Markers of Discourse Structure in Child-Directed Speech

Hannah Rohde
rohde@stanford.edu
Department of Linguistics
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Although the language we encounter is typically embedded in rich discourse contexts, existing models of sentence processing focus largely on phenomena that occur sentence internally. Here we analyze a video corpus of child-caregiver interactions with the aim of characterizing how discourse structure is reflected in child-directed speech and in children's and caregivers' behavior. We use topic continuity as a measure of discourse structure, examining how caregivers introduce and discuss objects across sentences. We develop a variant on a Hidden Markov Model to identify coherent discourses, taking into account speakers' intended referent and the time delays between utterances. Using the discourses found by this model, 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 cues used to signal topicality in adult discourse are also available in child-directed speech and that children's responses reflect joint attention in communication.

Keywords: Language acquisition; discourse; social interaction; reference tracking; Bayesian modeling.

Introduction

Extracting meaning from linguistic input requires listeners to infer a variety of dependencies both within sentences and across larger sequences of structured discourse. Within a sentence, listeners must infer relationships among words and phrases, deploying a body of knowledge regarding lexical and syntactic dependencies. Across sentences, listeners must figure out what common themes tie a series of utterances together into a coherent whole. These two types of inferential processes—one local and one global—are at the core of language understanding, and yet our processing models have tended to focus largely on the former. Although researchers have begun to capture the broader structures that characterize coherent discourse and dialog, that work has been limited in scope to adult discourse. Such work has made progress in formally modeling the properties of coherent discourses (Asher & Lascarides, 2003; Grice, 1975; Kehler, 2002; Mann & Thompson, 1988) and even in showing how such properties can influence sentence-internal structure building (Altmann & Steedman, 1988; Rohde, Levy, & Kehler, in press), but there has been little work asking whether child-directed language shows evidence of these properties as well.

Work in language acquisition has begun to explore the issue of processing in children (Snedeker & Trueswell, 2004; Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998), but this work remains at the level of classic sentence processing and treats sentences as independent units. The goal in this paper is to reconsider the child-directed sentence unit in context, recognizing the increased informativity of spoken language when it is encountered in a rich discourse context.

We examine discourse structure in child-directed speech through the lens of topic continuity. Topic continuity, observed in repeated references to a set of related discourse entities, is part of what allows listeners to infer that sentences they hear do not appear together arbitrarily but rather relate in meaningful ways. Research on reference tracking in adults has addressed a variety of questions, including how listeners identify referents, how speakers signal shifts in referents, and what inferences are involved in resolving ambiguities between sentences. Answers to these questions have highlighted the range of information sources that are brought to bear in sentence processing—from domain-general cognitive reasoning about events and causality to language-internal principles about grammatical roles and pronominal forms (Grosz, Joshi, & Weinstein, 1995; Kehler, Kertz, Rohde, & Elman, 2008). The importance of topic continuity in acquisition lies in word learning, a task which requires the establishment of mappings between referring expressions and real-world objects.

Topic continuity in word learning is central to a recent study by Frank, Goodman, Tenenbaum, and Fernald (2009). Frank et al. propose that by attributing utterance proximity to topic continuity, early word learners may be better able to aggregate information across multiple utterances and thereby make more effective inferences about speakers' referential intentions and the meanings of words. Their proposal is based on the generalization that, in a coherent discourse, utterances that are close in time are likely to refer to similar things. Frank et al.'s study showed that a discourse-continuity prior within a Bayesian word-learning model could provide enough information about probable topichood in child-directed speech to allow learning in otherwise ambiguous contexts.

In this paper, we follow this previous work by investigating the availability of topic-marking cues in child-directed speech. We ask (i) whether caregivers signal topicality in ways consistent with properties reported in adult-directed speech and (ii) how, over the course of a sequence of topically related utterances, caregivers' and children's behavior reflects joint attention. In the next section, we describe the video corpus we used and its annotations. We then introduce a variant of a Hidden Markov Model that we developed to identify discourses over time. We use these model-identified discourses to analyze how discourse markers change from the onset of a new topic over the course of subsequent utterances about a particular discourse topic (henceforth a *topical discourse* or simply a *discourse*). As we will show, lexical and syntactic properties of caregivers' speech undergo predictable

changes during a topical discourse, as do features related to the social interaction between caregiver and child.

Corpus & Annotation

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). A preliminary version of this corpus was analyzed by Frank et al. (2009), who selected it because the videos make it possible to identify both the objects being talked about and the objects present in the physical context. The play session settings are sufficiently restricted to permit annotation of the full set of alternative referents. The 24 available videos of English-speaking children range in length from 3 to 22 (M=12.2) minutes and contain between 56 to 397 utterances (M=202). Children in these videos fall into three age groups: 6 months (N=8), 11–14 months (N=8), and 18–20 months (N=8). Each video captures a single mother-child play session in which mothers were given several pairs of toys by the experimenter and asked to play with each pair for a 3-5 minute period.

The corpus was 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, restricted to the use of an object's name ("look at the *doggie*") or a pronoun ("look at *his* eyes and ears"). Excluded were cases in which the object was evoked only with a property like "red," a super-/subordinate terms like "animal", or a part term like "eye". Likewise, exclamations like "oh" were not judged to be referential, even if they were directed at an object.

We further added timestamp annotations that marked the onset time of each utterance. Because the goal of this study is to characterize topical discourses in child-directed speech, we excluded utterances that were not part of a minimal discourse, defined as at least 3 successive utterances on the same topic. The removal of utterances that did not participate in a minimal discourse was based on the smoothed topic assignments from the model described in the next section. One video was also excluded due to the limited amount of data available.

Model

Although the Frank et al. study identified topic continuity as a valuable information source in word learning, their treatment of topical discourse relied on a fairly coarse measure of topicality, based solely on the annotation of referential intention. For their analysis, a topical discourse consisted of any sequence of successive utterances that referred to the same entity. This approach may have both under-estimated and over-estimated the number of utterances that belong to particular topical discourses. 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 discourse. Alternatively, a long pause following

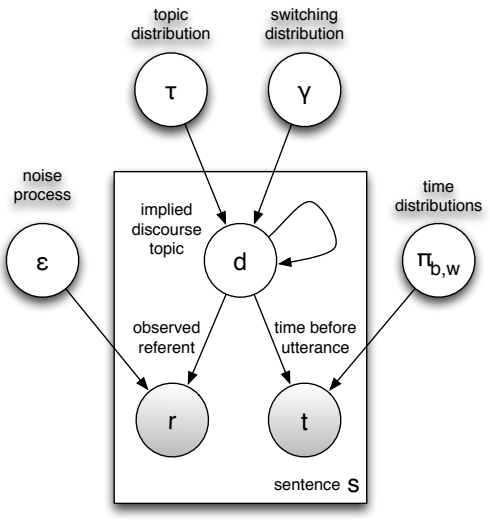


Figure 1: Schematic graphical model for the dependencies in our discourse-finding model.

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 discourse even if it mentions the referent of the previous discourse.

Figure 5 gives an example of this phenomenon. In this conversation, the mother occasionally pauses in her descriptions of the pig to encourage the child to look, saying "hi CHI" in order to bring his attention back to the pig. Simply identifying discourses as consistent sets of references to the same objects, as in Frank et al. (2009), may understate the continuity of these conversations. To investigate whether there were longer coherent discourse strings when these interruptions were taken into account, we created a probabilistic model designed to smooth across short interruptions to discover longer discourses. This model is not a cognitive model of discourse processing but instead a tool for data analysis, allowing us to identify discourse units in a principled way in order to examine corresponding linguistic and social cues.

Model details

To discover discourses, we created a variant of a Hidden Markov Model (HMM), shown in Figure 1. 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 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) discourse topic 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 discourse (switching process). If it starts a new discourse, 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 1 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 here, the model was run independently on the data for each video for 2000 Gibbs sweeps. Each sentence was assigned its modal discourse 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.

Model results

The resulting topic assignments from the model reduce the total number of topical discourses, in comparison with the number of discourses calculated with the raw reference annotations. The raw topics yield a total number of topical discourses per video that ranges from 8 to 65 (mean=31.6), whereas the model-assigned smoothed topics yield a total number per video that ranges only from 3 to 31 (mean=14.7).

Figure 2 shows a pair of histograms that display the differences in discourse length between the reference-annotated topics and the model-assigned topics. Discourses are considerably longer in the smoothed data. When we consider utterance onset times, we find that the gaps between utterance onsets are shorter within topical discourses than at discourse boundaries. Figure 3 shows the distribution of these gaps, based on the smoothed topic discourses.

To visualize the model results, Figure 4 shows a "Gleitman plot" (see Frank et al., 2009) for one video in the corpus. A sequence of green dots or a stretch of black bar indicates a topical discourse in the raw reference annotations or in the smoothed 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 had on the discourses: Topics extend through time even when intervening utterances do not reference the topic directly.

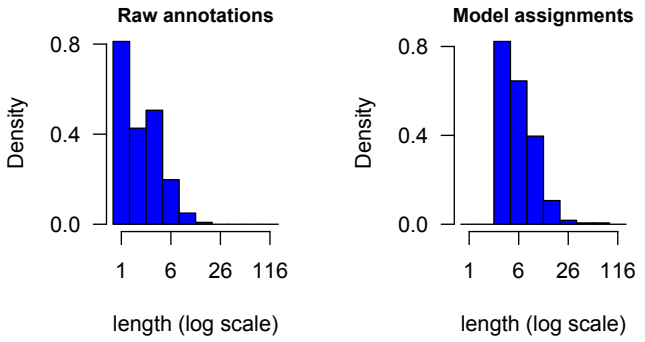


Figure 2: Mean discourse length (number of utterances). Left gives the distribution for raw annotations; right gives the distribution for the discourses found by the model. Model discourses are right-shifted because they were constrained to be 3 or more utterances in length.

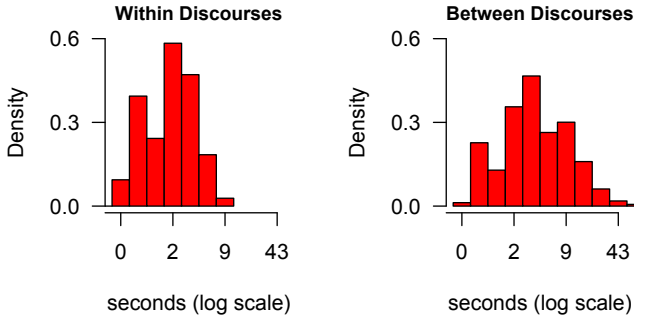


Figure 3: Mean time between utterance onsets. Left gives distribution of times between within-discourse utterances; right gives distribution of times at discourse boundaries.

Analyses of Topical Discourses

In order to determine what discourse markers change over the course of a topically related sequence of utterances, we consider caregivers' speech and social cues between caregiver and child. We analyze topical discourses generated from the raw topic annotations and the smoothed model-identified topics. The observed markers are modeled using mixed-effects multinomial regressions with random caregiver-specific and referent-specific intercepts (Gelman & Hill, 2007).

Figure 5 shows the behavior in the smoothed topic discourses of the 7 discourse markers we analyze. It also shows an excerpted transcript from one video in the corpus. In Table 1, we report the logistic- or linear-regression coefficient estimates and p-values for the factors child age (coded as *age*, a categorical factor) and utterance position within the topical discourse (coded as *time*, a numeric factor) and an interaction between the two. Both predictors were centered.

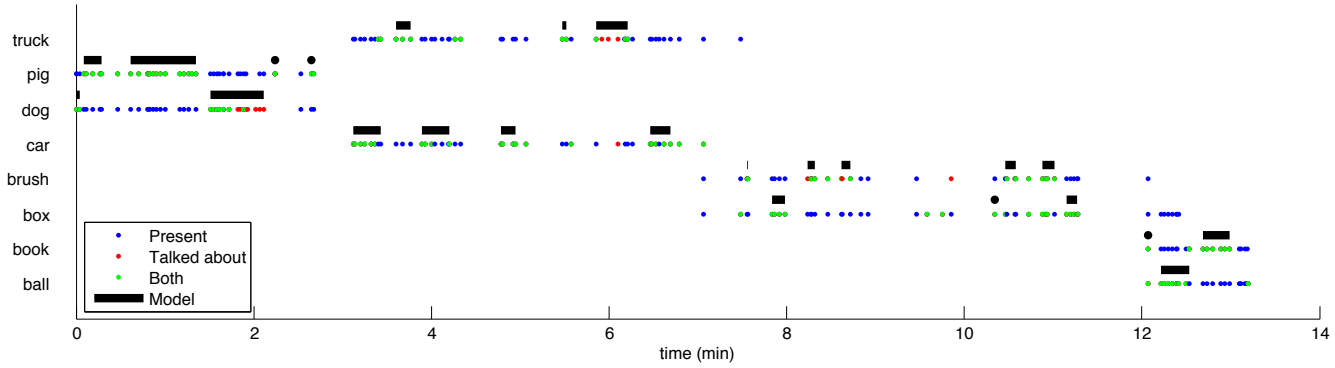


Figure 4: Sample Gleitman plot for a Fernald and Morikawa video. Rows denote objects; the x-axis marks time. Dots appear at utterance onset times; dot color reflects the raw video annotation of object presence and object reference. Blue denotes that the object was present when the utterance was uttered but not referenced; red denotes that the object was referenced but not present; green denotes that the object was present and referenced. The black bars show topical discourses identified by the model.

Discourse marker	β_{RAW}	p-val	β_{SMOOTH}	p-val
Pronoun use:				
time	0.156	*.005	0.280	*.001
age	-0.266	*.043	-0.170	.091
time×age	-0.035	.560	-0.049	.383
Sentence-final:				
time	-0.228	*.001	-0.338	*.001
age	-0.222	.153	-0.143	.318
time×age	-0.011	.855	-0.062	.300
Utterance length:				
time	-0.013	.818	-0.010	.848
age	0.140	.331	0.167	.356
time×age	-0.133	*.010	-0.073	.292
Children's eyes:				
time	0.000	.990	0.136	*.010
age	0.053	.604	0.085	.400
time×age	0.013	.827	0.039	.438
Children's hands:				
time	0.124	*.004	0.259	*.001
age	0.278	*.021	0.345	*.026
time×age	-0.106	.062	0.014	.795
Mother's points:				
time	0.094	.375	-0.139	.199
age	0.260	.206	0.235	.269
time×age	0.435	*.001	0.055	.598
Mother's hands:				
time	0.055	.323	0.050	.328
age	-0.137	.348	-0.085	.592
time×age	-0.100	.079	-0.025	.615

Table 1: Predictors for modeling discourse markers in mixed-effect models. The * marks significant predictor coefficients.

Lexical Trends Many studies have identified a correlation between the information status of a discourse entity and the linguistic form used to reference that entity, whereby familiar entities are more likely to be realized with pronouns, whereas unfamiliar entities tend to be realized with full noun phrases (Ariel, 1990). We therefore test whether the rate of pronominalization increases over the course of a topical discourse. The results confirm that in the raw and smoothed discourses, time is a significant factor for modeling the binary outcome of pronominalization, with more pronouns (3rd person nominative/accusative/possessive forms plus *one*) being used later in the discourse. The rate of pronominalization also varied by age with mothers of younger children using more pronouns overall, though this effect was marginal in the smoothed discourses. It is possible that caregivers use more adult-like rates of pronominalization with children who they believe are too young to be engaged in serious word learning.

Syntactic Trends The position of a referring expression within a sentence often 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). We therefore test whether references to a discourse topic occur later in a sentence in the early parts of a discourse. For this analysis, we consider the final word of each utterance with the prediction that topical entities are less likely to be referenced utterance-finally as a discourse progresses and an entity becomes more familiar. Fernald and Morikawa (1993) noted the strong prevalence of referential nouns at the ends of sentences in the English caregivers' speech, hence our choice to not test subject/object position. The results confirm that in the raw and smoothed discourses time is a significant factor for modeling the binary outcome of sentence-final mention, with fewer sentence-final references later in the discourse.

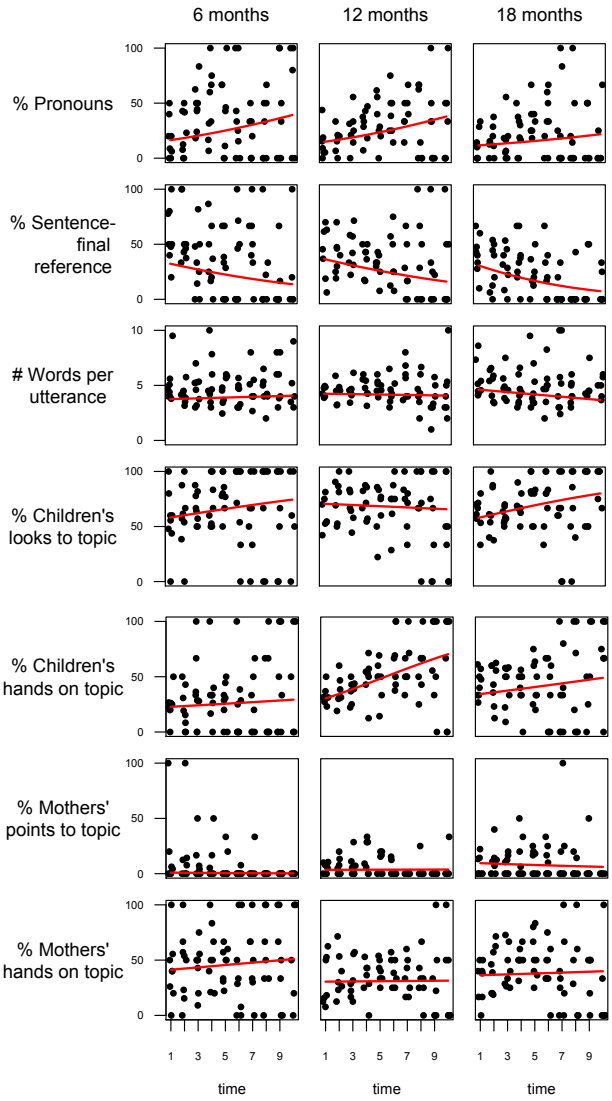
We also test whether sentence complexity increases as the topical discourse progresses. Information-theoretic models of language production (Genzel & Charniak, 2002; Levy & Jaeger, 2007) posit that, as discourse entities become more familiar, the processor is better equipped to handle longer and more complex structures. For this analysis, we measure complexity as mean utterance length to test the prediction that length increases over the course of a topical discourse. Contrary to prediction, the raw discourses showed an interaction between time and age whereby mothers of older children

slightly decreased their utterance length over time, and there were no reliable effects in the smoothed discourses. This lack of an effect 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 Figure 5).

Social Interaction Trends 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, Nagell, & Tomasello, 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 discourse, an effect apparent only in the smoothed discourses. Children also touched the referenced object more over the course of a discourse, an effect apparent in both the raw and smoothed discourses. It appears that children only gradually became engaged in the discourse, rather than shifting their attention immediately to the topic. For mothers' pointing, the raw discourses revealed an interaction between time and age whereby the rate of pointing to the topical object climbed most quickly for the oldest age group group. For mothers' hands, there were no reliable effects.

General Discussion

As one of the first 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 discourses, 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



Raw/Smooth	Utterance
pig/pig	this a pig
pig/pig	oink pig oh where you going
/pig	where you going
/pig	yes
pig/pig	what's he saying
/pig	look see
pig/pig	he's got little ears
pig/pig	see the eyes
pig/pig	can you see the piggy's eyes
/pig	what
/pig	CHILD look
/pig	ha
pig/pig	it squeaks
pig/pig	look it he's got a tail too
pig/pig	oh look i can put his hand in it
/pig	hi CHILD
pig/pig	look at his eye
pig/pig	can you see his eyes and his little nose
pig/pig	he gonna give you a kiss
/pig	yeah
pig/pig	what's this look
pig/pig	can you touch him
pig/pig	he's got fur
pig/pig	look it look at his tail
pig/pig	CHILD look at his tail
pig/pig	look at his tail
pig/pig	see his tail
pig/pig	its different huh
pig/pig	it feels different
pig/pig	there's a tail
dog/dog	can you say tail oh look at the puppy
dog/dog	look at the puppy
dog/dog	oh where'd it go

Figure 5: Discourse markers. Graphs on left plot caregiver means at successive utterance positions within smoothed discourses; points are jittered on the x-axis to avoid overplotting; superimposed regression lines correspond to logistic and linear models built with the smoothed-topic data. Topic assignments and transcript excerpt on right from a video with 12-month-old child.

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).

The benefit of our HMM modeling can be seen in the analysis of social cues, where we see a reliable effect of utterance position in children's looking in the smoothed discourses but not the raw discourses. 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.

As noted in the introduction, researchers studying processing in acquisition have tended to focus on word-level acquisition and, in so doing, have 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 remains unclear, but the progression of topical discourses may be apparent especially if children are hearing sentences with supporting context regarding what is present and what could be referred to.

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 growth of speaker-listener common ground over the course of a conversation (Clark, 1996). Our results are 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.

In sum, we take these exploratory results as an invitation to reconsider 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.

Acknowledgments

Thanks to Noah Goodman and Eve Clark for helpful discussion and Allison Kraus for research assistance. A Mellon postdoctoral fellowship to H. Rohde supported this research.

References

Altmann, G. T. M., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*, 191-238.

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

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

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.*

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)*.

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

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., & Fernald, A. (2009). Continuity of discourse provides information for word learning. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (p. 1418-1423).

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.*

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.

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.

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.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, *8*, 243-281.

Rohde, H., Levy, R., & Kehler, A. (in press). Anticipating explanations in relative clause processing. *Cognition.*

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.