Markers of topical discourse in child-directed speech

Hannah Rohde

Department of Linguistics and English Language, University of Edinburgh

Michael C. Frank

Department of Psychology, Stanford University

Some portions of this work were first presented as Rohde and Frank (2011). Thanks to Noah Goodman and Eve Clark for helpful discussion and Allison Kraus for valuable assistance in annotation and data coding. An Andrew W. Mellon postdoctoral fellowship to H. Rohde supported this research.

Please address correspondence to Hannah Rohde, `hannah.rohde@ed.ac.uk`.

Abstract

Although the language we encounter is typically embedded in rich discourse contexts, many existing models of processing focus largely on phenomena that occur sentence internally. Similarly, most work on children's language learning does not consider how information can accumulate as a discourse progresses. Research in pragmatics, however, points to ways in which each subsequent utterance provides new opportunities for listeners to infer speaker meaning. Such inferences allow the listener to build up a representation of the speakers' intended topic and more generally to identify relationships, structures, and messages that extend across multiple utterances. We address this issue by analyzing a video corpus of child-caregiver interactions. We use topic continuity as an index of discourse structure, examining how caregivers introduce and discuss objects across utterances. For the analysis, utterances are grouped into topical discourse sequences using three annotation strategies: raw annotations of speakers' referents, the output of a model that groups utterances based on those annotations, and the judgments of human coders. We analyze how the lexical, syntactic, and social properties of caregiver-child interaction change over the course of a sequence of topically-related utterances. Our findings suggest that many cues used to signal topicality in adult discourse are also available in child-directed speech.

Introduction

One of the characteristics of natural discourse is that the presence and ordering of utterances is non-arbitrary: Utterances appear together because they relate to each other in meaningful ways, often via a shared topic. Successful comprehension thus requires not only the interpretation of each of the individual utterances that comprise a discourse but also the inference of what we will call TOPICAL DISCOURSE SEQUENCES, stretches of discourse each consisting of a series of utterances that center around a shared topic. In the simple case, such a sequence consists of a pair of utterances that both transparently refer to the same object. In less explicit cases, a sequence may span a set of observations exchanged on a loose theme. As such, identifying these sequences may be trivial in some cases, while in other cases the natural flow of conversation is organic and difficult to segment.

For language learning, the identification of these topical discourses presents both a challenge and an opportunity. The challenge lies in tracking the ebb and flow of discourse topics and relating the content of one utterance to material in the larger discourse context. Identifying topical discourses almost always requires some inference: Only some of the related utterances in a topical sequence may explicitly mention the shared topic, but the utterances may cohere in other ways. On the other hand, the opportunity derived from inferring topical discourse sequences is that learners can profit from information that is split across multiple utterances. Utterances build on one another and contain unique information that can only be appreciated when they are interpreted together. Similarly, social cues like pointing typically are not used to mark reference for every sentence individually, but instead pick out a referent when it is first introduced in discourse. Without an appreciation of topical discourse, learners risk missing many linguistic and social connections that are available in a conversation.

Yet despite the challenges and opportunities afforded by the presence of topical

discourse sequences in conversation, work in language acquisition has typically treated sentences as independent units rather than integrated components of an unfolding discourse. This focus is reflected in the current expansions of acquisition work into the domain of sentence processing through research on children's comprehension and production abilities (Snedeker & Trueswell, 2004; Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998). The treatment of sentences as independent units echoes the emphasis on sentence-internal phenomena within classic sentence-processing research, whereas analyses of cross-sentence phenomena in adult discourse stem largely from insights in formal pragmatics and computational linguistics (Asher & Lascarides, 2003; Grice, 1975; Grosz, Joshi, & Weinstein, 1995; Hobbs, 1979; Kehler, 2002; Mann & Thompson, 1988; Polanyi, 1988; Genzel & Charniak, 2002).

The goal of this paper is therefore to consider child-directed speech in context, recognizing the increased informativity of spoken language when it is encountered in a rich discourse context. We extend work on the markers of topical discourse in adult speech to the child-directed context, analyzing a video corpus of child-caregiver interactions to identify properties of the topical discourse sequences.[1]

In the acquisition setting, one domain in which the inference of topical discourses may be particularly relevant is in the task of word learning. Word learning requires a child to establish relationships between referring expressions and real-world objects, and in turn infer meanings for the components of these expressions. As many authors have noted,

[1]We distinguish two senses of the term "discourse" in this work. The first sense is the broader one that is most commonly used in the literature, in which "discourse" refers simply to utterances that together constitute a coherent whole and instantiate a set of cross-sentence relationships. The second sense, which we employ throughout this paper, is a reference to a specific *topical discourse sequence*: a group of adjacent utterances united by a shared topic (see the example in Table 1). These sequences—and how they are identified in child-directed speech—are the primary focus of our work here, and can be viewed as an operationalization of the broader notion of "discourse."

Table 1

Artificial (constructed) transcript containing two topical discourse sequences (i.e., two groups of adjacent utterances united by a shared topic: a dog for the first 5 utterances and a pig for the subsequent 3)

Utterance	**Topic**
1. do you see the little dog?	dog
2. look, it's the one right here [pointing to dog]	dog
3. it's wagging its tail	dog
4. did you get licked?	dog
5. look at those long floppy ears	dog
6. now do you see the little pig? [pointing to the pig]	pig
7. the pig has a curly tail	pig
8. he says oink oink	pig

connecting words to referents and inferring their underlying meanings can both be arbitrarily complex tasks (Quine, 1973; Gleitman, 1990). In this light, identifying a word's referent (and perhaps even a particular feature or construal of a referent) can be far easier in a coherent discourse: A sequence of utterances, accompanied by gestures and supportive feedback around a shared communicative goal, can serve to constrain hypotheses about reference.

Consider the example in Table 1. A new word may appear in one utterance without any accompanying gesture to the real-world entity (utterance 1), whereas an adjacent utterance may be accompanied by a gesture but include neither the relevant lexical item (utterance 2) nor the names for both the entity and its feature (utterances 3, 5, 8). However, if a learner assumes that nearby utterances are likely to relate via a shared topic, a link can be posited between the gesture in one utterance and the new lexical item in another or between the name of an entity and the name of a feature of that entity, without requiring that the cues be co-present (as they are in utterances 6 and 7). These links can support powerful inferences about word meaning.

Despite the possible utility of topic continuity as a cue for learning, nearly all models of word learning treat sentences as independent events (Fazly, Alishahi, & Stevenson, 2010; Frank, Goodman, & Tenenbaum, 2009; Siskind, 1996; Yu & Ballard, 2007; Goldwater, Griffiths, & Johnson, 2009). Attempting to address this issue, Frank, Tenenbaum, and Fernald (2013) proposed that by assuming that proximate utterances are more likely to refer to the same objects, early word learners may be better able to aggregate information from social cues and make better guesses about what words mean. Their study provided evidence for the existence of topic continuity in child-directed speech: Caregivers were more likely to talk about objects that they had referred to in the previous sentence. But their analyses made two simplifying assumptions. First, they assumed that discourse topics were directly observed by learners. Second, they reduced the problem of the identification of topical discourses to a local notion of discourse continuity: The topic of one utterance is linked to the utterance immediately previous. The current study addresses both of these limitations, considering information sources that might lead to the discovery of discourse topics and moving beyond local dependencies to consider multi-utterance topical discourses.

The approach we take in our study is to test whether information sources vary with an utterance's position in a topical discourse sequence. If such variation is present, then these information sources can be used by learners as signals for the identification of topical sequences. Thus, the purpose of the current study is not to provide evidence that learners *do* use such cues—evidence for claims of this sort will require experimental evidence from children—but to identify sources of information that they *could* use, in order to motivate future experimental work.

In our study, we consider information relevant to how caregivers refer to objects by analyzing the rate of pronoun use and the syntactic position of particular referring expressions, both of which are known to reflect properties of the larger discourse context.

We also consider utterance complexity—operationalized as utterance length—as a measure of growing common ground. Lastly, we consider social cues such as eye gaze and hand position, both of which serve to index joint attention. Overall, our findings on each of these information sources suggest that many of the cues used to signal topicality in adult discourse are also available in child-directed speech.

Factors relevant to discovering topical discourse sequences

Below we review a set of markers of topic continuity that we will use in our analysis of child-directed speech. These markers have been discussed both in work in the adult literature on discourse processing and in work in acquisition on information structure and informativeness.

Research on how adults track topics across sentences has addressed a variety of questions, including how listeners initially identify referents, how speakers signal shifts in topic, and what inferences are involved in resolving referentially ambiguous expressions. Answers to these questions have highlighted the range of information sources that are brought to bear in coreference processing (Arnold, 2001; Caramazza, Grober, Garvey, & Yates, 1977; Hobbs, 1979; Kaiser, 2012; Kehler, Kertz, Rohde, & Elman, 2008; Koornneef & Van Berkum, 2006; Smyth, 1994; Stevenson, Crawley, & Kleinman, 1994) as well as the coreference conventions speakers adhere to across sentences (H. Clark, 1996; Grosz et al., 1995). Although much of this work has focused on complex, highly-structured discourses, some factors identified in this literature nevertheless can be applied to the simple referential discourses treated here, and we analyze pronoun usage, referent position, and utterance complexity.

In the acquisition context, issues of information structure in children's productions have received extensive treatment. For example, a wide variety of work has discussed the informativeness of children's early productions given the discourse context, especially

focusing on the omission of arguments. Studies suggest that children's early omissions (and their referential choices more generally, during the period when their linguistic abilities are limited) are motivated by the communicative demands of their environment (e.g., Allen, 2000; Bates et al., 1976; Clancy, 2004; Greenfield & Smith, 1976; Skarabela, 2007). Our work here is aimed at characterizing children's input rather than their productions, but we investigate many of the same cues used in this work as well. In particular, in addition to the linguistic cues that are emphasized in discourse processing in adults, we analyze the interactional cues used in child-directed speech to signal joint focus of attention (Baldwin, 1995; Carpenter, Nagell, & Tomasello, 1998), including eye-gaze, pointing, and holding objects.

Pronoun use. Research on adult discourse has established that conventions for appropriate reference to discourse entities include the use of different linguistic forms for the introduction and re-mention of a referent: First mentions tend to be longer and more explicit, presumably because those referring expressions serve the function of identifying a new referent, whereas reduced forms and pronouns are reserved for subsequent mentions of what has, at that point in the discourse, become a more accessible entity (Ariel, 1990; Gundel, Hedberg, & Zacharski, 1993; Prince, 1992). These conventions are so strong that a violation of the expectation for reduction (i.e., the use of a full noun phrase or name instead of a pronoun for a familiar entity) can lead to processing difficulty: Adults show longer reading times when a topical entity is rementioned with a repeated name than when it is referenced with a pronoun (Gordon, Grosz, & Gilliom, 1993).

In acquisition, researchers have examined both children's early pronominal productions and their ability to use discourse factors in disambiguating pronoun reference during comprehension. In production, children make use of the distinction between "given" and "new" referents in their choices about pronoun production quite early on (e.g. Guerriero, Oshima-Takane, & Kuriyama, 2006). In addition, a variety of recent work tests

whether children are able to use sentence position (e.g., first mention), though evidence on this question is mixed: Some studies find early evidence for position use in long processing windows (Song & Fisher, 2005, 2007; Pyykkönen, Matthews, & Järvikivi, 2010) while others find more limited use in older children (Arnold, Brown-Schmidt, & Trueswell, 2007).

Overall, our prediction is that child-directed discourses should show the same trends towards pronominalization as adult-directed discourses, providing a robust cue for discourse segmentation.

Referent position. Not only is the form of reference predicted to vary with discourse position, but the location of a referring expression within an utterance is also predicted to vary. This is based on the observation that the location of a referring expression correlates with the information status of the referenced discourse entity, such that (relatively) familiar entities are referenced earlier in a sentence whereas (relatively) unfamiliar entities are referenced later (Lambrecht, 1994). A body of work in acquisition has examined children's omission of "given" information in their own productions (much of this in languages that allow pervasive omission, e.g. Guerriero et al., 2006; Clancy, 2004; Narasimhan, Budwig, & Murty, 2005), although this research does not provide direct evidence about children's ability to make use of discourse structure in word learning in particular. In our study, we consider the final word of each utterance, with the prediction that entities are less likely to be referenced utterance-finally as a topical discourse progresses and an entity becomes more familiar.

Utterance complexity. Information-theoretic models of language production (Genzel & Charniak, 2002; Levy & Jaeger, 2007) posit that, as common ground increases and discourse entities become more familiar, speakers and listeners are better equipped to handle longer and more complex sentences. For our analysis, we measure complexity as mean utterance length to test the prediction that complexity increases over the course of a

topical discourse.

Social cues. When talking to a young child, speakers are likely to use social cues like eye gaze and pointing both to draw attention to a referent and to signal to the child that they are sharing attention. This joint focus of attention (often shortened to *joint attention*) is an important part of children's early word learning (Tomasello & Farrar, 1986). Some instances of joint attention come about due to adults directing attention to particular referents, whereas others are a result of the adult following the child's attention ("follow-in" labeling) (Baldwin, 1991), and these lead to somewhat different outcomes for learners (Tomasello & Todd, 1983). In addition, the degree to which children actively direct caregivers' attention also changes over the course of development (Carpenter et al., 1998), and the degree to which caregivers actively follow in changes depending on how difficult or complex the referential task is (Rohlfing, 2011).

Here we hypothesize that, regardless of whether the referent is established by caregiver or child, cues like pointing and eye gaze will be used more during the process of establishing the referent than later in the topical discourse about that referent. One recent experimental study finds that young children can integrate social cue use across multiple, independently ambiguous sentences in a short discourse, providing tentative support for this hypothesis (Horowitz & Frank, 2013). In our study, we follow this previous work by evaluating whether caregivers' use of social cues varies across a topical discourse.

Corpus and original annotations

The corpus we use consists of a set of videos showing mothers and children involved in object-centered play in their homes, collected by Fernald and Morikawa (1993).[2] A

[2]This publicly-available corpus is introduced in more depth in Frank et al. (2013). The corpus was selected because the play session settings are sufficiently restricted to have permitted annotation of the full set of object referents available in the context.

representative excerpt from the corpus is shown in Table 2 in the following section. Crucially, the excerpted transcript contains the properties of interest which were introduced in the constructed example in Table 1, namely several word-learning opportunities that depend on the inference of a shared topic across utterances. For example, the word "doggy" appears in one utterance ("where's the doggy CHI", with the child looking at another referent) while the pointing gesture to the dog occurs in a later utterance which contains no explicit use of the referent name ("what's that"). In addition, although the link between pigs and toes is made explicit in the last line ("does the piggy have toes"), that information is made available earlier if the 'pig' topic is surmised to start earlier ("there's the piggy") and extend through several subsequent utterances that mention toes ("toes", "where's the toes").

Although Fernald and Morikawa's original study analyzed videos of American and Japanese mothers, we focus only on the American data. The 24 available videos of English-speaking children range in length from 3 to 22 (M=12.2) minutes. Children in these videos fall into three age groups: 6 months (N=8, mean age 6 months), 11–14 months (N=8, mean age 12.6 months), and 18–20 months (N=8, mean age 18.9 months). Each video captures a single mother-child play session in which mothers were given several pairs of toys (e.g., dog and pig puppets, or a ball and a box) by the experimenter and asked to play with each pair for a 3–5 minute period. More details are available in Frank et al. (2013).

We restricted our analysis to portions of the play sessions in which the mother was free to play with and talk about any of the objects that were present; we eliminated the final segment of each session, which involved a directed hiding game and which restricted the type of language the mother used. With the inclusion of only free-play portions of the videos and the exclusion of one video with limited data, the dataset consists of 23 of the original 24 videos, with the number of utterances per video ranging from 56 to 397

(mean=202). The original corpus from Fernald and Morikawa (1993) consists of the transcriptions of the set of 24 videos, including the extraction of the individual utterances.

In Frank et al. (2013), each utterance was then annotated with the following properties: intended referent, objects present, mother's and child's points of gaze, location of the mother's and child's hands, and direction of mother's points. Intended referent was operationalized as an intention to refer linguistically to an object; this included the mention of an object by name ("look at the *doggie*") or pronoun ("look at *his* eyes and ears"). In cases where the object was evoked only with a property like "red," a super-/subordinate terms like "animal", or a part term like "eye," the referent was coded as the relevant object. Exclamations like "oh" were not judged to be referential, even if they were directed at an object.

Despite its strengths (and its relative uniqueness as a corpus with reference and social cues fully annotated), this original corpus nevertheless had a number of limitations. First, in any corpus of spoken speech, the demarcation of utterances is somewhat subjective. This effect is exaggerated in the current corpus due to the frequent lack of formal complete sentences. Second, the reliability of social cue annotations was variable, with hand position being the most reliable ($\kappa = .8$) and gaze being the least ($\kappa = .47$). We return to these limitations in interpreting our results below.

The intended-referent annotations from the 2013 version of the corpus appear in Table 2 in the "Raw referent" column. The additional columns of Table 2 show annotations that we added for the purposes of this study. The next section describes that annotation process, which uses the cross-utterance notion, introduced above, of an inferred topical discourse sequence.

Identification of topical discourse sequences

Establishing the "correct" structure that should be inferred from a discourse remains an open research problem. In contemporary computational work, the structure of discourse is often established by adjoining sentences to a growing discourse structure based on the inference of a coherent dependency between one sentence and another (e.g., Prasad et al., 2008; Wolf & Gibson, 2005). Dependencies can hold both locally between adjacent sentences or proximally between a pair of separated, but not overly remote, sentences. In either case, one of the cues that marks membership in a discourse sub-structure is shared topic. For our purposes, one of the advantages of the corpus used in this study is that the speech has relatively few crossing dependencies.[3] Instead, the utterances consist largely of sequences of comments on the same object (see Table 2).

A TOPICAL DISCOURSE SEQUENCE is the term we use here to describe a set of adjacent or near-adjacent utterances that share a joint topic. Since our goal is to analyze properties of the discourse at different points within such a sequence, we must first add topical sequence annotations to the utterances that appear in the corpus. The challenge is that such sequences are often implicit; participants in a discourse naturally infer these topical sequences and have intuitions about the transition from one sequence to a new one, but the utterances themselves may not always contain overt cues about which sequence they belong to. To address this, we use three different strategies: human-coded raw referents, human-inferred topical sequences, and model-inferred topical sequences.

[3] As an example, a structure with crossing dependencies can be seen in the following passage: *On Saturday, the dog ran away from home. He thought his owners didn't love him. He didn't like the food they fed him or the toys they gave him. Last week, they fed him liver, and the week before they had offered him a toy mouse that was meant for cats. However on Sunday, he realized he missed them and decided to return home.* Establishing the coherence of this passage requires the inference of a link between the first and last sentences—a link that must be posited despite the intervening elaborations and subpoints.

Table 2

Example topical sequences for an excerpt of speech to an 18-month-old. See text for details of annotations. The string CHI is the child's name.

	Utterance	**Raw referent sequence**	**Human-inferred sequence**	**Model-inferred seqence**
1.	where's the doggy CHI	dog	dog	dog
2.	where's the doggy	dog	dog	dog
3.	where's the doggy	dog	dog	dog
4.	where's the doggy	dog	dog	dog
5.	what's that	dog	dog	dog
6.	what's that	dog	dog	dog
7.	what's that CHI	dog	dog	dog
8.	what's that	child	dog	dog
9.	what's that	child	dog	
10.	nose	dog	dog	dog
11.	what's this	dog	dog	dog
12.	what's this	dog	dog	dog
13.	what's this	dog	dog	dog
14.	what's this	dog	dog	dog
15.	eye	dog	dog	dog
16.	eye	dog	dog	
17.	can you say eye		dog	
18.	are you eating the doggy's nose	dog	dog	dog
19.	poor nose	dog	dog	
20.	nose	dog	dog	
21.	hard nose	dog	dog	
22.	ha			
23.	there's the piggy	pig	pig	pig
24.	you eating the piggy's nose	pig	pig	pig
25.	look CHI		pig	pig
26.	see the piggy	pig	pig	pig
27.	should we do your toes		pig	pig
28.	this little piggy [singing]		pig	pig
29.	toes	pig	pig	pig
30.	where's the toes	pig	pig	pig
31.	tail	pig	pig	pig
32.	does the piggy have toes	pig	pig	pig

The annotated excerpt in Table 2 shows the topical sequence annotations based on those three strategies. We describe each strategy in turn below.

Raw referent sequences

The first inference strategy relies solely on the human-coded raw referent annotations (from Frank et al., 2013, as described in the previous section); a "raw referent sequence" is defined as a sequence of successive utterances that contain explicit mentions of the same object referent (or its properties or parts, e.g. "red" or "eye"). This follows the measure of topic continuity from Frank et al. (2013) and is a fairly coarse measure of topicality.

The raw-referent sequences may both under-estimate and over-estimate the number of utterances that belong to a particular topical discourse. For example, a series of same-referent utterances that are close in time may be interleaved with a small number of non-referential utterances that have the effect of fragmenting what might otherwise be interpreted as a single longer sequence. Alternatively, a long pause following a sequence of referentially related utterances may signal an intended topic break, such that a subsequent utterance may be more appropriately assigned to a new sequence even if it mentions the referent of the previous sequence as part of the transition to a new discourse topic. Note also that these referent sequences—as with the model- and human-identified sequences below—are a data-analytic construct: They include a number of references that would not necessarily be transparent to a child. We return to the question of how a child might identify the relevant discourse sequences in the General Discussion.

Table 2 gives an example of the possibility of underestimating the length of a topical sequence. In this conversation, the mother pauses in her description of a pig to encourage the child to look, saying "look CHI" (where CHI indicates the child's name) in order to bring his attention back to the pig. Simply identifying sequences as consistent sets of references to the same objects, as in Frank et al. (2013) and as shown in column 2 of Table 2, may understate the continuity of these conversations. To investigate whether there were longer stretches of discourse on a single topic when these interruptions were

taken into account, we created annotations of topical sequences—coherent sequences of utterances about a single referent, with some tolerance for occasional interjections.

Human-inferred sequences

Our second strategy for inferring topical sequences relies on an additional set of human judgments that we collected for this study. We asked human coders to mark topical sequences by following a basic set of instructions for grouping sentences with the same referent, in this case a toy. Because of the necessity of a careful explanation of this task and its inherent subjectivity, we used a group of trained in-lab annotators. Our goal was to create a set of annotations that indicated, for each utterance, which topical sequence it belonged to, if any. We instructed annotators to read the transcripts and assign each utterance to a particular sequence or otherwise mark it as "non-topical," with the constraint that topical sequences were continuous (and so non-topical utterances were not permitted within a sequence but there could be some non-topical material between sequences). An example annotation is shown in Table 2. Inter-coder reliability for these annotations and their relationship to the raw-referent sequences are given below in the section on Annotation Results.

A concern that arises with the use of human coders' inferred sequences is that the coders may be sensitive to precisely the kinds of discourse markers that we intend to evaluate, so this method raises the possibility of circularity. In the following section, we introduce a model for automatically identifying topical sequences; the model uses the raw-referent information combined with timing information.

Model-inferred sequences

We coded topical sequences using a smoothing technique. We note that this model of topical sequence discovery is not a cognitive model of discourse processing; instead it is used here as a tool for data analysis, allowing us to identify discourse units in principled

ways in order to examine corresponding linguistic and social cues.

Our unsupervised model of sequence discovery uses timing information, so we additionally annotated the timing of utterance onsets relative to the video data. In order to speed the laborious process of annotating transcripts with the precise timing of each utterance in the video data, we made use of Amazon Mechanical Turk (AMT). AMT is a "crowdsourcing" marketplace, where workers in the United States and around the world can be paid anonymously for small amounts of work; AMT is already being used widely as an experimental tool in computer science (Hsueh, Melville, & Sindhwani, 2009) and cognitive science (Munro et al., 2010). We posted each video and transcript and asked annotators to write the timestamp of each utterance beside it (with payment varying as a function of transcript length).

Because we were concerned about the quality of the work from this method, we posted each job three times and then took the mean of the two closest annotations for each sentence. This kind of voting-related method is similar to taking a median rather than a mean, ensuring that a single typo cannot produce a large effect on the overall estimate, and has been used by a number of other groups working with Mechanical Turk (Carterette & Soboroff, 2010; Irvine & Klementiev, 2010; Madnani, Boyd-Graber, & Resnik, 2010). For example, if the sequence of annotations was 12, 120, 121 (with the 12 presumably due to a typo), the overall mean would be 84, whereas our method would produce 120.5. The cost of this triple-entry was still significantly lower than hiring individuals to perform this annotation from within our lab. Spot-checking of these annotations suggested that the resulting data were of high quality, implying that a similar process could be used in the future on a larger scale.

To automate the sequence-discovery process, we created a variant of a Hidden Markov Model (HMM), as described in Appendix A. This model uses the raw referent annotations (from Frank et al., 2013) and the utterance onsets (from the AMT

annotation) to estimate each utterance's membership within a topical sequence. The last column in Table 2 shows the HMM-model-inferred sequences, which in the case of that transcript excerpt, identifies longer dog and pig sequences than the sequences inferred from the raw referents.[4]

In the next section we discuss the characteristics of the raw-referent sequences, the HMM-model-inferred sequences, and the human-inferred sequences.

Annotation results

To visualize the HMM-model-inferred and human-inferred sequences in comparison to the raw-referent sequences, we created a "Gleitman plot" (see Frank et al., 2013) for one video in the corpus, as shown in Figure 1. The colored dots indicate which referents are present and being talked about. The thin, medium, and thick black bars indicate topical sequences in the raw referents, model-inferred, and human-inferred topic assignments, respectively. The fact that some black bars are longer than any sequences of red or green (reference-marking) dots shows the effect that smoothing and human-annotating had on the discovery of topical sequences: Topics extend through time even when intervening utterances do not reference the topic directly. The visualization matches the characteristics of the topical sequences shown in Table 2, in which the human-inferred and model-inferred sequences correctly smooth over the presence of off-topic utterances (e.g., "look CHI" in Table 2).

[4]We also constructed a smoothed version of the raw-referent sequences, which allows a single off-topic utterance to occur between two utterances that contain overt ('raw') mentions of the entity. We will call this smoothing technique the RAW+1 model. Such sequences are not marked in Table 2 but would have the effect of filling in two single off-topic gaps in the dog utterances (utterance 10, "can you say eye") and the pig utterances (utterance 25, "look CHI"). We include the analysis based on this model in footnotes in the section on "Analyses of Topical Discourse Sequences". The RAW+1 results are largely consistent with the other reported results.

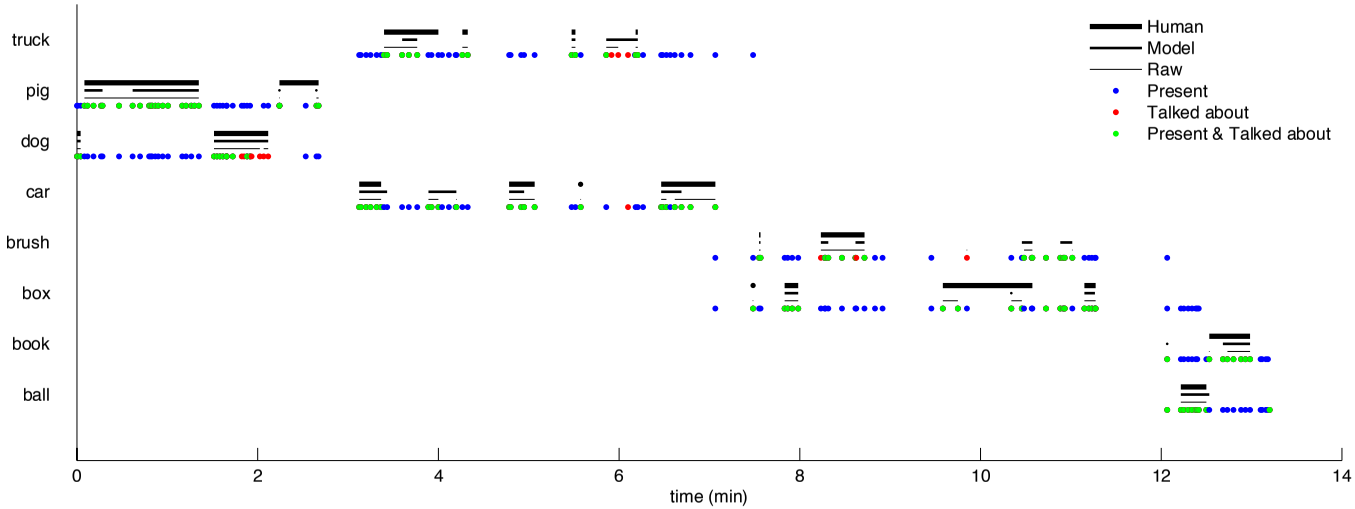

Figure 1. Sample Gleitman plot for a Fernald and Morikawa video. Rows denote objects; the x-axis marks time. Dots appear at utterance onset times; dot color reflects the raw video annotation of object presence and object reference. Blue denotes that the object was present when the utterance was uttered but not overtly referenced; red denotes that the object was overtly referenced but not present; green denotes that the object was present and overtly referenced. The black bars denote topical sequences: raw referent sequences (the union of red and green dots), human-inferred sequences, and model-inferred sequences.

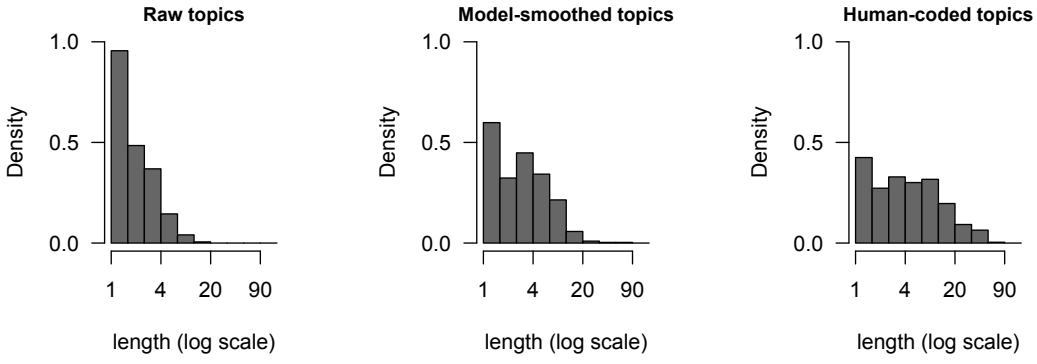

Figure 2. Mean topical sequence length (number of utterances). Left gives the distribution for raw-referent sequences; middle gives the distribution for the model-inferred sequences; right gives the distribution for the human-inferred sequences.

Next, we compared the distributional statistics of the raw-referent sequences, the HMM-model-inferred sequences, and the human-inferred sequences. The resulting topic assignments from the model and the human coders reduced the total number of topical sequences, in comparison with the number of sequences calculated with the raw-referent annotations. The raw topics yielded a total number of topical sequences per video that ranged from 10 to 88 (mean=45.0); the model-assigned topics yielded a total per video that ranged from 4 to 51 (mean=27.2); the human-coded topics yielded a total per video that ranged from 5 to 50 (mean=21.7).

Figure 2 shows three histograms that display the differences in sequence length among the raw-referent topics, the model-inferred topics, and the human-inferred topics. Discourses are considerably longer in the model-assigned and human-coded data than in the raw-referent sequences.

When we consider utterance onset times, we find that the gaps between utterance onsets are shorter within topical sequences than at sequence boundaries (raw: 3.3 vs. 4.9;

model: 2.4 vs. 5.9; human: 3.3 vs. 4.9), showing that the raw-referent and the human-inferred sequences are in keeping with Frank et al.'s (2013) claim that utterances that are close in time are likely to be close in topic. This feature is also upheld in the model-inferred sequences, where we see the largest difference between within-sequence and between-sequence gaps.

Finally, we make use of results from a related literature on discourse segmentation in computational linguistics to provide quantitative comparisons between the different means of identifying topical sequences. We use three metrics for the similarity of topic assignments across all files; all provide a score between 0 and 1, with 1 marking maximum agreement between measures.[5] The first metric is a simple proportion equivalence of sequence assignments, which we refer to as $a = b$. The second metric, p_k, is a moving-window method that was introduced by Beeferman, Berger, and Lafferty (1999) as a way of calculating the probability that two random utterances are correctly classified as being in the same sequence segment. The third metric, WindowDiff, was introduced by Pevzner and Hearst (2002) in response to the widespread use of p_k. WindowDiff addresses a number of issues with p_k, including different weighting of false negatives and false positives and a lack of "partial credit" given to boundaries that are placed close to the true sequence boundary. WindowDiff corrects these issues by comparing the number of sequence boundaries posited within some window to the true number of boundaries and then moving this window across the corpus.

Using these three metrics, we computed pairwise comparisons between topical sequences (shown in Table 3). First, we compared the human-inferred sequences to the

[5]Note that for this section there is some ambiguity between two tasks that a sequence or topic segmentation algorithm can provide: assignment of sentences to topics and finding boundaries between sequences. Our human annotators and model produced sequence assignments, but it is easy to convert this data into boundaries for evaluations.

Table 3

Measures of sequence-discovery accuracy between raw referent annotations, model-inferred topics, and human-inferred topics, and human-inferred topics with two coders.

Measure	raw–human	model–human	human–human (3 videos)
$a = b$	0.60	0.67	0.81
p_k	0.26	0.47	0.85
WindowDiff	0.13	0.32	0.69

raw-referent sequences and to the the model-inferred sequences. Then, in order to compute an upper-bound for these measures, a second annotator double-coded three randomly selected videos out of 24 and we compared human judgments to other human judgments (this also serves as a test of the coders' reliability in this task).

On all three measures, model results were closer to human annotations than the raw sequences were, suggesting that the model did generally identify more human-like topical sequences. The improvement over the raw baseline is relatively modest for some measures, compared with the human-human correlation. We note that both WindowDiff and especially p_k severely penalize over-segmentation (which is precisely what the raw sequences represent), hence the very low raw–human scores for these measures. Nevertheless, this set of metrics also suggests that our model is generally providing some value in finding appropriate sequences.

Analyses of Topical Discourse Sequences

In order to determine whether discourse markers change over the course of a topically related sequence of utterances, we consider the content of the caregivers' speech and the social cues between caregiver and child. We analyze raw-referent,

HMM-model-inferred, and human-inferred sequences. The observed markers are modeled using mixed-effects regressions (logistic or linear, as appropriate for the particular variable) with random caregiver-specific and referent-specific intercepts (Gelman & Hill, 2007). We coded sequence position (i.e., utterance number in a given topical sequence) on a log scale, both because this method provided better fits to the data and because previous work suggested that discourse features might be non-linear in their distribution over time (Dowman et al., 2008; Frank et al., 2013).

We excluded utterances that were not part of a minimal sequence, defined as at least 3 successive utterances on the same topic. The removal of utterances that did not participate in a minimal sequence was established separately for the raw-referent, model-inferred, and human-inferred sequences. The raw sequences yielded a remaining dataset containing 1313 utterances (4 to 154 utterances per video, mean=57.1); the model-inferred sequences contained 2220 utterances (16 to 222 utterances per video, mean=96.5); the human-inferred sequences contained 3593 utterances (14 to 376 utterances per video, mean=156.2).

Figures 3 and 4 show the behavior of the seven discourse markers (three linguistic, four social) that we analyze across the different types of topical sequences. In Tables 4 and 5, we report the logistic- and linear-regression coefficient estimates and p-values for the factors child age (coded as *age*, a numeric factor) and the log of the sequence position within the topical sequence (coded as the utterance number, or *logSeqPos*, a numeric factor) and an interaction between the two. The factor *age* was centered.

Pronoun use. We predicted that the rate of pronominalization would increase over the course of a topical sequence, and our results confirmed this prediction in the raw, model-inferred, and human-inferred sequences. Sequence position was a significant factor

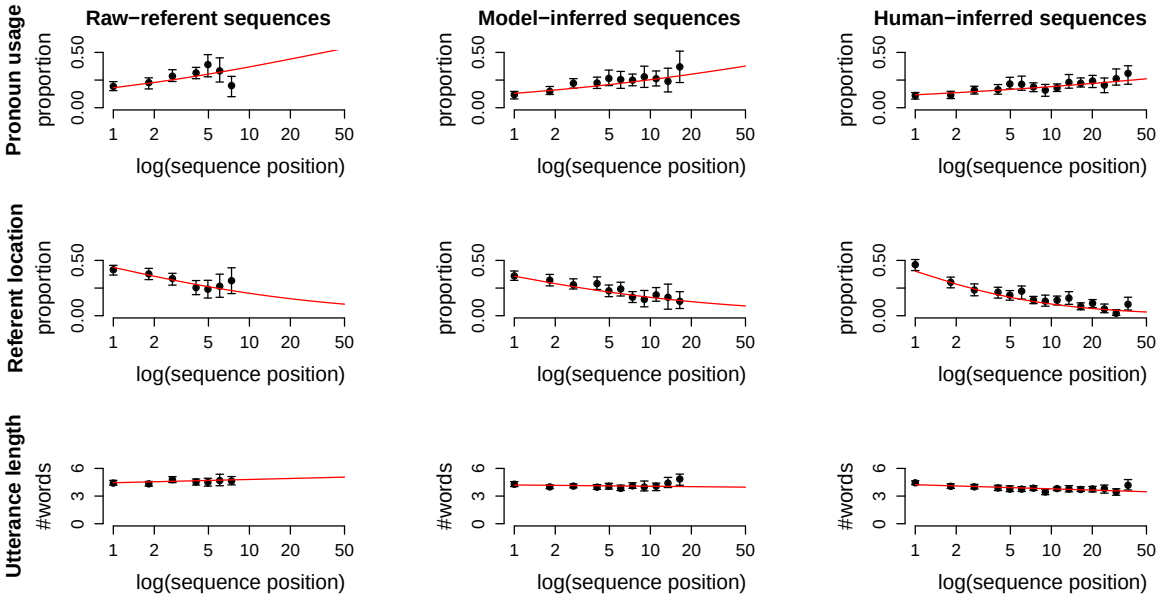

Figure 3. Graphs plot linguistic-cue means, collapsed across age, at successive sequence positions in raw referent, HMM-model-inferred, and human-inferred sequences; error bars show confidence intervals; regression lines correspond to logistic and linear models built with observed data. Note that the regression lines reflect a model fit only with log sequence position, not age; full model fits are given in Table 4.

for modeling the binary outcome of pronominalization using all three methods for identifying sequences, with more pronouns (3rd person nominative/accusative/possessive forms plus *one*) being used later in the sequences.

Referent location. To test our prediction that references to a sequence topic occur later in a sentence during the early utterances of a sequence, we considered the final word of each utterance. Topical entities were predicted to be less likely to be referenced utterance-finally (a location typically reserved for new information; see Ward & Birner, 2004) as a sequence progresses and an entity becomes more familiar. Fernald and Morikawa (1993) noted the strong prevalence of referential nouns at the ends of sentences

Marker	β_{RAW}	p-val	β_{HMM}	p-val	β_{HUMAN}	p-val
Pronoun use:						
logSeqPos	**0.422**	**.001**	**0.358**	**.001**	**0.254**	**.001**
age	-0.191	.259	-0.067	.644	-0.170	.231
logSeqPos × age	-0.006	.954	-0.070	.313	0.034	.477
Utterance-final mention:						
logSeqPos	**-0.476**	**.001**	**-0.447**	**.001**	**-0.752**	**.001**
age	-0.314	.100	-0.101	.523	-0.220	.147
logSeqPos × age	0.088	.406	-0.011	.876	-0.086	.086
Utterance length:						
logSeqPos	0.152	.103	-0.057	.335	**-0.194**	**.001**
age	0.207	.294	**0.338**	**.045**	0.160	.280
logSeqPos × age	-0.112	.261	**-0.135**	**.032**	-0.030	.464

Table 4

Predictors for modeling linguistic markers in mixed-effect models (bolding indicates significance).

in the English caregivers' speech, hence our choice to not target subject versus object arguments. Our results confirmed that in the raw, model, and human-inferred sequences, sequence position was a significant factor for modeling the binary outcome of sentence-final mention, with fewer sentence-final references later in the sequence.

Utterance complexity. We also tested whether sentence complexity increases as the topical sequence progresses. Measuring complexity as mean utterance length revealed an effect of sequence position only in the human-inferred sequences, and the effect was in the opposite direction than predicted, with utterances decreasing slightly in length over successive utterances. In the model-inferred sequences, there was an effect of age, whereby older children heard slightly longer utterances, and a sequence position × age interaction whereby the slight decrease over sequence positions was only reliable for the older children.

This lack of an effect of increased complexity over the course of the topical discourse may be due in part to the nature of the video transcripts and the difficulties in identifying sentence units in naturally-occurring speech (see excerpt in Table 2), or there may be

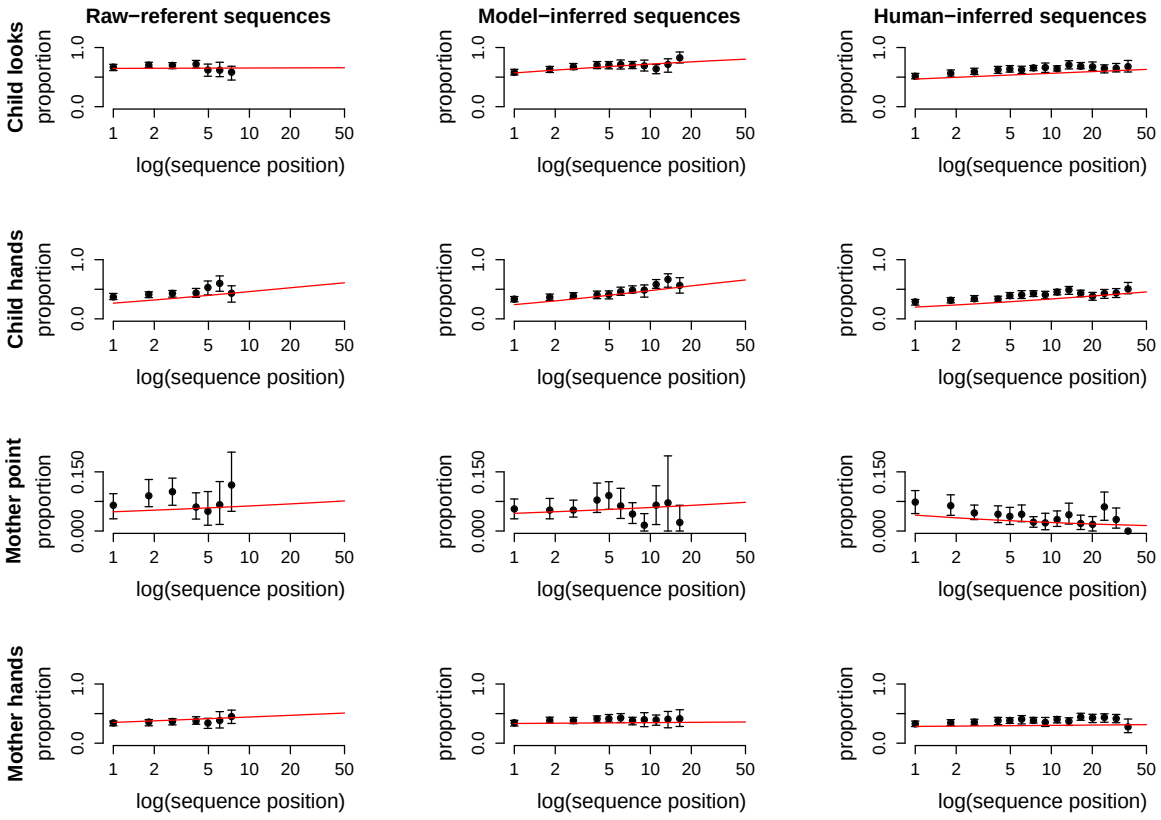

Figure 4. Graphs plot social-cue means, collapsed across age, at successive sequence positions in raw-referent, HMM-model-inferred, and human-inferred sequences; error bars and regression lines as in Figure 3.

unique constraints that arise from discourse that occurs in person or from speech that is situated in the presence of concrete objects (in contrast to the edited news text and adult telephone conversation used in previous work).

An effect of sequence position on complexity was similarly absent in a follow-up analysis in which we excluded utterances that contained only onomatopoetic words ("woof woof woof") or exclamatives ("oh", "good boy"). The exclusion of these utterances was intended to focus the analysis on utterances which, at a first approximation, could be

Cue	β_{RAW}	p-val	β_{MODEL}	p-val	β_{HUMAN}	p-val
Child's eyes:						
logSeqPos	0.012	.901	**0.288**	**.001**	**0.170**	**.001**
age	0.044	.788	-0.059	.625	-0.044	.740
logSeqPos × age	-0.013	.898	**0.153**	**.010**	-0.002	.953
Child's hands:						
logSeqPos	**0.375**	**.001**	**0.446**	**.001**	**0.309**	**.001**
age	**0.363**	**.037**	0.255	.149	**0.282**	**.047**
logSeqPos × age	-0.108	.296	0.108	.092	0.018	.646
Mother's points:						
logSeqPos	0.055	.761	0.128	.243	**-0.280**	**.003**
age	-0.022	.937	0.227	.376	0.206	.429
logSeqPos × age	**0.429**	**.032**	0.007	.951	0.000	.997
Mother's hands:						
logSeqPos	0.166	.098	0.028	.645	0.038	.348
age	-0.072	.710	-0.113	.536	-0.244	.204
logSeqPos × age	-0.146	.151	0.017	.773	**0.104**	**.006**

Table 5

Predictors for modeling social cues in mixed-effect models of raw-referent, model-inferred, and human-inferred sequences (bolding indicates significance).

expected to contain informative material, but there was still no effect of sequence position on utterance length.

Closer inspection of some of the transcripts offers a potential explanation for the observed decrease in complexity over sequence positions in the human-inferred sequences. In some of these sequences, there is a pattern that late in the sequences, mothers produce one-word utterances (which are shorter than the multi-word utterances early in the sequence). These one-word utterances include cases in which the mother comments on the derailing session ("oh", "fall"), acknowledges the end of the interaction with a particular toy ("bye"), or, more interestingly, names an attribute of the larger object ("nose", "toes", "tail", "wheels", etc.). The first is an understandable side effect of child play and the second seems to be an artifact of the experimental setting in which toys were offered and removed every 3 to 5 minutes. In the case of the object attributes, though, the

content of utterances might in fact reflect a growing complexity in the discourse: The more utterances into a topical sequence a mother gets, the more detailed (and hence semantically complex) the information can be that she offers to the child. Quantifying such trends will require substantially more sophisticated metrics of complexity, however.

In sum, the pattern of linguistic cues over the course of a topical sequence was most consistent for pronoun usage and referent location; the utterance complexity results may reflect the noisiness of this data or properties unique to in-person child-directed discourse. The pronoun and referent location effects were apparent in all three sequence types: raw-referent, HMM-model-inferred, and human-inferred.[6]

Social Cues. When a new topic is introduced, speakers are likely to draw attention to that entity, both in their words and with other social cues. We therefore evaluated cues related to joint attention (Baldwin, 1995; Carpenter et al., 1998), namely the position of mothers' and children's hands and their points of gaze.

The results show that children looked more to the referenced object over the course of a topical sequence, an effect apparent only in the somewhat longer model and human-inferred sequences. Children also touched the referenced object more over the course of a sequence, an effect apparent in the raw, model, and human-inferred sequences. Based on both of these metrics, it appears that children only gradually became engaged in the discourse, rather than shifting their attention immediately to the topic, and in these videos they did not start touching or holding the toy until later in the topical sequence

[6]The RAW+1 model mentioned in footnote 4 likewise patterned with the other models for both pronoun usage (main effect of sequence position: β=0.300, p<0.001; no effect of age: β=-0.269, p=0.096; no interaction: β=0.005, p=0.943) and referent location (main effect of sequence position: β=-0.493, p<0.001; no effect of age: β=-0.114, p=0.493; no interaction: β=-0.037, p=0.616). For utterance complexity, the RAW+1 model patterned with the raw-referent sequences in showing no effects (no effect of sequence position: β=-0.050, p=0.442; no effect of age: β=0.110, p=0.526; no interaction: β=-0.007, p=0.916).

(perhaps because many of the children were so young, as older children would likely have taken the toys more quickly). In the human-inferred sequences, there was also a main effect of age, whereby older children touched objects more than younger children.

These findings provide an interesting counterpoint to earlier work that has primarily made binary distinctions between referents that are "given" and those that are "new" (Clancy, 2004). Our data suggest graded increases in familiarity and givenness, leading to long-lasting changes in joint attention to objects (see Gundel et al., 1993; Skarabela, 2007). Considerably more work is needed to establish the generality of the pattern we observed, but our findings are nevertheless consistent with an older body of work on "social referencing" behavior that found graded effects of adult affect on children's engagement with novel toys (Hornik, Risenhoover, & Gunnar, 1987; Hornik & Gunnar, 1988; Walden & Ogan, 1988). It may be the case that although the young children in our sample did not grasp the specifics of the discourse that was being constructed, the overall level of parent engagement with the toys had a positive effect on the child's own willingness to engage.

For mothers' pointing, the human-inferred sequences showed a decrease over the course of a sequence. This likely corresponds to an attempt to get a child's attention when a new object first becomes the topic of the discourse. In the raw-referent sequences, an interaction between sequence position and age emerged, whereby the rate of pointing to the topical object rose most quickly for the oldest age group. This may reflect the raw-referent sequences' poor estimate of an utterance's position with the true underlying discourse sequence. For example, the third utterance about a toy may be labeled as sequence-final in the raw-referent sequence even if the topic in fact extends for many more utterances; if the mother points to the toy during that third (and seemingly final) utterance, those gestures could contribute to an apparent rise in the pointing behavior over the course of raw-referent sequences.

For mothers' hands, the only reliable effect was a sequence position $\times$ age

interaction in the human-inferred sequences, whereby the mothers of the oldest children decreased their object touching over the course of a sequence. This may reflect the older children's own increased touching of the object later on in the interaction, but both this and the result on pointing should be interpreted with caution due to the limited convergence across discourse identification methods.[7]

General Discussion

As one of the first quantitative investigations of discourse structure in an acquisition setting, the study presented here shows that topical discourse is characterized both by linguistic markers of topichood and by social cues related to joint attention. Across the topical sequences, we see patterns of pronominalization and sentence-final reference that are consistent with patterns observed in adult discourse: Less familiar information is referenced later in an utterance, and more familiar information is likely to be referenced with a pronoun. Also, across the discourse segments, children's patterns of hand and eye movements show increased attention to the topical object; mother's hand and eye movements are less reliable (potentially due to their concurrent task of monitoring the child).

[7]The RAW+1 model mentioned in footnote 4 confirmed that children looked more to the referenced object over the course of a topical sequence, in keeping with the HMM-model-inferred and human-inferred sequences, though it missed the sequence position × age interaction found in the HMM-model-inferred sequences (main effect of sequence position: β=0.214, p<0.001; no effect of age: β=-0.028, p=0.826; no interaction: β=0.108, p=0.116). The RAW+1 model matched all three other models in finding a main effect of sequence position in the analysis of children's hands, but it missed the effect of age found in the human-inferred sequences (main effect of sequence position: β=0.485, p<0.001; no effect of age: β=0.248, p=0.115; no interaction: β=0.087, p=0.225). Lastly, the RAW+1 model patterned with the HMM-model-inferred sequences for mother's points (no effect of sequence position: β=0.022, p=0.860; no effect of age: β=-0.004, p=0.986; no interaction: β=0.182, p=0.183) and mother's hands (no effect of sequence position: β=0.075, p=0.271; no effect of age: β=-0.169, p=0.340; no interaction: β=0.001, p=0.990).

In comparing the observed patterns of linguistic and social cues over the raw, model-inferred, and human-inferred sequences, it appears that effects in the raw sequences are, as predicted, noisier. Overall, however, the patterns are quite consistent across models, lending support to our choice of metrics for the various social and linguistic cues and also suggesting that the HMM model we propose succeeded in identifying relevant sequences. The benefit of smoothing (in the model-inferred and human-inferred topical sequences) is really only evident in the analysis of social cues, where we see a reliable effect of sequence position in children's looking in the model-inferred and human-inferred sequences, but not in the raw-referent sequences. This may be attributed to the fact that eye gaze is not manifested only at individual utterance times and may instead span multiple utterances, only some of which may have been identified as topical within the raw annotation.

The human-inferred sequences revealed the largest number of effects (roughly a superset of the ones found in the raw-referent and model-inferred sequences). Those sequences were the longest, which meant that if an underlying effect of sequence position was present, it would be easier to identify such effects with better estimates of which utterances are early versus late in a topical sequence. Given the additional sensitivity that the human-inferred sequences provide, we conclude that they (and the effects observed in the analysis of them) represent the most reliable source for drawing conclusions about the structure of child-directed discourse.

Using human-inferred sequences raises the possible circularity that the cues that we were analyzing (e.g., the number of pronouns in an utterance) were precisely the cues that contributed to the human annotators' own decisions, in which case the observed effects would be unsurprising. However, it is worth noting that the results with social cues are immune to this concern because the social cues were not available to the annotators who only had access to the text transcripts. The linguistic markers of topichood could have

influenced the annotators' decisions, and for that reason it is encouraging that the raw-referent and model-inferred sequences also showed effects of sequence position on pronoun usage and utterance-final mention.

If one of the functions of language is to provide the structure necessary to permit meaningful communication, one might hypothesize that discourses would be structured to increase the amount of information a speaker can convey. This is the argument put forward in work on the strategies that speakers employ to achieve communicative efficiency (Levy & Jaeger, 2007), on the complexity of sentences found later in a discourse (Genzel & Charniak, 2002), and on the establishment of speaker-listener common ground over the course of a conversation (H. Clark, 1996). Our results are largely consistent with these models of language use: Speakers use reduced referring expressions such as pronouns when topical entities are easily retrievable and listeners show signs of engaging in joint attention to entities that have become part of the common ground.

Nevertheless, our study has several limitations that should be addressed in future work. First and most prominent among these is a general issue for studies that link sentential and super-sentential information in acquisition: how to determine utterance boundaries. In our study, we relied on the transcripts that accompanied the corpus we studied, but these transcripts were made without explicit attention to prosodic phrase boundaries. It may be that this issue led to the relatively null effects of discourse position on utterance length (although other explanations, including those described above, are of course possible). Second, our study was limited to very simple referential contexts. This simplifying assumption was what allowed us to identify topical discourse sequences with relatively high reliability, but it also keeps us from drawing strong conclusions about topical discourse that does not rely on object reference. Future work should investigate the extension of the current methods to discourse whose utterances are not linked solely by joint reference but rather require other types of inference.

As noted in the introduction, researchers studying word learning have often treated sentences as largely independent units. The results presented here establish that larger discourse-level regularities are available in child-directed speech, such that children may have access to the topical nature of human discourse even if they cannot understand individual sentences in their entirety. The full extent of children's understanding of discourse structure will be a question for future experimental work, but our results point to a variety of ways that children might make use of discourse structure. At a minimum, a longer episode of discourse about a particular referent might allow young learners to integrate naming events with social cues even if the two were not presented in precise temporal synchrony (and preliminary evidence indicates that this sort of "smoothing" is possible, at least for older learners; Horowitz & Frank, 2013).

More generally, topical discourse may create a powerful learning context, in which the referent is fixed and mutually known and a parent can elaborate on details, including part and property terms, super- and subordinate labels, and generic features of a kind. All of these complex details can be difficult to convey in just a single utterance: Consider shoehorning reference, class, and kind information into a single sentence ("this toy here is a dog, which is a kind of animal that has four legs and barks"). Instead the progression of discourse allows these distinct tasks to be distributed throughout a collaborative conversation (E. V. Clark, 2003).

In sum, we take these exploratory results as an invitation to consider discourse-level phenomena in the acquisition setting, even for very young children. Discourse topics wax and wane over the course of a conversation with subtle repercussions in communication and common ground, and our results suggest that child-directed speech presents a new and rich domain for analyses of discourse structure.

References

Allen, S. E. (2000). A discourse-pragmatic explanation for argument representation in child inuktitut. *Linguistics*, *38*(3), 483–521.

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. London: Routledge.

Arnold, J. E. (2001). The effects of thematic roles on pronoun use and frequency of reference. *Discourse Processes*, *31*, 137-162.

Arnold, J. E., Brown-Schmidt, S., & Trueswell, J. C. (2007). Children's use of gender and order-of-mention during pronoun comprehension. *Language and Cognitive Processes*, *22(4)*, 527-565.

Asher, N., & Lascarides, A. (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.

Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, *62(5)*, 874-890.

Baldwin, D. A. (1995). Understanding the link between joint attention and language. In C. Moore & P. J. Dunham (Eds.), *Joint attention: its origins and role in development.*

Bates, E., et al. (1976). *Language and context: The acquisition of pragmatics* (Vol. 13). Academic Press New York.

Beeferman, D., Berger, A., & Lafferty, J. (1999). Statistical models for text segmentation. *Machine learning*, *34*(1), 177–210.

Caramazza, A., Grober, E., Garvey, C., & Yates, J. (1977). Comprehension of anaphoric pronouns. *Journal of Verbal Learning and Verbal Behaviour*, *16*, 601-609.

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)*.

Carterette, B., & Soboroff, I. (2010). The effect of assessor error on ir system evaluation.

In *Proceedings of the 33rd international acm sigir conference on research and development in information retrieval* (pp. 539–546).

Clancy, P. M. (2004). The discourse basis of constructions: Some evidence from korean. In *Proceedings of the 32nd stanford child language research forum. stanford, ca: Csli publications.*

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

Clark, H. (1996). *Using language.* Cambridge Univ Press.

Dowman, M., Savova, V., Griffiths, T., Kording, K., Tenenbaum, J., & Purver, M. (2008). A probabilistic model of meetings that combines words and discourse features. *Audio, Speech, and Language Processing, IEEE Transactions on*, *16*(7), 1238–1248.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*(6), 1017–1063.

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

Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the second year. *Psychological Science*, *9(3)*, 228-231.

Frank, M., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578.

Frank, M., Tenenbaum, J., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. In *Language, learning, and development* (p. 1-24).

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

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models* (Vol. 625). Cambridge University Press Cambridge.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the*

40th Annual Meeting of the Association for Computational Linguistics.

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

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54.

Gordon, P. C., Grosz, B. J., & Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, *17*, 311-347.

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

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech acts.* NY: Academic Press.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, *21*, 203-225.

Guerriero, A. S., Oshima-Takane, Y., & Kuriyama, Y. (2006). The development of referential choice in english and japanese: a discourse-pragmatic perspective. *Journal of child language*, *33*(04), 823–857.

Gundel, J., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, *69*, 274-307.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science*, *3*, 67-90.

Hornik, R., & Gunnar, M. R. (1988). A descriptive analysis of infant social referencing. *Child development*, 626–634.

Hornik, R., Risenhoover, N., & Gunnar, M. (1987). The effects of maternal positive, neutral, and negative affective communications on infant responses to new toys. *Child Development*, 937–944.

Horowitz, A., & Frank, M. C. (2013). Young children's developing sensitivity to discourse continuity as a cue to reference. In *Proceedings of the 35th Annual Meeting of the*

Cognitive Science Society.

Hsueh, P., Melville, P., & Sindhwani, V. (2009). Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the naacl hlt 2009 workshop on active learning for natural language processing* (pp. 27–35).

Irvine, A., & Klementiev, A. (2010). Using mechanical turk to annotate lexicons for less commonly used languages. In *Proceedings of the naacl hlt 2010 workshop on creating speech and language data with amazon's mechanical turk* (pp. 108–113).

Kaiser, E. (2012). Taking action: a cross-modal investigation of discourse-level representations. *Frontiers in Psychology*(3(156)), 1-13.

Kehler, A. (2002). *Coherence, reference, and the theory of grammar.* Stanford, CA: CSLI Publications.

Kehler, A., Kertz, L., Rohde, H., & Elman, J. (2008). Coherence and coreference revisited. *Jo. of Semantics*, *25*, 1-44.

Koornneef, A. W., & Van Berkum, J. J. A. (2006). On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye-tracking. *Journal of Memory and Language*, *54*, 445-465.

Lambrecht, K. (1994). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents.* Cambridge: Cambridge University Press.

Levy, R., & Jaeger, T. (2007). Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems.* Cambridge: MIT Press.

Madnani, N., Boyd-Graber, J., & Resnik, P. (2010). Measuring transitivity using untrained annotators. In *Proceedings of the naacl hlt 2010 workshop on creating speech and language data with amazon's mechanical turk* (pp. 188–194).

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a

functional theory of text organization. *Text*, *8*, 243-281.

Medina, T., Snedeker, J., Trueswell, J., & Gleitman, L. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014.

Munro, R., Bethard, S., Kuperman, V., Lai, V., Melnick, R., Potts, C., et al. (2010). Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the naacl hlt 2010 workshop on creating speech and language data with amazon's mechanical turk* (pp. 122–130).

Narasimhan, B., Budwig, N., & Murty, L. (2005). Argument realization in hindi caregiver–child discourse. *Journal of Pragmatics*, *37*(4), 461–495.

Pevzner, L., & Hearst, M. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, *28*(1), 19–36.

Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of Pragmatics*, *12*, 601-638.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., et al. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)* (pp. 2961–2968).

Prince, E. F. (1992). The zpg letter: Subjects, definiteness, and information-status. In S. Thompson & W. Mann (Eds.), *Discourse description: Diverse analyses of a fundraising text* (p. 295-325). Amsterdam/Philadelphia: John Benjamins.

Pyykkönen, P., Matthews, D., & Järvikivi, J. (2010). Three-year-olds are sensitive to semantic prominence during online language comprehension: A visual world study of pronoun resolution. *Language and Cognitive Processes*, *25*(1), 115–129.

Quine, W. (1973). *Word and object* (Vol. 4). The MIT Press.

Rohde, H., & Frank, M. (2011). Markers of discourse structure in child-directed speech. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society.*

Rohlfing, K. J. (2011). Exploring" associative talk": When german mothers instruct their two year olds about spatial tasks. *Dialogue & Discourse*, *2*(2).

Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1-2), 39–91.

Skarabela, B. (2007). Signs of early social cognition in children's syntax: The case of joint attention in argument realization in child inuktitut. *Lingua*, *117*(11), 1837–1857.

Smyth, R. J. (1994). Grammatical determinants of ambiguous pronoun resolution. *Journal of Psycholinguistic Research*, *23*, 197-229.

Snedeker, J., & Trueswell, J. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, *49*, 238-299.

Song, H.-j., & Fisher, C. (2005). Whos she? discourse prominence influences preschoolers comprehension of pronouns. *Journal of Memory and Language*, *52*(1), 29–57.

Song, H.-j., & Fisher, C. (2007). Discourse prominence effects on 2.5-year-old children's interpretation of pronouns. *Lingua*, *117*(11), 1959–1987.

Stevenson, R., Crawley, R., & Kleinman, D. (1994). Thematic roles, focusing and the representation of events. *Language and Cognitive Processes*, *9*, 519-548.

Tomasello, M., & Farrar, M. (1986). Joint attention and early language. *Child Development*, *57(6)*, 1454-1463.

Tomasello, M., & Todd, J. (1983). Joint attention and lexical acquisition style. *First language*, *4*(12), 197–211.

Walden, T. A., & Ogan, T. A. (1988). The development of social referencing. *Child development*, 1230–1240.

Ward, G., & Birner, B. (2004). Information structure and non-canonical syntax. In L. R. Horn & G. Ward (Eds.), *The handbook of pragmatics* (p. 153-174). Oxford: Basil Blackwell.

Wolf, F., & Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, *31*(2), 249–287.

Yu, C., & Ballard, D. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13-15), 2149–2165.

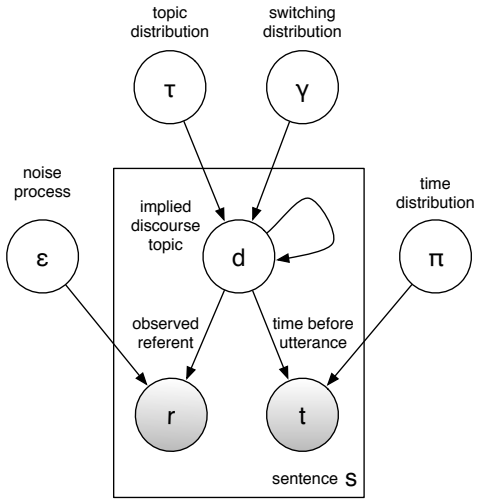

Figure 5. Schematic graphical model for the dependencies in our discourse-finding model.

Appendix A: Automatic topical sequence identification

To automate the sequence-discovery process, we created a variant of a Hidden Markov Model (HMM), shown in Figure 5. For each sentence s in the corpus, we assume that we observe both what the referent r_s is (if any; many sentences have no explicitly referenced object), and the time interval t_s preceding the sentence. On the basis of this information, our goal for each sentence is to infer the implied (hidden) topical discourse sequence d_s.

The model assumes that for each sentence, d_s is generated by the following process. First, flip a coin with weight γ to decide whether d_s will be the same as d_{s-1} or will start a new sequence (switching process). If it starts a new sequence, draw the new topic from the topic distribution τ and draw wait time t from the between-topic waiting time distribution π_b. If not, $d_s = d_{s-1}$ and draw t from the within-topic distribution π_w. Now flip a coin with weight ϵ to decide whether r_s will be the same as d_s, or whether r_s will be another topic from τ chosen uniformly at random. Aside from the time distributions, this

model resembles an HMM in that it encodes an immediate sequential dependency between hidden states.

Because this procedure contains many exponential-family distributions (the noise distribution ϵ, the switching distribution γ, the topic distribution τ, and the two time distributions π_b and π_w), we assign conjugate prior probability distributions to each and replace each with an integrated conjugate distribution (Gelman, 2004), so that the topic distribution is a multinomial-dirichlet, the switching and noise distributions are beta-binomial, and the time distributions are gamma-poisson (with corresponding parameter values for each).

Inference within this model can then be accomplished via a Gibbs sampler: a Markov-chain Monte-Carlo algorithm for estimating the posterior distribution over values of d for each sentence. Because model performance proved to be sensitive to the hyperparameter values of the conjugate distributions, we implemented a hyperparameter inference scheme in which, after each Gibbs sweep, a Metropolis-Hastings sampler modified hyperparameters for each distribution (we omit this step from Figure 5 and the generative process description above for simplicity). All hyperparameters were assumed to be drawn from an exponential distribution with rate 2, except for the Dirichlet parameter α_t, which was assigned rate 10 (so as not to promote excessive sparsity in the topic distribution).

For the simulations reported in this paper, the model was run independently on the data for each video for 2000 Gibbs sweeps. Each sentence was assigned its model sequence topic from the posterior samples (for discrete categorization tasks, this method is an estimator of the maximum a posteriori category assignment). In cases where no topic was favored in more than 50% of samples, the topic was set to be null, as with the "non-topical utterances" set by the human coders.