Children's interpretation of ambiguous pronouns based on prior discourse

Manuel Bohn[1,2], Khuyen Nha Le[1], Benjamin Peloquin[1], Bahar Köymen[3], & Michael C. Frank[1]

[1] Department of Psychology, Stanford University

[2] Leipzig Research Center for Early Child Development, Leipzig University

[3] School of Health Sciences, University of Manchester

Author Note

Abstract

In conversation, individual utterances are almost always ambiguous, with this ambiguity resolved by context and discourse history (common ground). One important cue for disambiguation is the topic under discussion with a particular partner (e.g., "want to pick?" means something different in a conversation with a bluegrass musician vs. with a book club partner). Here, we investigated 2- to 5-year-old American English-speaking children's ($N = 131$) reliance on conversational topics with specific partners to interpret ambiguous or novel words. In a tablet-based game, children heard a speaker consistently refer to objects from a category without mentioning the category itself. In study 1, 3- and 4-year-olds interpreted the ambiguous pronoun "it" as referring to another member of the same category. In study 2, only 4-year-olds interpreted the pronoun as referring to the implied category when talking to the same speaker but not when talking to a new speaker. Thus, children's conception of what constitutes common ground in discourse develops substantially between age 2 and 5.

Keywords: Discourse, Pragmatics, Social cognition, Language development, Conceptual development

Children's interpretation of ambiguous pronouns based on prior discourse

Research highlights

- Conversational topics are used to organize discourse with specific partners
- Children make inferences about conversational topics from 3 years onwards
- From 4 years onwards, children see conversational topics as speaker specific
- Discourse can structure communicative interactions and thereby help children to resolve ambiguity and learn language

Introduction

Language learning and use is a fundamentally social behavior – when children hear ambiguous or novel words they can rely on social information to infer meaning (e.g. Bohn & Frank, 2019; Clark, 2009; Tomasello, 2008). One such information source is the social context of the utterance: as discourse unfolds, interlocutors build up common ground – a set of shared knowledge and beliefs – that serves as a background against which new utterances are interpreted (Bohn & Köymen, 2018; Clark, 1996).

Children's interpretation of pronouns provides a good case to investigate their reliance on common ground because pronouns represent a referent that is mentioned or implied in the previous discourse. While children have difficulties with interpreting pronouns until about age 6 (Ambridge & Lieven, 2011; Matthews, Lieven, Theakston, & Tomasello, 2009), recent evidence suggests that discourse structure helps children as young as 18 months to disambiguate pronouns. For example, children interpret the pronoun it as referring to a previously mentioned object – even when it was absent from the scene (Ganea & Saylor, 2007; Lidz, Waxman, & Freedman, 2003). In a similar way, 3 to 5-year olds use discourse context to interpret personal pronouns (Hartshorne, Nappa, & Snedeker, 2015; Song & Fisher, 2005). Beyond interpreting pronouns, children can also use discourse

structure and common ground to learn the meaning of novel words(e.g. Horowitz & Frank, 2015; Sullivan & Barner, 2016; Yuan & Fisher, 2009).

In most studies, the loose construct of "discourse" or "common ground" is operationalized as including the objects that have been explicitly mentioned. In conversation, however, discourse is often organized in terms of an overarching topic rather than a specific object referent. Further, this topic is not always stated explicitly but often has to be inferred based on the content of utterances (Clark, 1996). For example, when a child reports from a school trip "I saw an elephant, a tiger, a giraffe, and another one that was really scary" the parent can infer that another one refers to yet another animal. One study comes closer to this conception of discourse topic. Akhtar (2002) had children interact with an adult who repeatedly drew their attention to either the shape or the color of an object. When the adult later used a novel word to describe an object, children seemed to infer that the word referred to the property (color or shape) that was emphasized during the preceding discourse. In that study, children could solve the task by keeping track of which part of the objects they were jointly attending to (in the sense of looking at) with the speaker previously. Such visual anchoring, however, is not always possible: the topic may be absent and/or an abstract entity (such as animals in the example above). In such cases, the topic has to be inferred by reasoning about how the individual conversational bits relate to one another – how they are organized in conceptual terms.

Another neglected aspect of discourse inference, at least in experimental work, is its social nature. Common ground built up during discourse is often partner-specific, meaning that it is bound to this conversational partner and should not be assumed to be known by others. Common ground inferences therefore incorporate both conversational topic(s), but crucially also social information – the participants engaging in the conversation. Children can keep track of their conversational partners from a very young age (Bohn et al., 2018). In one relevant study, Saylor and Ganea (2007) had 14 to 20 month old children play ball with two different experimenters. Depending on who later asked for "the ball", children

retrieved the object that featured in their previous play with that person. From 3 years onwards, children use partner specific referential expressions (referential pacts), both with adults and peers (Köymen, Schmerse, Lieven, & Tomasello, 2014; Matthews, Lieven, & Tomasello, 2010). Interestingly, speaker identity influences children's interpretation of ambiguous utterances at an early point in language processing (Khu, Chambers, & Graham, 2020; San Juan, Khu, & Graham, 2015). However, these social inferences based on the identity of the conversational partners have rarely been studied in conjunction with discourse inferences. Thus, it is unclear whether children's discourse inferences involve social reasoning at all.

In the present set of studies, we investigated the impact of partner specific discourse topics on ambiguous pronoun resolution in 2- to 5-year-old children. First, we examine how children interpret ambiguous utterances in light of an overarching topic that guides conversation, but was never explicitly mentioned (and must be inferred in context). Second, we examine how their interpretation changes depending on the common ground shared with specific partners. In this way, we fill the gap between discourse topic inferences and partner-specific reasoning.

Our studies are presented as simple, storybook-like tablet games. In study 1, children interacted with a speaker who repeatedly mentioned objects from a particular category (e.g., horse and dog from the category animals). At a later time in the discourse, the speaker used the ambiguous pronoun it in the context of three new objects – one object from the category mentioned earlier (e.g., a cat) and two other objects (e.g., a hat and a train). Thus, we operationalized the topic as the category label – or the subordinate concept – because children are aware of object categories starting from infancy (Bornstein & Arterberry, 2010; Mareschal, Powell, & Volein, 2003). We found that children from age 3 onwards reliably interpret the ambiguous pronoun as referring to the object from the same category.

In study 2 we tested whether these judgments were specific to particular speakers. Children heard one speaker talk about objects from a particular category. Then, either the same or a different speaker used the ambiguous pronoun. We found that children adjusted their inference to the identity of the speaker from age 4 onwards.

Study 1

All experimental procedures, sample sizes and statistical analysis were pre-registered (see https://osf.io/9ypxn and https://osf.io/fyaxq). All deviations from the pre-registration are explicitly mentioned. The experimental material can be found in the associated online repository at https://github.com/manuelbohn/disCon.

Methods

Participants. Our final sample included data from 71 children, including 30 2-year-olds ($M = 2.63$, $range = 2.00$ - 2.98, 14 girls), 21 3-year-olds ($M = 3.56$, $range = 3.13$ - 3.97, 9 girls) and 20 4-year-olds ($M = 4.50$, $range = 4.00$ - 4.97, 9 girls). We tested a larger sample of 2-year-olds because we expected a weaker effect in this age group. In addition, 28 children were recruited but not included because they were reported to have less than 75% of English exposure at home (12 children); did not finish the task (10 children); were correct on less than 5/6 training rounds (3 children, see below), or equipment failed (3 children). All children were recruited from the floor of a Children's museum in San José, California, USA. The population from which this sample is drawn is characterized by diverse ethnic backgrounds (predominantly White, Asian, or mixed ethnicity) and relatively high levels of parental education and socioeconomic status. Parents consented to their children's participation and provided demographic information. Data was collected between January and September 2019. All experiments were approved by the Stanford Institutional Review Board (protocol no. 19960).
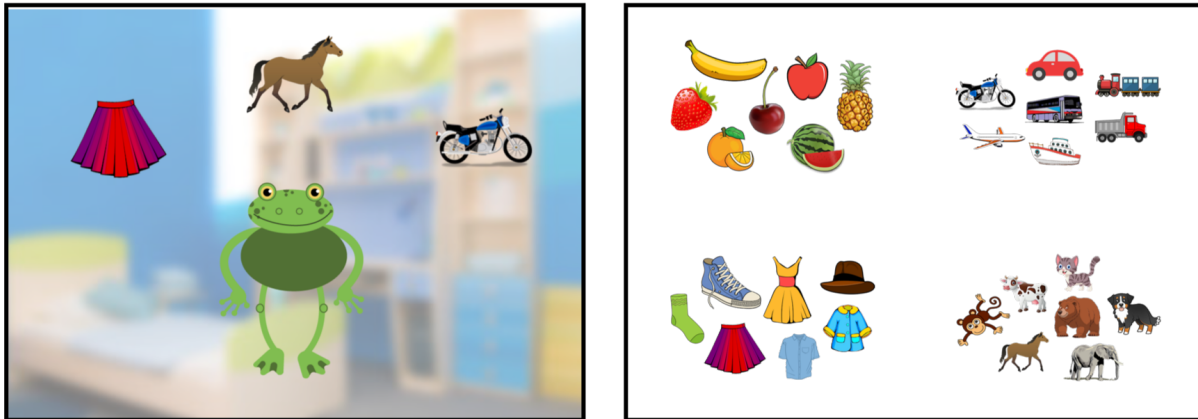
Figure 1. Left: Screenshot from the experimental setup. Right: Stimuli pictures for the four categories: fruits, vehicles, clothes and mammals.

Materials.   Study materials were presented as a picture book on a tablet computer (Frank, Sugarman, Horowitz, Lewis, & Yurovsky, 2016). Children responded by touching objects on the screen. Responses were automatically saved. We selected four categories of objects: animals, vehicles, clothes, and fruits. For each category, we chose pictures for seven different objects from that category (e.g. for vehicles: car, truck, train, bus, airplane, boat and motorbike). The right panel of Figure 1 shows all pictures used in the study grouped by category (see supplementary material for a list of the corresponding words). All utterances made by the different speaker animals were pre-recorded from native English speakers, with one speaker per animal.

Procedure.   The experimenter introduced the study as a visit to the house of the little animals, during which the animals would show the child the things they have at home. On each trial, children saw one animal (the speaker) in the middle of the screen with three objects above them (Figure 1, left). Each of the objects belonged to a different category. Trials started with six training rounds, in which the speaker named one of the three objects displayed above them and asked the child to touch it ("Look at that, can you touch the X?"). From one round to the next, the pictures changed but the three categories remained the same and the speaker consistently asked the child to touch the object from

one category. For example, on the first training round, children saw a skirt, a horse, and a motorcycle and the speaker asked "Can you touch the motorcycle?" On the second training round, children saw a jacket, a dog, and a bus and the speaker asked "Can you touch the bus?". After the six training rounds, children moved to the test and saw a new set of three (e.g. a hat, a cat and a train). However, instead of using a label, the speaker used an ambiguous pronoun to refer to one of the objects ("Look at that, can you touch it"). Children could identify the referent of the pronoun by assuming that the speaker continues to talk about objects from the same category as they did previously (vehicles in the example above). However, in order to do so, children had to infer the category by which all of the previous objects the speaker mentioned were grouped.

Children received four sets of trials (six training trials and one test trial) with a different category as the target on each one. The categories, the order of the objects within each category and the position of the target object (left, right, middle) were randomized across children.

Results

The dependent variable in all analyses was whether the object chosen at test was from the same category as the objects named throughout the training trials. All analyses were computed in R (R Core Team, 2018). As a first step, we aggregated responses across test trials for each child and compared the proportion of correct responses to a level expected by chance (33% correct) within each age bin. We used the function `ttestBF` from the R-package `BayesFactor` (Morey & Rouder, 2018) to compute a Bayes factor (BF) in favor of the hypothesis that performance is above chance (0.33 correct). Figure 2A shows the corresponding posterior distribution for each age bin. We found little evidence that 2-year olds performed above chance (mean proportion correct = 0.42, BF_{10} = 0.59) but found substantial evidence for 3-year-olds (mean proportion correct = 0.60, BF_{10} = 90.77) and 4-year-olds (mean proportion correct = 0.55, BF_{10} = 10.39). This result is robust to

changes in the prior on the standardized effect size (see supplementary material for details).

To analyze responses continuously across age we used logistic Bayesian generalized linear mixed models (GLMM) fit via the function `brm` from the R-package `brms` (Bürkner, 2017). All models had default priors and included random effects for participant and category[1], with random slopes for age. Inference was based on comparing models that differed in whether they included the key predictor of interest, which was age in study 1. Following McElreath (2016), we compared models using WAIC (widely applicable information criterion) scores and weights. The WAIC score is an indicator of the model's predictive accuracy for out of sample data; model's with lower scores are preferred. WAIC weights are an estimate of the probability that this model (compared to all other models considered) will make the best predictions on new data. In addition, we inspected the posterior distribution for the key parameters in the model via it's mean and 95 % credible interval (CI).

The model comparison favored the model including age as a predictor (Table 1). The mean model estimate for age was positive ($\beta = 0.34$, 95% CI = -0.54 - 1.27), suggesting an increase of performance with age (see Figure 2B). However, the small difference in model weights and the fact that the 95% CI for the model estimate also overlapped with zero speak against substantial developmental gains in the age range considered.

Discussion

This study presents evidence consistent with the idea that at least around age 3, children make inferences about conversational topics. Based on hearing a speaker consistently refer to objects from a certain category, children interpreted the ambiguous pronoun "it" as referring to an object of the same category. This suggests that children

---

[1] In the first preregistration (data collection with 3 and 4 year olds) we forgot to include random effects for category. All subsequent preregistration included it.
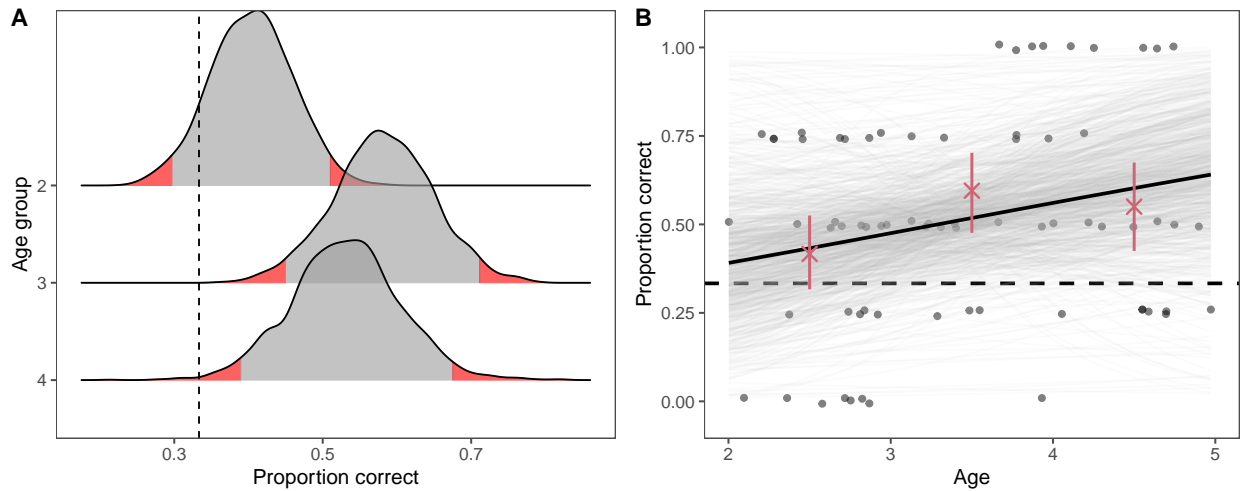
Figure 2. (A) Posterior probability distribution for the mean for each age bin based on one sample Bayesian t-test. Grey regions indicate 95% credible intervals for each age group. (B) Correct responses for age continuously. Transparent dots show data aggregated for each individual participant. Red crosses show mean within age bin with 95% confidence intervals based on non-parametric bootstrap of the data. Black line shows the mean of the posterior distribution of the model including age. Grey lines show 1000 random draws from the model posterior to depict uncertainty in the model. Dotted line indicates a level of performance expected by chance.

Table 1

Model comparison for Experiment 1

Predictors	WAIC	SE	weight
Age	381.75	9.47	0.57
Intercept only	382.34	9.06	0.43

Note. All models included random intercepts for participant and category and random slopes for age within category.

track common ground with a speaker not just in terms of remembering what has been talked about previously, but also in the form of an overarching topic that guides the conversation and allows predictions about what will be talked about in the future. In the supplementary material, we report an experiment that shows that adults make the same inference in a similar setup.

In a follow-up study, we tested whether the number of training trials affected children's ability to make the inference. We contrasted one training trial with six training trials. The results suggest that fewer training trials do not necessarily mean worse performance. However, the data were not conclusive and we therefore present the procedure and the analyses of this study in the supplementary material.

We also conducted an exploratory analysis (not pre-registered) to investigate whether the inference varied with the category that had to be inferred. Results point to differences between categories in that, for example, children across the entire age range were more likely to be correct when fruit is the target category compared to clothes. Details of this analysis are presented in the supplementary material. We return to this point in the general discussion. In the next study, we continued with the same number of training trials and categories and focused on whether the inference found in study 1 is partner specific.

Study 2

Children carried out the same task as in study 1. However, in half of the trials it was the same speaker who produced the ambiguous pronoun ("Can you touch it?") as the speaker who presented the training trials and in the other half of the test trials, it was a new speaker. Because we found limited evidence that 2-year-olds make category-based discourse inferences in study 1, we only tested 3- and 4-year-olds in study 2. The preregistration for this study can be found at https://osf.io/5e9pk; study materials are in the associated online repository.

Methods

Participants.   Sixty children, including 30 3-year-olds ($M = 3.48$, *range* $= 3.00$ - 3.98, 16 girls) and 30 4-year-olds ($M = 4.34$, *range* $= 4.00$ - 4.89, 12 girls) participated in this study. In addition, eight children were tested but not included because they had less than 75% English exposure at home (4 children); they did not finish the task (2 children), or the equipment failed (2 children). Data was collected in August and September of 2019. For details on population characteristics and ethical approval see study 1.

Materials.   Study materials were the same as in study 1.

Procedure

The general procedure was the same as in study 1 except for the following changes. The speakers were introduced as showing children their favorite things (in contrast to study 1, which used the category of things they have at home). We included this to focus children's attention on the individual speakers. Conversely, when making a request, speakers said: "I like that. Can you touch [object/it]". The speaker change manipulation was implemented in the following way: After the six training trials, the speaker announced that they had to leave and left the scene by walking off the left edge of the screen. After a short pause, either the same or a different animal returned to the scene, positioned themselves in the middle, and made a request using the ambiguous pronoun. If the same speaker returned, they entered from the same side as they left; if a new speaker appeared, they entered from the other side. This measure served to emphasize that the different speaker was new to the scene and therefore unfamiliar with the preceding discourse. If the speaker was the same as during the training trials, children could infer the referent of the pronoun in the same way as in study 1. However, if the speaker changed, no such inference was licenced if children assumed that the discourse topic is specific to a speaker.

Selection and positioning of objects and categories was randomized in the same way

Table 2

Model comparison for Experiment 2

Predictors	WAIC	SE	weight
Age * Condition	334.47	9.41	0.62
Age + Condition	336.69	8.16	0.20
Age	337.00	7.42	0.18

Note. All models included random intercepts for participant and speaker and random slopes for condition within participant and within condition and age within category.

as in study 1. Each child received four test trials, two with the same and two with a different speaker returning, in an order randomized for each child.

Results

We tested the effect of speaker change on children's discourse inferences via a model comparison. We compared a base model including only age as a fixed effect to models including speaker type, either as a main effect or as an interaction with age. Models were fitted and compared in the same way as in study 1. Models now also included random slopes for condition. The model comparison clearly favored the interaction model (Table 2). The interaction model was three times more likely than the other models considered to make better predictions on new data. The interaction term in the model itself was large and reliably different from zero ($\beta = 1.55$, 95% CI = 0.23 - 2.95). Figure 3 visualizes the data and the interaction model and shows that while younger children did not take into account speaker identity, older children (starting at around age 4) only interpreted the ambiguous
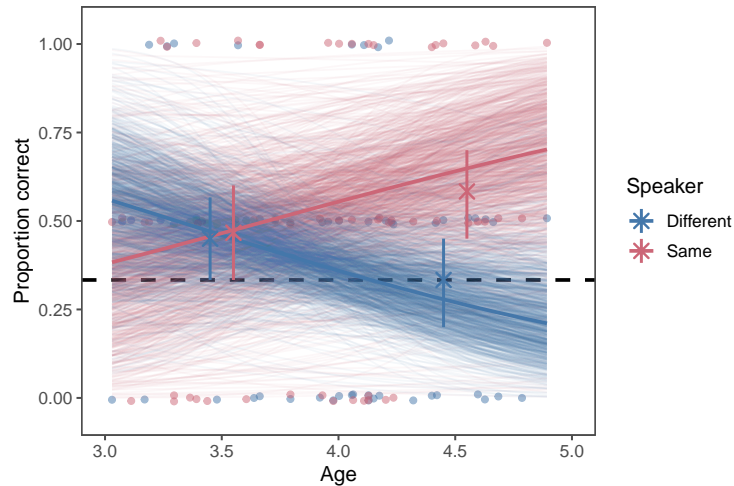
Figure 3. Correct responses in study 2 by age and condition. Transparent dots show aggregated data from individual participants. Blue and red crosses show data mean for by-year age bins with 95% confidence interval based on non-parametric bootstrap. Colored lines show the mean of the posterior distribution for each condition based on the interaction model. Lighter lines show 1000 random draws per condition from the model posterior to depict uncertainty. Dotted line indicates a level of performance expected by chance.

pronoun in light of the previous discourse topic when the speaker remained the same.

Discussion

In study 2, we found evidence that children from 4 years onward assume that a conversational topic is specific to the identity of the speaker. When one speaker repeatedly referred to objects from the same category, they expected the same speaker, but not a different one, to continue communicating about the same category. 3-year-olds, on the other hand, did not take into account speaker identity.

General Discussion

We studied how children use conversational topics with specific partners to interpret ambiguous pronouns. Our results suggest that, by age 3, children can identify the referent

of an ambiguous pronoun by inferring the topic of a conversation based on preceding discourse. Moreover, starting around age 4, children also associated this inference with specific partners. These findings illustrate how children can leverage the structure of the social interactions in which conversations take place to disambiguate utterances.

Our findings speak to preschoolers' ability to reason about conversational topics. Around age 3, children highlight the relevance of their utterance to the prior discourse by adjusting the informativeness of their referring expressions. For example, in English, children use a pronoun instead of a noun phrase to mark the topic (Matthews, Lieven, Theakston, & Tomasello, 2006; see also Chen, 2011; De Cat, 2009; Kurumada, 2009 for cross-linguistic evidence). However, these studies do not necessarily show whether children consider partner-specificity aspect of topic-marking: topic-marking is necessary with the conversational partners who are familiar with the topic, but not with new conversational partners who do not share this common ground. In fact, it has been shown that when children interact with a new partner, who is unaware of the topic, their referring expressions oftentimes do not change (Bannard, Rosner, & Matthews, 2017). Thus, our findings expand these and suggest that young children not only track the topic of a conversation but also pay attention to with whom they share this common ground.

We did not find substantial evidence in study 1 that children under the age of 3 are reliably able to infer the topic under discussion and use it to resolve ambiguity. However, we suspect that this result is not due to a lack of inferential abilities; instead, we suspect that it follows from difficulties with processing the information the inference was based on. In an exploratory analysis we compared children's performance with the different categories (see supplementary materials). We found that children across the entire age range were more likely to make the inference when *fruit* was the target category compared to when it was *clothes. Animals* and *vehicles* resulted in intermediate levels of performance. On the one hand, younger children might not conceptualize e.g. clothes as a cohesive category. On the other hand, the individual objects in some categories might be less familiar to children

and provide limited evidence for the category itself. From a developmental perspective, this pattern suggests that the main loci of change might well be children's vocabulary and their conceptual abilities, rather than their understanding of conversational topics. Of course, the exploratory nature of these results needs confirmation in future studies.

Study 2 showed more pronounced developmental effects. While 3-year-olds tended to choose the object from the same category for both speakers (see Figure 3), 4-year-olds only did so when the speaker was the same. This lack of sensitivity to speaker identity seems to conflict with earlier work which has found speaker-specific communicative adjustment in younger children (Bohn et al., 2018; Saylor & Ganea, 2007). However, in these studies, speakers only communicated about specific objects that were communicated about before. Children could identify the referent by remembering which speaker had communicated about which object. In our study, the objects at test had never been mentioned before. The connection to the prior discourse only lay in the relation of the objects to the topic under discussion. Thus, children not only needed to track what objects had been talked about but also had to infer that talking about these objects is a product of the particular speaker intending to talk about a specific topic.

Our setup provides a method to study the interplay between linguistic knowledge (necessary to make inferences about categories), conceptual knowledge (necessary to make inferences about additional category members) and social cognition (necessary to determine whether inferences are licensed). This interplay opens up questions about the processes involved in balancing and integrating these different information sources. Computational models provide a productive way to formalise theories of how information may be integrated. For example, Bohn and colleagues (2019) used a Bayesian model of pragmatic reasoning (Frank & Goodman, 2012) to formalise and predict the process by which children integrate information from common ground with information that is provided by an utterance in the moment. Such integration mechanisms could in principle be extended to make use of discourse topics as well (Luong, Frank, & Johnson, 2013;

Rohde & Frank, 2014).

An open question is how these results generalize outside the scope of the present context with data collected using tablet computers in a western, affluent study population. Regarding the first point, Frank and colleagues (2016) found similar results when children were tested in word recognition and word learning tasks either on a tablet or in a live interaction. Thus, we do not expect that the tablet format specifically should compromise the validity of the paradigm, but further work will be necessary to confirm that experimental findings apply to naturalistic conversational interactions. Regarding the second point, the cross-cultural generalizability of experimental pragmatics of the type we studied here is an open challenge for future research. While some semantic/pragmatic phenomena do appear to be general (Katsos et al., 2016), there are relatively few cross-cultural developmental studies on inference specifically, with some finding similar (Su, 2013; Su & Su, 2015; Zhao, Ren, Frank, & Zhou, 2019) and others different (Fortier, Kellier, Flecha, & Frank, 2018) trajectories compared to western settings.

Taken together, the studies reported here illustrate the development of children's ability to balance different layers of discourse. Thinking of discourse as organized by overarching topics allows listeners to predict what will be talked about next and, assuming that these predictions hold, increases the likelihood of successful comprehension. But discourse is also a form of social interaction and thus requires making inferences about the common ground shared between specific conversational partners. Here we show that children can use these processes to identify the referent of an ambiguous utterance. Because referent identification is a first step in language learning, we expect the same processes may also be recruited for language learning.

References

Akhtar, N. (2002). Relevance and early word learning. *Journal of Child Language*, *29*(3), 677–686.

Ambridge, B., & Lieven, E. V. (2011). *Child language acquisition: Contrasting theoretical approaches.* Cambridge University Press.

Bannard, C., Rosner, M., & Matthews, D. (2017). What's worth talking about? Information theory reveals how children balance informativeness and ease of production. *Psychological Science*, *28*(7), 954–966.

Bohn, M., & Frank, M. C. (2019). The pervasive role of pragmatics in early language. *Annual Review of Developmental Psychology*, *1*(1), 223–249.

Bohn, M., & Köymen, B. (2018). Common ground and development. *Child Development Perspectives*, *12*(2), 104–108.

Bohn, M., Tessler, M. H., Merrick, M., & Frank, M. C. (2019). Predicting pragmatic cue integration in adults' and children's inferences about novel word meanings. PsyArXiv. https://doi.org/10.31234/osf.io/xma4f

Bohn, M., Zimmermann, L., Call, J., & Tomasello, M. (2018). The social-cognitive basis of infants' reference to absent entities. *Cognition*, *177*, 41–48.

Bornstein, M. H., & Arterberry, M. E. (2010). The development of object categorization in young children: Hierarchical inclusiveness, age, perceptual attribute, and group versus individual analyses. *Developmental Psychology*, *46*(2), 350.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Chen, A. (2011). Tuning information packaging: Intonational realization of topic and focus in child dutch. *Journal of Child Language*, *38*, 1055–1083.

Clark, E. V. (2009). *First language acquisition.* Cambridge: Cambridge University Press.

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

De Cat, C. (2009). Experimental evidence for preschoolers' mastery of "topic". *Language Acquisition, 16*(4), 224–239.

Fortier, M., Kellier, D., Flecha, M. F., & Frank, M. C. (2018). Ad-hoc pragmatic implicatures among shipibo-konibo children in the peruvian amazon. PsyArXiv. https://doi.org/10.31234/osf.io/x7ad9

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336*(6084), 998–998.

Frank, M. C., Sugarman, E., Horowitz, A. C., Lewis, M. L., & Yurovsky, D. (2016). Using tablets to collect data from young children. *Journal of Cognition and Development, 17*(1), 1–17.

Ganea, P., & Saylor, M. M. (2007). Infants' use of shared linguistic information to clarify ambiguous requests. *Child Development, 78*(2), 493–502.

Hartshorne, J. K., Nappa, R., & Snedeker, J. (2015). Development of the first-mention bias. *Journal of Child Language, 42*(2), 423–446.

Horowitz, A. C., & Frank, M. C. (2015). Young children's developing sensitivity to discourse continuity as a cue for inferring reference. *Journal of Experimental Child Psychology, 129*, 84–97.

Katsos, N., Cummins, C., Ezeizabarrena, M.-J., Gavarró, A., Kraljević, J. K., Hrzica, G., . . . others. (2016). Cross-linguistic patterns in the acquisition of quantifiers. *Proceedings of the National Academy of Sciences, 113*(33), 9244–9249.

Khu, M., Chambers, C. G., & Graham, S. A. (2020). Preschoolers flexibly shift between speakers' perspectives during real-time language comprehension. *Child Development, 91*(3), e619–e634.

Köymen, B., Schmerse, D., Lieven, E., & Tomasello, M. (2014). Young children create
    partner-specific referential pacts with peers. *Developmental Psychology*, *50*(10),
    2334–2342.

Kurumada, C. (2009). The acquisition and development of the topic marker wa in l1
    japanese. In H. O. R. Corrigan E. A. Moravcsik (Ed.), *Formulaic language* (Vol. 2,
    pp. 347–374). Amsterdam, the Netherlands: John Benjamins.

Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but
    couldn't have learned: Experimental evidence for syntactic structure at 18 months.
    Cognition, *89*(3), 295–303.

Luong, M.-T., Frank, M. C., & Johnson, M. (2013). Parsing entire discourses as very long
    strings: Capturing topic continuity in grounded language learning. *Transactions of
    the Association for Computational Linguistics*, *1*, 315–326.

Mareschal, D., Powell, D., & Volein, A. (2003). Basic-level category discriminations by
    7-and 9-month-olds in an object examination task. *Journal of Experimental Child
    Psychology*, *86*(2), 87–107.

Matthews, D., Lieven, E., Theakston, A., & Tomasello, M. (2006). The effect of perceptual
    availability and prior discourse on young children's use of referring expressions.
    Applied Psycholinguistics, *27*(3), 403–422.

Matthews, D., Lieven, E., Theakston, A., & Tomasello, M. (2009). Pronoun co-referencing
    errors: Challenges for generativist and usage-based accounts. *Cognitive Linguistics*,
    20(3), 599–626.

Matthews, D., Lieven, E., & Tomasello, M. (2010). What's in a manner of speaking?
    Children's sensitivity to partner-specific referential precedents. *Developmental
    Psychology*, *46*(4), 74–760.

McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in R and*

Stan (pp. xvii, 469). Boca Raton: CRC Press.

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of bayes factors for common designs.* Retrieved from https://CRAN.R-project.org/package=BayesFactor

R Core Team. (2018). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rohde, H., & Frank, M. C. (2014). Markers of topical discourse in child-directed speech. *Cognitive Science*, *38*(8), 1634–1661.

San Juan, V., Khu, M., & Graham, S. A. (2015). A new perspective on children's communicative perspective taking: When and how do children use perspective inferences to inform their comprehension of spoken language? *Child Development Perspectives*, *9*(4), 245–249.

Saylor, M. M., & Ganea, P. (2007). Infants interpret ambiguous requests for absent objects. *Developmental Psychology*, *43*(3), 696–704.

Song, H., & Fisher, C. (2005). Who's "she"? Discourse prominence influences preschoolers' comprehension of pronouns. *Journal of Memory and Language*, *52*(1), 29–57.

Su, Y. E. (2013). Scalar implicatures and downward entailment in child mandarin. *Journal of East Asian Linguistics*, *22*(2), 167–187.

Su, Y. E., & Su, L.-Y. (2015). Interpretation of logical words in mandarin-speaking children with autism spectrum disorders: Uncovering knowledge of semantics and pragmatics. *Journal of Autism and Developmental Disorders*, *45*(7), 1938–1950.

Sullivan, J., & Barner, D. (2016). Discourse bootstrapping: Preschoolers use linguistic discourse to learn new words. *Developmental Science*, *19*(1), 63–75.

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

Yuan, S., & Fisher, C. (2009). "Really? She blicked the baby?" Two-year-olds learn combinatorial facts about verbs by listening. *Psychological Science*, *20*(5), 619–626.

Zhao, S., Ren, J., Frank, M. C., & Zhou, P. (2019). The development of quantity implicatures in mandarin-speaking children. PsyArXiv. https://doi.org/10.31234/osf.io/e8qc5