attention (as in our example), they can use this information to make inferences about the speaker's referential intentions and hence the meanings of words.

Theoretical accounts of early word learning emphasize the role of sharing attention through social cues to joint attention (St. Augustine, 397/1963; Bloom, 2002; Clark, 2003), and a wide variety of empirical evidence supports the view that children use signals like the eye-gaze of speakers to infer what the speaker is talking about (Baldwin, 1993; M. Carpenter, Nagell, & Tomasello, 1998; Hollich, Hirsh-Pasek, & Golinkoff, 2000). Our recent computational work has elaborated this idea—that inferring the intentions of a speaker can give a sophisticated word learner leverage in figuring out the meanings of the words the speaker uses (Frank, Goodman, & Tenenbaum, 2009). In addition, a wide variety of work has attempted to characterize the nature of caregiver-child interactions and their links to language development (Bruner, 1975; M. Carpenter et al., 1998; Stern, 2002). However, there has been comparatively little work on the micro-structure of referential cues: which particular cues matter to determining reference in an individual social interaction.

Going back to our dinner party, a learner who assumes the guest's utterance is about the trout is making use of immediate social information about the speaker's intentions: that pointing is a signal of an intention to refer to some aspect of a particular object. However, another source of information is relevant as well: if a second guest speaks up immediately afterwards, the learner could guess with some certainty that this remark also has to do with the trout (or if not, at least the asparagus or the salad). This kind of aggregation of information across time makes use of the continuity of discourse in conversation. If the second guest's remark had come an hour or even a minute after the first remark, the learner would have had much more uncertainty about the topic.

Speech to children is highly repetitive and includes many partial repetitions of phrases (Snow, 1972); this feature may be important for learning for a variety of reasons. Repetition may allow effective decoding of phonetic material that would otherwise be difficult to decode (Bard & Anderson, 1983); it may also allow for the extraction of structural regularities via minimal pairs in adjacent sentences (Onnis, Waterfall, & Edelman, 2008). Work by Hoff-Ginsberg also has suggested that, for slightly older children at least, repetitions and reformulations are related to question-asking by parents, a feature that positively predicts language outcomes in children months later (Hoff-Ginsberg, 1986, 1990). Finally, this kind of partial repetition and

reformulation has also been argued to give a form of indirect negative evidence (Chouinard & Clark, 2003).

Another way to think about the perceived repetitiveness of child-directed speech is that it creates supportive discourse contexts, in which even partial understanding can nevertheless lead to identification of the topic (as in our example above). General properties of discourse structure have been well-studied in psycholinguistics (P. Carpenter, Miyake, & Just, 1995; Graesser, Millis, & Zwaan, 1997; Wolf & Gibson, 2006), but research on word learning from the perspective of word-meaning mapping has largely neglected the role of discourse. There has been some investigation of the role of the given/new distinction for learning, for example, Akthar, Carpenter, and Tomasello (1996) and Guerriero, Oshimo-Takane, and Kuriyama (2006), but this discussion has not translated into widespread appreciation for the role that discourse continuity might play in early word learning.

One measure of the relative under-appreciation of discourse factors comes from recent computational work. Although many computational models use cross-situational information about the co-occurrence of words and referents for word learning, nearly all of these models assume that utterances are sampled independently from one another with respect to time, throwing away important information about the order of utterances (Siskind, 1996; Yu & Ballard, 2007; Frank, Goodman, & Tenenbaum, 2009).[1]

Recent experimental work has investigated adults' and children's abilities to make cross-situational mappings between words and objects (Yu & Ballard, 2007; L. Smith & Yu, 2008; Vouloumanos, 2008; Vouloumanos & Werker, 2009). These studies have found that learners at all ages are able to learn associations between words and objects based on consistent co-occurrence in individually ambiguous situations. Because these studies have been focused on controlling for extraneous factors, they have largely randomized the order of presentation of word-object pairings in their stimuli, intentionally removing any information about discourse continuity. To date, only one study has focused on the effects of temporal structure on mapping accuracy (Kachergis, Yu, & Shiffrin, 2010), finding that temporal contiguity between instances of a pairing did aid word-object mapping. It seems likely that this effect will be only more pronounced in richer, more naturalistic learning situations.

Our goal in the work reported here is to investigate the utility of social cues and discourse continuity for determining reference in child-directed speech. We conduct this investigation from the perspective of an ideal observer (Marr, 1982; Geisler, 2003): we hope to quantify the amount of information that can be brought to bear on the problem of early word learning on the basis of both explicit social cues to intention and the continuity of discourse via their contribution to determining intentions. This approach allows us to understand the structure of the environment in which early word learning proceeds and to quantify the relative utility of different information sources for the learner (Yu & Ballard, 2007; Frank, Goodman, & Tenenbaum, 2009; K. Smith, Smith, & Blythe, in press). Although the approach does not itself make claims about the use of

[1] An important exception to this trend comes from a model by Roy and Pentland (2002), who used a recurrence filter to take into account temporal contiguity in evidence for word-object mappings. Although this work justified its choice in terms of the dynamics of short-term memory, rather than the structure of discourse, it raises the interesting possibility that memory mechanisms may explain some aspects of the temporal dynamics of word learning (Frank, Goldwater, Griffiths, & Tenenbaum, 2010).

these information sources by human learners, such an analysis can be used to inform empirical studies of word learning.

Although a detailed, realistic model of discourse might contain abstract topics such as "the quality of the food served in the main course," in the current work we consider a simplified version of talking about the same topic that may be more appropriate for young children: talking about the same object. Although this approach almost certainly omits a good deal of abstract information about the kind of activity or action that the child and caregiver are jointly involved in, it is more likely to include the kind of information available to even the youngest word learner. In addition, it is easy to operationalize and does not require the development of a coding scheme for actions or intentions that goes beyond the level of object categories. Finally, even for adults in a situation like our dinner party, continuity of reference may be a more powerful cue than continuity in topic. Thus, in this initial descriptive work, we focus on reference continuity and use the terms "continuity of reference" and "discourse" interchangeably.

Our current study follows work by Yu and Ballard (2007), who used an associative model of word learning to integrate social and prosodic information with information provided by the cross-situational co-occurrence of words and their referents. They investigated these variables in a small hand-annotated corpus of videos from CHILDES (MacWhinney, 2000) and found that performance was improved by the addition of both social and prosodic information. Their work provides inspiration though our study is broader in scope and somewhat different in aim. While they were interested in the improvement in word learning that was brought by integrating social and prosodic cues, here we make a direct attempt to characterize the structure of various social cues and their potential contributions to determining the speaker's intended referent.[2]

The method for our investigation is a corpus study. The approach of coding the information available from videotapes of caregiver-child interaction allows us to analyze the learning environment directly, facilitating our ideal observer approach. A limitation of this type of study, however, is that it does not measure what information learners are able to extract from a particular learning environment. There are many possible reasons why a particular source of information might not be exploited, including learners' biases or even basic cognitive limitations on memory and attention. But our hope is that by pursing this ideal observer approach, the measurements we conduct will motivate future work on the abilities of children in comparable learning situations.

The plan of the article is as follows. We first introduce the corpus we studied. We next discuss the reliability of cross-situational statistics in this corpus, as a motivation for our future analyses. We then discuss social cues and discourse continuity as sources of information about speakers' intentions. We conclude by using a supervised classifier to investigate how much information about speakers' referential intentions can jointly be extracted from these information sources with a simple model of cue combination.

[2]An influential body of work has suggested the utility of words as cues to other words' meanings (Gleitman, 1990; Fisher, 1994; Gillette, Gleitman, Gleitman, & Lederer, 1999). Our focus was on the beginnings of lexical acquisition, rather than the process of learning once some initial words are already known, so we chose to focus on social and discourse information, since this information is likely available to young learners prior to information about linguistic context.

CORPUS MATERIALS

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

We selected a corpus which fulfilled these requirements: a set of videos of object-centered play between mothers and children in their homes, collected by Fernald and Morikawa (1993). Although the original study considered videos of American and Japanese mothers, in the current study we only made use of the American data. The children in these videos fell into three age groups: 6 months (N = 8, 4 males), 11–14 months (N = 8, 5 males), and 18–20 months (N = 8, 4 males). All families were Caucasian.

The corpus was collected by a pair of female observers who made visits to the homes of participants and audio- and video-recorded mother-child dyads as they played. After an introductory period, sets of standardized toy pairs were introduced, including a stuffed dog and pig, a wooden car and truck, and a brush and a box. The mother was given each pair of toys for three to five minutes and asked to play "as she normally would." Although all three toy pairs were given to all dyads, several dyads also played with other toys that were present in the home. Toward the end of the session, to elicit multiple productions of each of the object names, the experimenter asked the mother to hide several of the objects and request that the child search for them. Although the original study made use of only five minutes of data from each video (due to the particular aims of the study), we coded all available data on play centered around pairs of objects. Descriptive data for the corpus are given in Table 1. Participant codes are included so that readers can reference the raw data.

For the purposes of the current study, we made the decision not use data from the hiding game at the end of each video. We made this decision for several reasons. First, and most importantly, the use of referential cues was quite different in this context. Although these differences might potentially be of interest, the game was usually quite short (hence there was not enough data to analyze separately). Second, one of our goals was to study the efficacy of social cues in a relatively restrictive context. The hiding game is an example of a case where social cues have a complex relationship to reference, one that you can only appreciate if you understand the nature of the shared task (to find something that cannot be seen). Accordingly, we believe our basic analyses of social cues are inappropriate in this kind of context. Finally, the hiding game was not performed for the six-month-olds, so its inclusion would compromise comparisons across age groups.

Coding

For each utterance we first coded the toys present in the field of view of the learner at the time of the utterance. Over the course of the video, the union of these sets of toys form the total set of possible object referents for our analyses.[3] A sample frame from the videos is shown in Figure 1.

[3]The assumption not to include objects such as the child's shoes, the rug, the furniture, etc., is of course a simplifying assumption. But all of the other objects in the scene (aside from the experimentally-manipulated toy sets) generally

TABLE 1
Descriptive statistics for each file in the FM corpus. Obj = object, utt = utterance.

Age Grp	*Code #*	*Gend*	*Age*	*Utts*	*Length*	*Obj types*	*Objs/utt*	*Word tokens*	*Word tokens/utt*
6mos	31	M	6	238	14:48	10	1.26	912	3.83
	32	F	6	142	12:08	9	1.56	713	5.02
	33	M	6	257	11:32	12	2.28	974	3.79
	35	F	6	224	14:51	9	1.95	1232	5.50
	36	F	6	109	5:41	14	1.27	396	3.63
	38	M	6	85	7:32	6	2.28	315	3.71
	39	F	6	158	11:03	5	1.85	722	4.57
	40	M	6	244	15:28	8	2.66	845	3.46
12mos	28	M	11	296	14:49	7	2.30	949	3.21
	2	M	12	288	17:06	7	1.84	1252	4.35
	3	M	12	336	21:29	8	2.23	1279	3.81
	4	M	12	154	11:18	6	2.93	476	3.09
	8	F	12	145	13:16	21	1.93	572	3.94
	12	F	14	56	2:53	3	1.18	180	3.21
	14	M	14	65	4:14	9	1.25	216	3.32
	16	F	14	155	8:37	5	1.25	660	4.26
18mos	17	F	18	197	10:02	4	2.00	746	3.79
	18	M	18	232	10:19	5	1.95	801	3.45
	26	F	18	189	12:18	4	1.80	704	3.72
	29	M	18	178	10:14	5	1.60	646	3.63
	22	M	19	120	11:59	17	1.94	427	3.56
	19	F	20	397	12:40	4	2.00	1339	3.37
	20	F	20	266	15:31	9	1.91	1075	4.04
	21	M	20	232	22:00	9	1.57	1030	4.44

The only toy judged to be in the field of view of the child at the time of the utterance most proximate to this frame was the dog. We also coded, for each utterance, the object or objects in the context that were being looked at, held, and pointed to by the mother. These cues were sparse: in many cases, no object was being looked at, held, or pointed to and so these fields were marked "none." This method of coding was chosen because it was practical for the large amount of video data we were working with (a total of approximately five hours of video).[4]

One potential downside of this coding method is that it does not make use of the temporal coordination between, for example, eye-gaze and language production (Griffin & Bock, 2000). The use of eye-tracking during natural interaction is outside of the scope of the current study and may prove difficult more generally (but c.f. Merin, Young, Ozonoff, & Rogers, 2007; Gredebäck,

remain present throughout the entire videos. That does not mean that they cannot be discourse referents—it simply means that they must be pointed out explicitly and introduced into the discourse. In the current study, we limited our coding of objects to those objects that actually were referents in the discourses we coded.

[4]Note that coders had access to the audio at the same time as they annotated social cues. While this meant that they could conceivably be biased towards increasing the accuracy of cues, it would have been difficult for them to synchronize their codes with individual utterances otherwise. Though a technical solution could be found for this issue, given the results reported below we do not believe that bias in favor of cue accuracy was likely to be a major issue. In addition, any bias would apply across cues (allowing us to interpret the relative differences between cues).

FIGURE 1 A sample frame from the FM corpus. (Color figure available online.)

Fikke, & Melinder, 2010; Franchak, Kretch, Soska, Babcock, & Adolph, 2011). In addition, a child observing a caregiver's eyes during natural interaction may be only slightly more accurate in identifying the object of their eye-gaze than an observer who has multiple opportunities to code the same gaze from video. Nevertheless, if technical methods are developed that allow for the automated collection of this kind of data, such data would enable comparisons with the current dataset. Collecting this type of data would be a valuable goal for future work.

Though the data arguably are not a part of the same ideal observer analysis, we also coded two other cues: the object or objects that were being looked at or held by the child. We refer to these information sources as "attentional cues" in the sense that they are information sources that can help us determine reference. In this initial exploratory analysis, we treat them similarly to social cues produced by the mothers in our study. Although this comparison may make attentional and social factors appear superficially more similar than they actually are, we believe it is important to assess the utility of these attentional factors (an issue to which we return in the General Discussion).

We next coded the speaker's *intended referent* for each utterance. We coded an utterance as referring to an object when the utterance contained either the name of the object or a pronoun referring to that object. For example, in the sentence "look at the doggie," the intended referent would clearly be the dog. Likewise, in the utterance "look at his eyes and ears" (where the caregiver was pointing at the dog), the intended referent would also be the dog—though the coder would need to make reference to the videotape to determine the pronoun's referent. We did not specifically mark the use of property terms such as "red," super-/subordinate terms such as "animal" or "poodle," or part terms such as "eye," instead simply coding their object referent in the current discourse. Exclamations such as "oh" were not judged to be referential, even if they were directed at an object. Objects that were not present were still judged to be intended referents, for example, "do you like the doggie" would still be judged to have the referent dog even if the child could not see a dog or a dog was not present in the scene at all.

TABLE 2
Values of Cohen's κ for coding of corpus features

Cue	κ
Intended referent	.83
Mother's eyes	.47
Mother's hands	.80
Mother's points	.77
Child's eyes	.55
Child's hands	.83

In order to evaluate the reliability of our hand-coding scheme, a second coder produced independent annotations for two representative videos. We then calculated a single value of Cohen's κ (a measure of reliability in an n-alternative decision that corrects for chance guessing of frequent options, see Table 2 for data). Since coders were free to assign multiple objects to each coded category, we assumed that utterances for which multiple objects were indicated for a particular category contained multiple opportunities for agreement. This assumption gave the opportunity for "partial credit" (e.g. if one coder assumed that both the dog and the pig were being looked at by the mother while the other assumed that only the dog was being looked at). While reliabilities were high (around .8) for the objects being referred to, mother's hands, points, and child's points, they were considerably lower (in the range of .5) for ratings of the objects looked at by the mother and child. (There may be many causes of this lower reliability, from the relative difficulty of assessing eye-gaze from video of this type to the issue of how to code the temporal coordination of eye-gaze and speech). We consider how the lower reliabilities might affect our analysis in subsequent sections.

All in all, the end product of this coding effort was a corpus of approximately 5,000 utterances and 18,000 words, for which each utterance was annotated with the objects present in the field of view of the learner, the intended referent(s) of the speaker, and the social and attentional cues given by the mother and child. The annotated transcripts for this corpus are available at: http://langcog.stanford.edu/materials/FMcorpus.html. A short excerpt from the corpus is given in Appendix A.

CROSS-SITUATIONAL ASSOCIATIONS

We begin our analysis of the corpus data described above by considering the problem of a word learner attempting to map labels to objects. For all of the objects in the three consistent toy pairings used in the corpus (dog/pig, car/truck, and brush/box), we computed the total probability of a co-occurrence between each object and the corresponding words ("dog" / "doggie", "pig" / "piggie" etc.), both for each participant and for the corpus as a whole. We did this by calculating $P_{co-occur} = C(w)/C(o)$, where $C(w)$ was the count of utterances containing relevant word tokens and $C(o)$ was the count of utterances where the relevant object was present. Results for the entire corpus are shown in Figure 2.

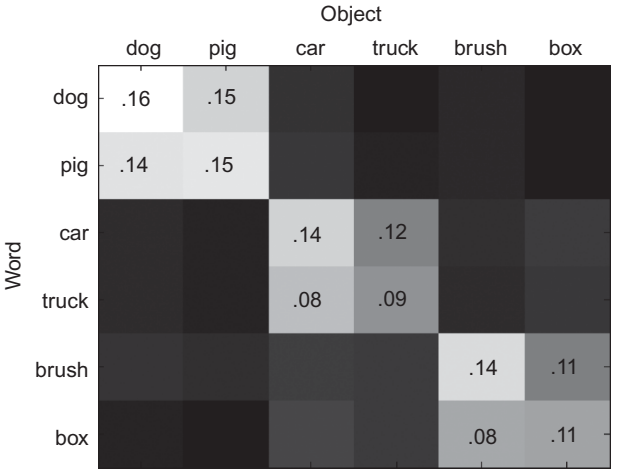

FIGURE 2 A heatmap showing co-occurrence probabilities between words and objects, across the entire corpus. Lighter colored squares indicate higher probabilities (with probabilities greater than .03 marked within each square).

The primary finding of this analysis was that the toy pairs were truly ambiguous in their co-occurrence with the relevant lexical items: "pig" was heard almost exactly as many times with the dog toy as it was with the pig toy. This finding held true across individuals and across toys. There were very slight trends towards greater associations between toys and their correct labels, but they were nowhere near the level that would likely be needed for even the most sensitive statistical learning system.[5]

Because the procedure for data collection for the corpus involved distributing pairs of toys to caregivers, we assumed that there would be some degree of ambiguity in the co-occurrences between words and toys, but we were nevertheless struck by the mismatch between the degree of ambiguity shown in this analysis and the overall impression from the video footage that accompanies our corpus (see e.g. the sample video available along with the corpus transcripts). Although the analysis above shows almost perfect ambiguity between toys in a pair, the impression given from the videos is that very few utterances are truly ambiguous. Instead, the parents in the sample take care to avoid referential ambiguity. This impression is confirmed by the high reliability of coders' estimates of what the mother's intended referent is. Had there been true referential ambiguity in the corpus, agreement would be much lower.

The mismatch between our naïve cross-situational analysis and the human-coded referents form the motivation for the analyses that follow. Using the corpus annotations, we investigate in detail how it is that human coders disambiguate the speaker's intended referent reliably for nearly every sentence.

[5]We return to this issue in the general discussion, when we discuss the results of a cross-situational word learning model that was recently evaluated on this corpus.

SOCIAL AND ATTENTIONAL CUES

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

Signal detection analyses

We began by measuring the utility of each cue in predicting object reference independently. We chose the framework of signal-detection theory as our base for constructing these measures, treating each cue as a predictor to the signal (object reference). Imagine a cue such as the mother looking at objects (referred to as "mother's eyes"). If, for a particular utterance, a look correctly signals the object being talked about, this is counted as a "hit." If an object is talked about but not looked at, this utterance is classified as a "miss." If an object is looked at but not referred to, it is a "false alarm."

From these measures, we calculated two standard scores for summarizing performance. The first was "precision" (hits / hits + false alarms) and the second was "recall" (hits / hits + misses.[6] Precision measures the proportion of the time when the cue was correct, while recall measures the proportion of opportunities for detecting an intended referent when the cue was present. These two measures can be combined into a single number, F_0 (their harmonic mean, which we also refer to as an F-score), for easy comparison.

For example, imagine a case where a mother says "look at the doggie" while looking at the dog, then says "you like the doggie," and looks at the child. In the third utterance, she says "he's so furry," and continues looking at the child. In the fourth, she says "you want to play with something else?" and looks back at the dog. In this hypothetical corpus with only four utterances, one cue and one object, we can demonstrate the use of each of our measures. In the first sentence, looking at the dog is counted as a hit. In the second, looking at the child (but not the dog) is a miss. The third is another miss, and the fourth, where she looks at the dog but doesn't refer to it, is a false alarm. Thus, the precision is 1 hit / (1 hit + 1 false alarm) = .50. The recall of the mother's eyes as a cue to reference in this case is 1 hit / (1 hit + 2 misses) = .33, and F-score is thus approximately .40 (the harmonic mean of .33 and .50). In this example, paying attention to the mother's eyes to figure out her intended referent would not be a good idea.

Results of the signal detection analysis for the broader corpus are plotted in Figure 3 and mean values are given in Table 3. The majority of cues had approximately equal precision and recall (with values centered around .45). Among these, the child's eyes had the best F-score, while the child's hands had the worst. The only major exception to this trend was the mother's pointing, which had a very low recall but high precision. Caregivers' points are relatively few and far between, even in the kind of context that would be most open to ostenstive word teaching. But when these points are present, they are very strong and reliable cues that a particular object is being talked about.

[6]These measures are also referred to as "completeness" and "accuracy."

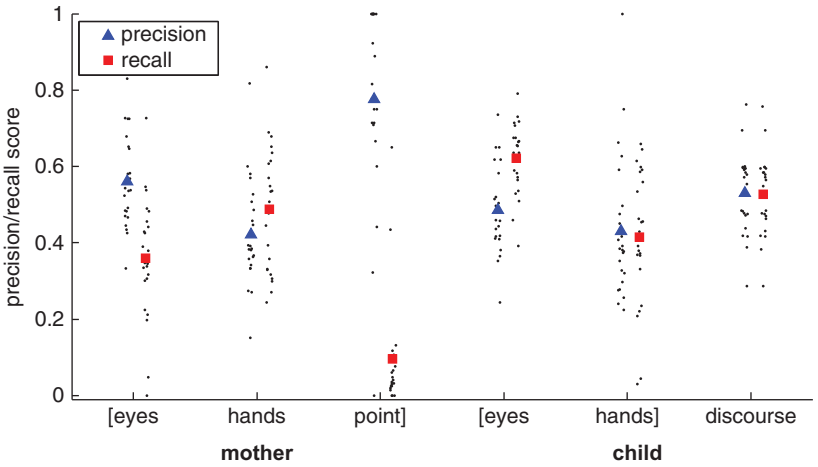

FIGURE 3 Precision and recall for each cue relative to its value in recovering the mother's intended reference in the corresponding utterance. Each dot shows the value for a single dyad, while blue triangles show mean precision and red squares show mean recall. Points are jittered slightly on the horizontal axis to avoid overplotting. (Color figure available online.)

TABLE 3
Precision, recall, and F-scores for all cues

Cue	*Precision*	*Recall*	*F-score*
Mother's eyes	.55	.36	.43
Mother's hands	.42	.49	.44
Mother's points	.78	.10	.14
Child's eyes	.49	.62	.54
Child's hands	.43	.41	.38
Discourse continuity	.53	.52	.53

To test for developmental trends in the F-score of each of these cues, we constructed simple regressions predicting F-score as a function of age. The only predictor that increased significantly with age was the mother's eyes ($\beta = .27$, $p = .005$), with a trend toward a developmental increase in the reliability of the child's eyes as well ($\beta = .46$, $p = .08$).

Looking-related cues for both the mother and the child were relatively good predictors among the group, despite the low inter-coder reliability shown by annotations of this factor. Several interpretations of this result are possible. One is that these cues are even more informative with respect to the speaker's reference, but that they are hard to code from video and hence errors in coding lowered their informativeness in our analysis. A contrasting interpretation is that these cues had low reliability because only some looking behavior is truly meaningful as a signal of reference. On this account, other behavior—scanning the scene, monitoring the other conversational participant—is both difficult to code reliably and relatively uninformative with respect to reference. A third possible explanation is that the difficulty that our coders had in identifying looking

behavior is actually somewhat reflective of the general difficulty of extracting the moment-to-moment location of another person's gaze. While in any given instant it may be easy to determine where someone is looking, it is extremely unusual (and socially aversive) to engage in continuous monitoring of another person's eyes. We believe that this issue is best resolved by future work, perhaps using techniques such as eye-tracking or head-mounted cameras in order to determine the availability of gaze-related information to learners in the moment (Aslin, 2009; Yoshida & Smith, 2008).

Summarizing this analysis, individual social and attentional cues were noisy and did not fully disambiguate the referential ambiguity between toys. Most cues were present often but correct half or less than half the time, while pointing was rare but correct much more of the time. In the current dataset, individual social and attentional cues were not alone sufficient for the determination of reference. To make more accurate guesses, learners must do some kind of extra processing or integration of this information.

Timecourse analyses

The goal of the next analyses was to explore temporal dynamics in the cues we measured. In particular, we were interested in whether some were used more often at the beginning of talking about objects. To perform this analysis, for each age group we aggregated all the examples of discourses—continuous runs of talking about an object. We defined a discourse to be three continuous references to an object, though results did not change qualitatively when we explored other reasonable values for this number. We aligned each of these discourses and averaged the social cues for the object that was being referred to. There were 88, 110, and 107 such discourses for the 6-, 12-, and 18-month-olds, respectively. Since relatively few lasted longer than 5 or 10 utterances, data were too sparse to calculate cue probabilities accurately for longer discourses. Therefore, we excluded lengths for which we had fewer than 10 datapoints. Results are plotted in Figure 4.

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

The major result of this analysis is that points appear to be used to introduce new discourse topics. Although they are infrequent, they are more frequent at the beginning of discourses about objects, at least for young children. A learner who identified a point would thus do well to assume that the object being pointed to is the topic of discourse for the next several utterances topic, if the

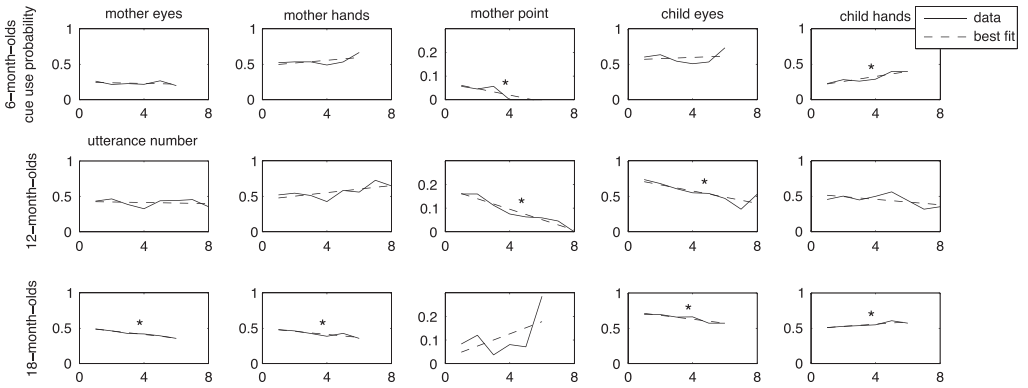

FIGURE 4 Each plot shows the probability of a particular social cue being present, plotted by the number of utterances a particular object had been talked about. Empirical data are shown with a solid line and the best linear fit to these data are shown with a dashed line. Significant linear trends are shown with a star. Each row of plots shows an age group and each column of plots shows a particular social cue.

discourse topic did not shift. In the next set of analyses we investigate this strategy by measuring the dynamics of topic shifting in our sample of child-directed speech.

DISCOURSE INFORMATION

The next goal of our study was to quantify the role of continuity of discourse (here defined as continuity of reference) in predicting to which objects caregivers were referring. In this section, we first develop a visualization of reference in child-directed speech. We next show some descriptive results about the magnitude and temporal dynamics of reference continuity. Finally, we end by comparing reference continuity to the social cues examined above using the same signal detection analyses.

Visualizing continuity of reference

The first step we took towards understanding the prevalence of discourse continuity was to visualize the results of coding the speakers' intended referent. We introduce what we call a "Gleitman plot": a visualization of a stretch of discourse based on (1) what objects are present and (2) what objects are being talked about. We have chosen this name because Gleitman (1990) was concerned with the relationship between what is present in a learner's experience and what is being talked about; we believe our plotting method provides insight into this question.

A representative Gleitman plot for one mother-child dyad in the corpus is shown in Figure 5. Rows show references to individual objects over time, such that an object that is present is shown in blue; one that is talked about is shown in green; and one that is present and talked about is in red. This view of the corpus allows us to examine trends in the timecourse of reference and to

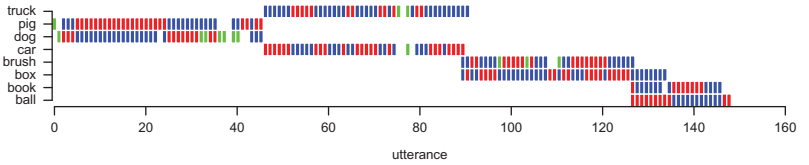

FIGURE 5 Example Gleitman plot. Each row represents an object, each column represents an utterance. A blue mark denotes that the object was present when the utterance was uttered but not mentioned; a green mark denotes that the object was mentioned but not present; and a red mark denotes that the object was present and mentioned. Horizontal stretches of red marks indicate continuous sets of utterances referring to a particular object that was visible to the child. (Color figure available online.)

visualize a complex set of data in a compact form. For example, it shows us at a glance that the corpus interactions were structured around pairs of objects, since the interaction can be divided into sets of twin stripes for the pig/dog, truck/car, and brush/box pairs, as well as a short segment on the book/ball pair.

We can draw two anecdotal conclusions on the basis of viewing the Gleitman plots for each mother-child dyad in the corpus. First, within the corpora we studied, mothers talk primarily about objects that are present in the field of view of the children. This can be seen by examining the small amount of green within the plots. Unsurprisingly, for a word learner guessing the meaning of a novel noun, the best guess will likely be that the word refers to an object that is present (Pinker, 1989; Yu & Smith, 2007; Siskind, 1996). (Although this generalization may be true for nouns, it is much less likely to be true for verbs; see Gleitman, 1990.) Of course, the generality of this conclusion is limited by the restricted task that the mothers in our sample were asked to perform.

Second, we can see clear evidence of discourse continuity (again, defined as continuity of reference). For example, in Figure 5, rather than being distributed evenly throughout the span of time when an object is present, references to an object are "clumpy": they cluster together in bouts of reference to a single object followed by a switch to a different object. This can be seen for example in the dog / pig portion (first 45 utterances), where the mother alternates several times between the two objects, talking about each for several utterances before switching.

Measuring reference continuity

In our visualizations, we observed clumps of references to a particular object rather than a more uniform distribution of references over time. To quantify this trend, we first defined a measure of reference continuity, P_{RC}: the probability of referring to a particular object, given that it was talked about in the previous utterance. We go into some detail about how this measure was calculated in order to be clear about how we calculated our baseline measure, since an appropriate baseline is crucial for determining whether P_{RC} is greater than chance.

For an object o, we defined the reference function $R_t(o)$ as a delta function returning whether or not that object was referred to at time t. We then define $P_{RC}(o)$ (the probability of reference continuity for a particular object):

$$P_{RC}(o) = \frac{\sum_t R_t(o)R_{t-1}(o)}{\sum_t R_t(o)}. \quad (1)$$

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

We then estimated a baseline value for P_{RC} via permutation analysis. Intuitively, this analysis asks what a "chance" value for P_{RC} would be if utterances were completely independent of one another. This analysis is important because the distribution of individual objects is very uneven in time and some objects are more likely to be talked about than others. We calculated this baseline value for each dyad in the corpus by recomputing $P_{RC}(o)$ for 10,000 random permutations of the times at which each object was talked about.[7] For the Gleitman plots in Figure 5, this analysis would be represented by randomly shuffling all the red and blue squares in each row so that the same overall set of squares were red and blue but their ordering was different.

The results of this analysis are shown in Figure 6. As predicted based on our visualizations, P_{RC} was outside of the 95% confidence interval on chance for all but 3 of the 24 dyads. A simple linear regression showed no relationship between P_{RC} and age ($r^2 = 0.01$, $p = .56$). Thus, it appears that reference is considerably more continuous than would be expected by chance in child-directed play situations of the type in our corpus.

Note that our baseline is very dependent on the number of objects that are present. With a mode of two objects present, the baseline calculation is as conservative as possible. The result that discourses are significantly more continuous than expected by chance, even in the extremely restricted experimental situation for our corpus, suggests that in a noisier environment, discourse continuity could be an even more powerful cue. In other words, in the absence of other information about what is being talked about, a good bet for a child is that mom is still talking about the same thing she was a moment ago.

Temporal properties of reference

We next examined the temporal properties of reference: how the recency of mention for an object affects whether it will be talked about again. This analysis can be thought of as a generalization of the analysis above; the new analysis asks about the probability of an object being talked about given that it was referred to some number of utterances ago.

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

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

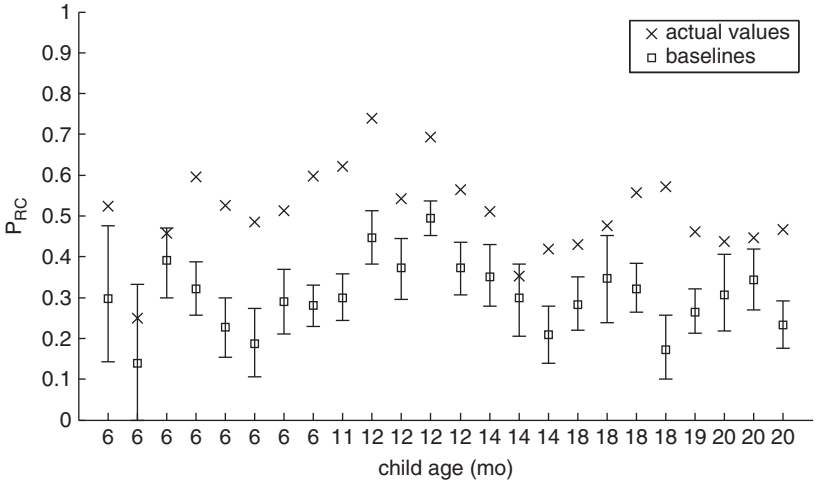

FIGURE 6 Probability of reference continuity (P_{rc}) for each child, shown in age order on the horizontal axis. Box with error bars shows 95% confidence intervals for a permuted baseline.

$$P^t_{RC}(o) = \frac{\sum_t R_t(o) R_{t-n}(o)}{\sum_t R_t(o)}. \tag{2}$$

The result of this analysis are plotted in Figure 7. It is clear from this visualization that very recent utterances are disproportionately correlated with the probability of referring again—this observation summarizes the previous analysis. The influence of a particular object in discourse declines slowly, however.

We quantified this property by fitting two functions to the resulting data: an exponential and a power-law function. Both functions were fit by adjusting two parameters (intercept and decay) in order to minimize mean squared error. We found that the power-law (MSE = .14) fit considerably better than the exponential function (MSE = .75). The key portion of the curve on which the power law gave better fit was the sharp initial decrease from a very high probability of referring again to a much lower one a few utterances later.

This dynamic may be due to the more general phenomenon of power-law decays in human memory (Anderson & Schooler, 1990), or comparable dynamics governing discourse topics (e.g., boredom or shifting attention). Recent work has also found similar distributions for word frequency repeats in much larger corpora, though this work suggested that with more data it was possible to distinguish a power law from a stretched exponential distribution (Altmann, Pierrehumbert, & Motter, 2009). Thus, our conclusions about the specific form of the distribution here are tentative. Regardless, this pattern clearly shows a slow drop-off in the probability of bringing up a previously-mentioned referent, suggesting that a learner who takes this bias towards previously-mentioned references into account will make better guesses about the current referent.

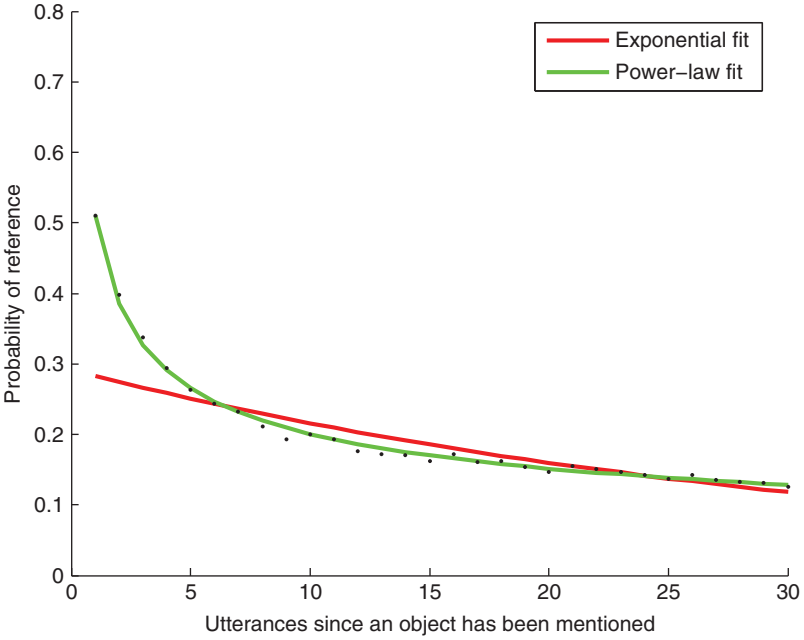

FIGURE 7 Probability of reference given that an object was referred to *n* utterances ago, where *n* is plotted on the horizontal axis. (Color figure available online.)

Signal detection analysis

We conducted the same analysis for discourse continuity as a cue to reference as we did for each social cue. We analyzed the precision, recall, and F-score for the scenario in which the learner guesses that a particular utterance will refer to the same object that the utterance before did. Results are plotted alongside the social cues in Figure 3. We found that discourse continuity had an average F-score comparable to knowing what the child was looking at. In other words, a learner with perfect information about previous referents would do as well guessing the current reference based on continuity as they would based on the most informative social/attentional cue. This analysis confirms the results of the reference continuity analyses above, demonstrating again that discourse information, when available, can be quite useful in guessing speakers' intended referent.

JOINT CLASSIFICATION ANALYSIS

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

To carry out our supervised classification analysis, we used a Naïve Bayes classifier (a standard technique in statistical machine learning; see e.g., Hastie, Tibshirani, & Friedman, 2001) to combine each of the cues in order to make a judgment about what object was being talked about. This classifier has the advantage of being simple and computationally efficient, although it makes the assumption that all cues are conditionally independent from each other. We use a method that makes this assumption rather than a more sophisticated method that naturally exploits mutual information between cues in order to explore whether the information sources in our corpus were in fact truly independent (something which might be hidden in the operation of a more sophisticated classifier).

Our classifier was a standard Naïve Bayes classifier:

$$p(O|C_1, ..., C_n) = \frac{1}{Z} p(O) \prod_i p(C_i|O), \tag{3}$$

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

Because many mother-child dyads played with somewhat different sets of objects (and also to ensure the generality of our results), we constructed a separate classifier for each mother-child dyad. The classifiers were evaluated using a tenfold cross-validation scheme in order to ensure that results were not due to overfitting. Results reported here are averaged across all 10 test sets.[8]

Figure 8 shows the results of this analysis. The baseline probability of reference to an object (calculated as the proportion of utterances with a coded intention that was not "none") was relatively low in all three groups (6-month-olds, .60; 12-month-olds, .52; 18-month-olds, .50). We therefore report classification performance only for those sentences which had an intended referent. We evaluate classifiers created by fully crossing three sources of information: social cues exhibited by the mother, including eyes, hands, and pointing; markers of the child's attention, including the child's eyes and hands; and discourse cues. All classifiers included baseline information about which objects were present in the field of view of the child. We did not find systematic age-related differences in classifier accuracy, so we consolidated data across all 24 dyads.

For all dyads, baseline performance was low, indicating that the physical presence of objects was not enough to predict reference effectively. While mothers often referred to objects that were present, sometimes they did not, and they also sometimes referred to objects that were not present (Gleitman, 1990). Adding social/attentional information (whether social cues from the mother or attentional cues from the child) nearly doubled classifier accuracy. Adding discourse information also resulted in a boost in classification accuracy, though not quite as large as that caused by adding social/attentional information.

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

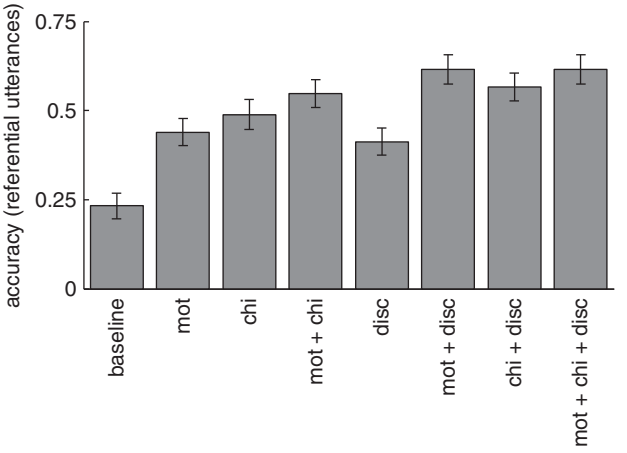

FIGURE 8 Classifier performance on those utterances for which there was an intention to refer to an object by cues used in classification. Error bars show standard error of the mean across all children. "mot" = mother's social cues, "chi" = child's social cues, and "disc" = discourse cues.

Combining any two information sources resulted in an additional boost, but adding the third did not add any additional accuracy. This result suggests that cues were nonindependent: there was overlapping information about reference between the different sets of cues (hence the gain from having both was less than the classifier gain expected by having only one or the other). In the case of the mother/child cue interaction, it seems likely that this overlap is due to cases of joint attention in which both participants are directly focused on a single object. The interaction between the child's attention and discourse continuity is less clear but may suggest that children's attention in this task is "sticky," staying on the current focus of conversation and switching more gradually than the mother's attention. Overall performance with all information sources was 61.5%, suggesting that even with imperfect information, there are many utterances for which social information suffices for the identification of the speaker's intended referent.

Summarizing this analysis, social cues and discourse together represent overlapping sources of information for determining what is being talked about. Taken together, these information sources (as captured by our coding) were far from perfect but nevertheless allowed for relatively good guesses about the topic of an utterance without any additional linguistic information.

GENERAL DISCUSSION

The goal of this study was to measure the contributions of various sources of nonlinguistic information to determining reference—what object a speaker is talking about—in child-directed speech. To address this question we introduced a corpus of videos of child-directed speech across a range of ages, which we annotated with information about the objects visible to the child, the speakers' intended referent, and the various social interactions of the child and caregiver with

the objects. We found that, with the exception of pointing, social cues such as eye-gaze and hand position were at best noisy indicators of reference, and that no individual cue revealed the speaker's referent more than a portion of the time, even in this highly constrained corpus. Discourse continuity (the assumption that the speaker was talking about the same thing as in their previous utterance) provided an additional source of information about what was being talked about that was as reliable as any of these individual cues. A final set of simulations with a supervised classifier suggested that, despite their overlap, aggregating information across these information sources together provided a better (though still imperfect) estimate of the speakers' intended referent.

Limitations

The current study has a number of limitations. Each of these was exposed in the process of conducting this descriptive study, and might not have been obvious without the effort taken to develop a coding scheme for the factors of interest. Although each may limit the strength of the generalizations possible from this particular study, we hope that each also points the way towards future work.

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

Second, we have spoken throughout our analyses as though complex physical gestures can be individuated into discrete "cues" which can easily be associated with a particular utterance. This approximation will almost certainly miss nuances of gestural communication (e.g., anecdotally, caregivers in our sample often moved the object they were holding and talking about more than one that they were not talking about), but this approximation was necessary to code the volume of data reported here. Technical advances such as motion capture or motion recognition from computer vision may provide some traction on these questions (L. Smith, Yu, & Pereira, 2009).

Finally, we have equated an eye-movement by the caregiver (which may or may not be visible to the child) with an eye-movement by the child (which controls what is visible to him or her). From the perspective of the child, this equivalence is not valid: the child's own eye-movements control what is being looked at, while the adult's eye-movements constitute an ephemeral signal to another person's attention. Nevertheless, in order to understand the relative validity of the child's own attention compared with external social information, we believe it is important to include these cues. In fact, the relative informativeness of what the child looks at and touches may provide support for a hypothesized egocentric belief: that words refer to aspects of the child's own perceptual experiences rather than signaling the speaker's referential intentions (Baron-Cohen, Baldwin, & Crowson, 1997; Hollich et al., 2000). Since there is significant cultural variation in the amount that caregivers accommodate their labeling behavior to children, future work on this

topic should sample across a wider range of cultural and socio-economic contexts (e.g., Ochs, 1988; Hart & Risley, 1995; Fernald, 2010).

CONCLUSIONS

In studies of the role of social cognition in language learning, researchers tend to manipulate the presence or absence of social information. When this manipulation results in a significant difference in the child's reasoning about or retention of a word, we assume that a particular cue is a viable source of information or even one that is particularly important for word learning. But these inferences may not always be warranted. Our study takes a first step towards measuring the microstructure of cue presence "in the wild." From this perspective, the single most important insight from this study is that no individual cue would consistently allow an observer of our corpus to infer what the speaker was talking about. Instead, an efficient learner would do far better by combining social information sources and aggregating this information over time—treating speakers' behavior as signals from a noisy source—than they would by monitoring any particular cue.

A simple view of this process is that social cues each could provide a probabilistic filter for possible referents. Since recent work suggests that children are able to use co-occurrence to link words to referents (Yu & Ballard, 2007; L. Smith & Yu, 2008; Vouloumanos, 2008; Vouloumanos & Werker, 2009), this kind of "filtering" account would be a way for learners to integrate cross-situational and social information. Both experimental (Baldwin, 1993; Akhtar et al., 1996) and computational (Yu & Ballard, 2007) studies suggest cross-situational evidence is more effective when supplemented with social information.

Perhaps, however, cross-situational and social learning are not separate proceses at all (Frank, Goodman, Tenenbaum, & Fernald, 2009). Perhaps the kind of "statistical learning" that is seen in cross-situational word learning experiments actually operates over fundamentally social representations. A statistical word learner trying to interpret an utterance may be balancing uncertainty about the speaker's intended referent (or even, intended meaning) with uncertainty about what words themselves mean. This kind of "communicative inference" proposal would suggest that social cues are not a filtering step prior to a cross-situational analysis. Instead, they are a fundamental part of the inference itself.

Our recent work implements such a model, in which the relative weights of social cues are learned jointly along with word-object mappings, and assesses it on the corpus described here (Johnson, Demuth, & Frank, 2012). While word learning performance on the corpus is still far from perfect, the model performs especially well at the utterance-by-utterance process of identifying the speaker's intended referent (the topic of our study here). Most notably, this system learns the probabilities that the various social cues in the corpus are signaling reference, replicating the descriptive analyses reported here in a completely unsupervised learner.

The analysis presented here thus points to a distinction between two different conceptions of early word learning: one in which different sources of information like cross-situational statistics and social cues are independent (a "filtering" model), and one in which they are joint (a "communicative inference" model). Distinguishing these two conceptions will not be easy. But the development of tools for analyzing the microstructure of children's social input will be a critical step along the way.

REFERENCES

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

Altmann, E., Pierrehumbert, J., & Motter, A. (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS One*, *4*(11), e7678.

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

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

Baldwin, D. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language*, *20*, 395–395.

Bard, E., & Anderson, A. (1983). The unintelligibility of speech to children. *Journal of Child Language*, *10*(2), 265–292.

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

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

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

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*(4).

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

Chouinard, M., & Clark, E. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language*, *30*(3), 637–669.

Clark, E. (2003). *First language acquisition*. New York, NY: Cambridge University Press.

Fernald, A. (2010). Getting beyond the convenience sample in research on early cognitive development. *Behavioral and Brain Sciences*, *33*(2–3), 91–92.

Fernald, A., & Morikawa, H. (1993). Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child Development*, *64*, 637–656.

Fisher, C. (1994). Structure and meaning in the verb lexicon: Input for a syntax-aided verb learning procedure. *Language and Cognitive Processes 9*, 473–517.

Franchak, J., Kretch, K., Soska, K., Babcock, J., & Adolph, K. (2011). Head-mounted eye-tracking of infants natural interactions: A new method. *Child Development 82*, 1738–1750.

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

Frank, M. C., Goodman, N. D., Tenenbaum, J. B., & Fernald, A. (2009). Continuity of discourse provides information for word learning. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

Geisler, W. (2003). Ideal observer analysis. *The Visual Neurosciences*, 825–837.

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

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

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

Gredebäck, G., Fikke, L., & Melinder, A. (2010). The development of joint visual attention: A longitudinal study of gaze following during interactions with mothers and strangers. *Developmental Science*, *13*(6), 839–848.

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

Guerriero, A., Oshima-Takane, Y., & Kuriyama, Y. (2006). The development of referential choice in English and Japanese: A discourse-pragmatic perspective. *Journal of Child Language*, *33*(4), 823.

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

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

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

Hoff-Ginsberg, E. (1990). Maternal speech and the child's development of syntax: A further look. *Journal of Child Language*, *17*(1), 85–99.

Hollich, G., Hirsh-Pasek, K., & Golinkoff, R. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, *65*(3), 1–135.

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

Kachergis, G., Yu, C., & Shiffrin, R. (2010). Temporal contiguity in cross-situational statistical learning. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

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

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

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

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

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

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

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

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

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

Smith, L., Yu, C., & Pereira, A. (2009). Not your mother's view: The dynamics of toddler visual experience. *Developmental Science*, *14*, 9–17.

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

St. Augustine. (397/1963). *The confessions of St. Augustine* (R. Warner, Ed.). New York, NY: Clarendon Press.

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

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

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

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

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

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

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

APPENDIX A: CORPUS EXCERPT

We include a brief excerpt from the first session in our corpus (from a 12-month-old, A2). M and C refer to mother and child, respectively. We have removed punctuation for purposes of automated analysis and comparison.

Utterance	*Objects*	*Referent*	*M eye*	*M hand*	*C eye*	*C hand*	*M point*
is that the doggy	dog pig	dog	none	none	pig	none	none
you see the doggy	dog pig	dog	dog	none	pig	none	none
theres the doggy	dog pig	dog	dog	none	dog	pig	none
thats a pig	dog pig	pig	child	none	pig	pig	none
pig	dog pig	pig	child	none	pig	pig	none
is that soft	dog pig	pig	pig	pig	pig	pig	none
is that a puppet pig soft	dog pig	pig	child	pig	pig	pig	none
love the pig	dog pig	pig	pig	pig	dog	none	none
love the pig	dog pig	pig	child	none	pig	pig	none
pigs say oink oink oink	dog pig	pig	pig	pig	pig	pig	none