What's she talking about? Category based discourse inferences in young children

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

[1] Stanford University

[2] Leipzig University

[3] University of Manchester

Author Note

We thank Megan Merrick and Sabina Zacco for their help with the data collection. MB received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 749229.

Correspondence concerning this article should be addressed to Manuel Bohn, Jahnallee 59, 04109 Leipzig, Germany. E-mail: manuel.bohn@uni-leipzig.de

Abstract

tbd...

Keywords: keywords

Word count: X

What's she talking about? Category based discourse inferences in young children

Experiment 1

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

Participants

We obtained valid data from 71 children, including 30 2-year-olds (mean = 2.63, range = 2.00 - 2.98, 14 girls), 21 3-year-olds (mean = 3.56, range = 3.13 - 3.97, 9 girls) and 20 4-year-olds (mean = 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, 12 children were recruited but not tested because their parents reported less than 75% of English exposure at home. Ten children started the experiment but did not finish it because they became impatient (7) or the equipment broke (3). Three children were tested but excluded because they were correct in less than 5/6 training trials (see below). 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 background (predominantly self identifying as White, Asian or of mixed ethnicity) and high levels of parental education and socioeconomic status. Parents gave informed consent and provided demographic information. Data was collected between January and September 2019. All experiments reported in this paper were approved by the Stanford Institutional Review Board (protocol no. 357 19960).

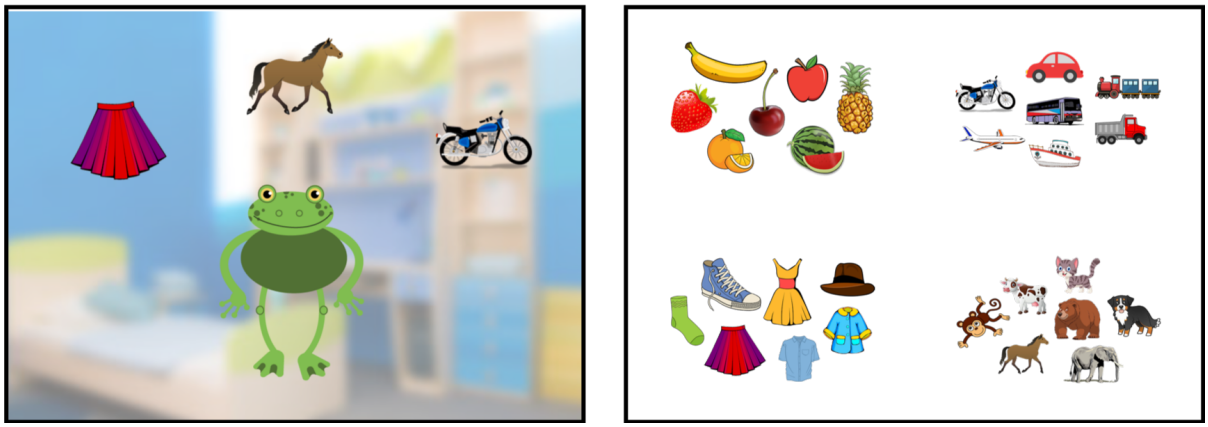

Figure 1. Left: Screenshot from the experimental setup. Right: Stimuli pictures for the four categories: fruits, vehicles, clothes and mammals.

Method

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. The experimenter guided children through the procedure and read out general instructions. The study was framed as visit to the house of the little animals during which the animals would show the child the things they have at home. Utterances made by the different animals were pre-recorded from native English speakers, with one speaker per animal. On each trial, children saw one animal in the middle of the screen with three objects above them (Figure 1, left). Each objects came from a different category (mammals, vehicles, clothes and fruits). For each category, we selected pictures of seven different category members (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.

The trial started with six training rounds, in which the animal named one of the objects above them, asking the child to touch it (e.g. “Look at that, can you touch the horse”). From one round to the next, the pictures changed but the categories remained the

same. For example, children saw a skirt, a horse and a motorcycle on the first training round and a jacket, a dog and a bus on the second. During training, the speaker consistently named objects from the same target category. After six training rounds, children received a test round in which the speaker used an ambiguous pronoun to refer to one of the objects ("Look at that, can you touch *it*"). Categories were randomly selected at the beginning of each trial and so was the order of pictures within each category. The position of each picture (left, right middle) was also randomly determined on each round. Children received four trials, one with each category as the target.

Results

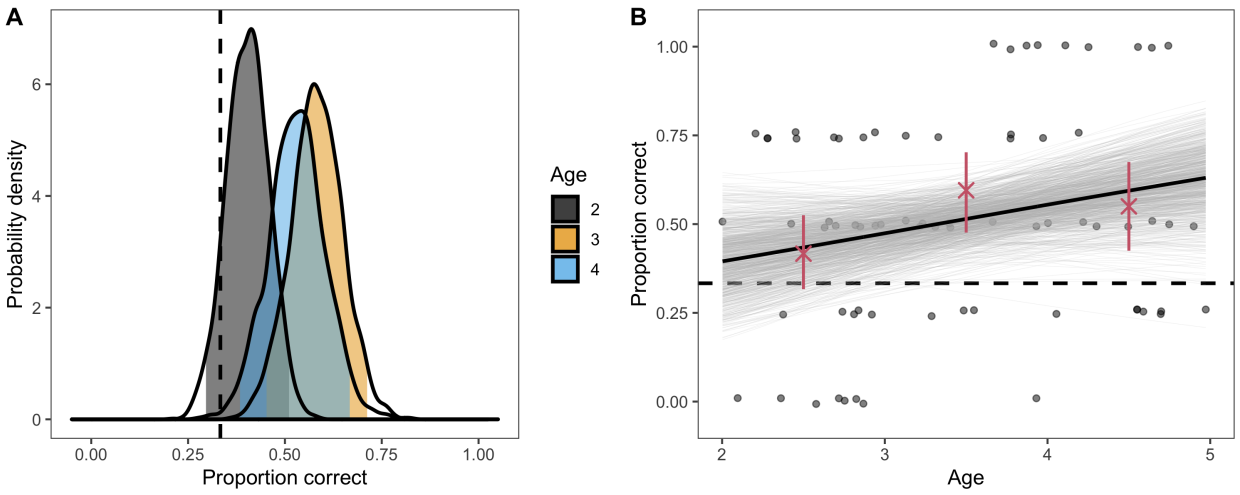

Figure 2. A: Posterior probability distribution for the mean for each age bin based on one sample Bayesian t-test. Shaded regions indicate 95% credible intervals for each age bin. B: Correct responses for age continously. Transparent dots show data aggregated for each individual participants. Red crosses show mean within age bin with 95% confidence intervals based on non-parametric bootstrap. 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 level of performance expected by chance.

The dependent variable in all analysis was whether the touched object at test was from the same category as the objects named throughout the training rounds. All analysis were computed in R (R Core Team, 2018). As a first step, we aggregated responses across 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. 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, $BF_{10} = 10.39$)[1].

To analyse responses continuously across age we used 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 id and speaker. Inference was based on comparing models that differed in whether they included the key predictor of interest, in this case age. 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).

For experiment 1, the model comparison favored the model including age as a predictor (Table 1). The mean model estimate for age was positive ($\beta = 0.32$, 95% CI = -0.08 - 0.75),

[1]This result is robust to changes in the prior on the standardized effect size. Choosing a wider prior results in slightly smaller Bayes factors, see online repository for details

Table 1

Model comparison for Experiment 1

Predictors	WAIC	SE	weight
Age	387.28	7.67	0.56
Intercept only	387.78	7.13	0.44

Note. All models included random intercepts for participant and speaker.

suggesting an increase in performance with age (see also 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 experiment presents evidence that children make inferences about conversational topics. Based on hearing a speaker consistently refer to objects from a certain category, children first inferred this category and then used it to interpret an ambiguous pronoun in a subsequent utterance. This suggests that children track common ground with a speaker not just in terms of remembering what has been talked about previously, but also in form of an overarching topic that guides the conversation and allows predictions about what will be talked about in the future.

In a follow-up experiment we tested whether the number of training rounds affected children's ability to make the inference. WE contrasted one training round with six training rounds. The results suggest that fewer training rounds do not necessarily mean worse performance. However, the data is not conclusive and we therefore present the details of this

experiment in the supplementary material. In the next experiment, we focus instead on whether this inference is conditional on the identity of a speaker.

Experiment 2

Here we test whether children expect a conversational topic to be specific to a speaker. Because we only found limited evidence that 2-year-olds make category based discourse inferences in experiment 1, we only tested 3- and 4-year-olds in Experiment 2. The preregistration for this experiment can be found at https://osf.io/5e9pk and the study material are in the associated online repository.

Participants

We tested 60 children, including 30 3-year-olds (mean = 3.48, range = 3.00 - 3.98, 16 girls) and 30 4-year-olds (mean = 4.34, range = 4.00 - 4.89, 12 girls). Four additional children were recruited but not included because parents reported less than 75% English exposure at home. Another four children were tested but ended up not contributing data because the equipment failed (2) or because they did not finish the experiment (2). Data was collected in August and September of 2019. For details on population characteristics and ethical approval see experiment 1.

Method

The general setup and procedure were the same as in experiment 1 except for the following changes. The animals were introduced as showing children their *favorite* things (experiment 1: 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

Table 2

Model comparison for Experiment 2

Predictors	WAIC	SE	weight
Age * Condition	325.25	10.45	0.59
Age + Condition	327.13	9.52	0.23
Age	327.66	8.94	0.18

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

way: After the six training rounds, the speaker announced that they had to leave and left the scene by walking off the left edge of the screen. Then, either the same or a different animal returned to the scene 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. We tested conditions within participant in a randomized order. Each child received four trials, two with the same and two with a different speaker returning.

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 also including condition, either as a main effect or in form of an interaction with age. Models were fitted and compared in the same way as in experiment 1. The model comparison clearly favored the interaction model (Table 2). The interaction term in the model itself was large and reliably different from zero ($\beta = 1.77$, 95% CI = 0.23 - 3.50). Figure 3 visualizes the

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

Discussion

In Experiment 2, we found evidence that children from four 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. Younger children did not take into account speaker identity.

General Discussion

References

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

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.

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.

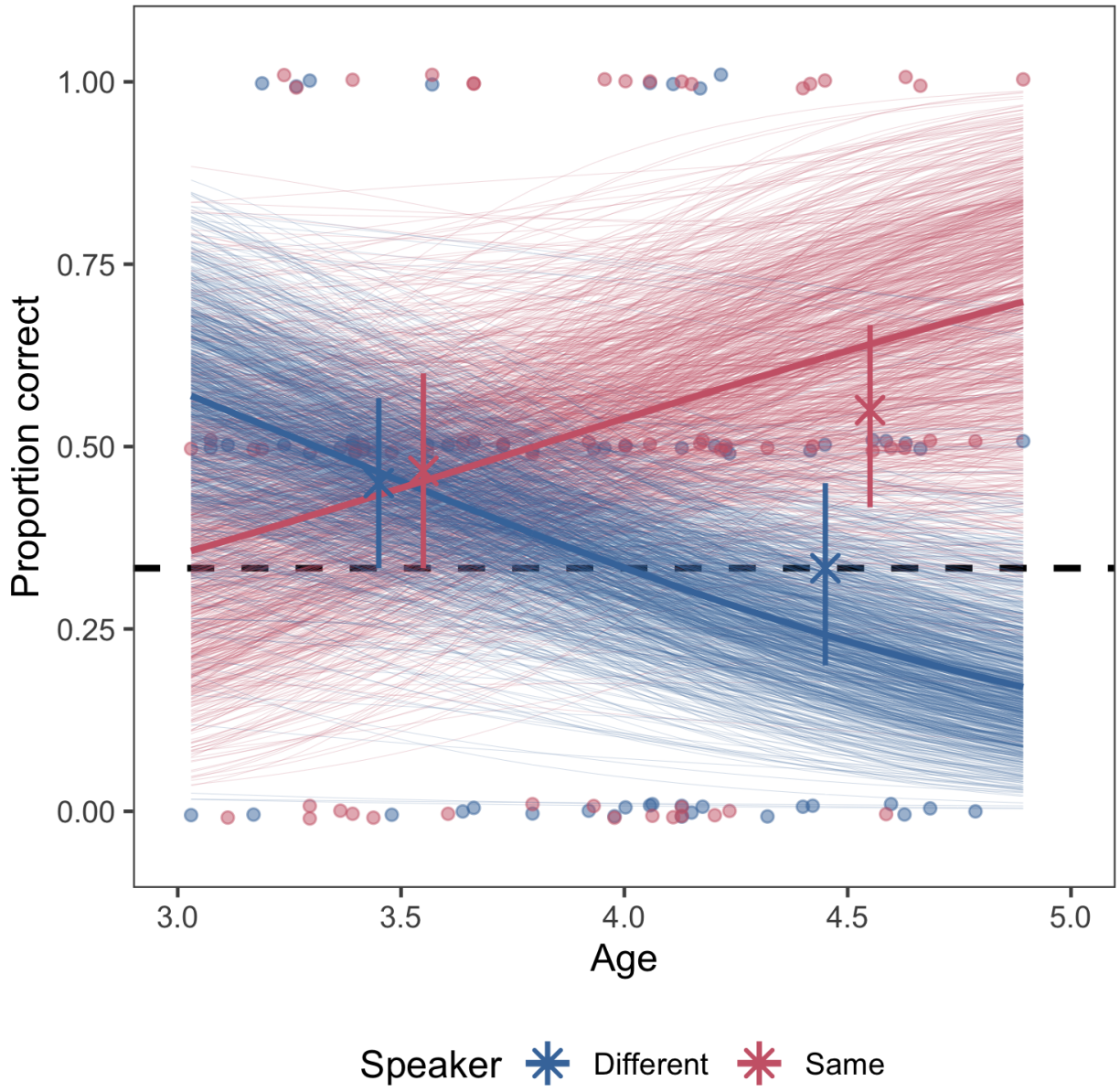

Figure 3. Correct responses in experiment 2 for age continously by age and condition. Transparent dots show aggregated data from individual participants. Blue and red crosses show mean within age bin based on aggregated data with 95% CI based on non-parametric bootstrap. Colored line shows the mean of 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 level of performance expected by chance.