Predicting the integration of social information in pragmatic word learning across development

Manuel Bohn[1,2], Michael Henry Tessler[3], & Michael C. Frank[1]

[1] Department of Psychology, Stanford University

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

[3] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Author Note

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, Department of Psychology, 450 Serra Mall, Stanford, CA 94305. E-mail: bohn@stanford.edu

Abstract

Enter abstract here. Each new line herein must be indented, like this line.

Keywords: language acquisition, social cognition, pragmatics, Bayesian modeling

Word count: X

Predicting the integration of social information in pragmatic word learning across development

Introduction

. . .

Adults

Assessment of general and speaker specific expectations (Experiment 1 and 2)

As a first step, we assessed whether adults make general and common ground inferences when tested separately. In experiment 1, participants could infer which object a novel word referred to by assuming that the speaker communicated in an informative way. The speaker was located between two tables, one with two novel objects, A and B, and the other with only object A (Fig. 1A). When the speaker turned to the table with the two objects (A and B) and used an novel word to request one of them, participants could infer that the word referred to object B. This follows from the counter-factual inferences that, if the (informative) speaker would have wanted to refer to object A, they would have turned to the table with the single object, because this would have been the least ambiguous way to refer to this object. As a consequence, turning to the table with two objects most likely reflects an intention to refer to object B. In the control condition, both tables contained both objects and no inference could be made based on the speaker's behavior.

Participants made the inference consistent with the assumption that the speaker communicated informatively: They selected object B above chance (0.5) in the test condition (t(39) = 5.51, $p < .001$, see Fig. 2) and more often compared to the control condition ($\beta =$ 1.49, se = 0.50, $p = .003$).

In experiment 2, we tested if participants use common ground information that is specific to a speaker to infer the referent of a novel word. In the preference manipulation (Fig. 1B . left), the speaker expressed preference for one of two objects. When the speaker

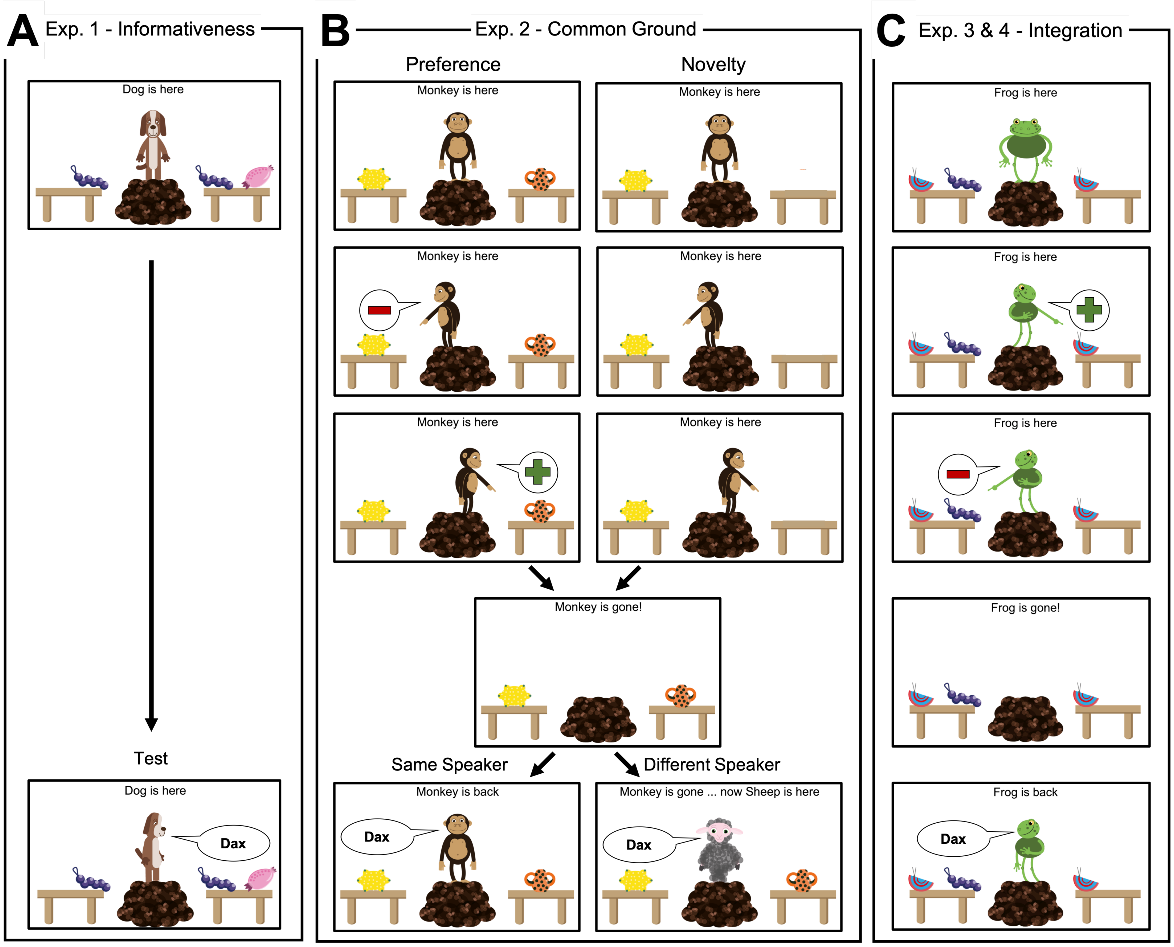

Figure 1. Schematic experimental procedure with screenshots from the adult experiments. In all conditions, at test (bottom), the speaker ambiguously requested an object using a non-word (e.g. "dax"). Participants clicked on the object they thought the speaker referred to. Informativeness (Experiment 1, left) translated to making one object less frequent in context. Common ground (Experiment 2, middle) was manipulated by making one object prefered by or new to the speaker. Green plus signs represent utterances that expressed preference and red minus of dispreference (see main text for details). Experiment 3 (right) combined manipulations. When expressing e.g. preference for an object on a table with two objects (panel 3 from top), the respective object was temporarily enlarged. Condition for Experiment 3 shown here: preference - same speaker - incongruent.

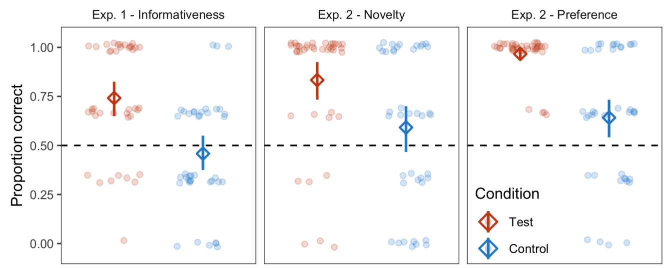

Figure 2. Results from experiment 1 and 2 for adults. For preference and novelty, control refers to a different speaker (see Fig. 1B). Transparent dots show data from individual participants, diamonds represent condition means, error bars are 95% CIs. Dashed line indicates performance expected by chance.

later used a novel word to requested an object, participants could infer the referent by assuming that the speaker was talking about the object they liked. In a control condition, a different speaker, whose preferences were unknown, made the request. Adults selected the preferred object above chance when the same speaker made the request (t(39) = 29.14, $p <$ .001, see Fig. 2) and more so compared to when a different speaker asked for an object ($\beta =$ 2.97, se = 0.70, $p < .001$).

In the novelty manipulation (Fig 1B - right), the speaker encountered one object on one of the tables. Later, a second object appeared. The second object was therefore new to the speaker. When the same speaker later used a novel word to requested an object (in a slightly excited way), participants could infer that it was the new object they are referring to. In contrast, when a different speaker, tho whom both objects were equally new, made a request, no inference could be made based on the speaker's state of knowledge. In line with this reasoning, participants selected the new object above chance with the same speaker (t(39) = 6.77, $p < .001$, see Fig. 2), and more often compared to the different speaker ($\beta =$ 1.70, se = 0.39, $p < .001$). Taken together, experiment 1 and 2 showed that adults use general and speaker specific expectations to identify the referent of a novel word. In the following, we studied how these expectations are integrated when combined experimentally.

Model predictions for information integration evaluated against new data (Experiment 3).

We modeled the integration of general and common ground expectations as a process of probabilistic inference. We used the results of experiment 1 and 2 to inform key parameters in the model's architecture. Within the RSA framework, general expectations that speakers (or listeners) communicate in a cooperative and informative way are reflected in the model's recursive structure. When interpreting an utterance, the hypothetical listener imagines that it was generated by a pragmatic speaker whose goal is to produce the most informative utterance. The pragmatic speaker is modeled as reasoning about what the most informative utterance would be for a naive listener, who does not know the novel words. Importantly, the utterance in this case includes not just the novel word but also the turning to one of the tables. The conditional probability that the listener interprets an utterance as referring to a particular referent is defined as:

$$P_L(r_s|u) \propto P_S(u|r_s)P_S(r_s)$$

Here, the term $P_S(u|r_s)$ denotes the likelihood that the pragmatic speaker will produce a particular utterance u to refer to a referent r. It is defined in terms of a utility function $U_S(u;s)$ consisting of the surprisal of u for a naive listener, who interprets u according its literal semantics:

$$P_S(u|r_s) \propto exp(\alpha U_S(u;s))$$

This utility is maximized in order to decide which utterance to use to communicate about a particular referent, reflecting the expectation that the speaker communicates in an informative way. The absolute strength of $P_S(u|r_s)$ depends on a scalar value, α, which can be interpreted as an indicator of how rational the speaker is in choosing utterances (i.e. as α increases, the speaker is more likely to choose the most informative utterance). We used the data from experiment 1 to specify how rational participants thought speakers were in the current setup. That is, we inferred which value of α would generate model predictions (assuming equal prior probability for each object) that corresponded to the average proportion of correct responses measured in experiment 1.

We treated common ground expectations as changes in the prior probability that a particular referent is being referred to ($P_S(r_s)$ in the first equation). For example, if the speaker expresses preference for object A, and the same speaker later requests an object, then A has a higher probability of being the referent of a subsequent utterance, regardless of any additional pragmatic considerations. We used the proportions measured in experiment 2 to specify the probability distribution across referents for the different common ground manipulations (novelty or preference, same or different speaker).

With α and $P_S(r_s)$ fixed based on experiment 1 and 2, we generated model predictions for situations in which general and common ground expectations had to be integrated (Fig. 1C). Expectations could either be congruent or incongruent. If, for example, object A was the more informative object (unique to that table) and also the one the speaker expressed preference for, expectations were congruent. Next, the speaker who made the request could either be the same as the one in the previous interaction or a different one. Finally, common ground information could be manipulated in the form of preference or novelty. This resulted in a total of 2 (novelty or preference) x 2 (same or different speaker) x 2 (congruent or incongruent) = 8 conditions for which we generated model predictions. In order to capture that behavioral data is to some extend noisy, all models included a noise parameter. Noise parameters ranged between 0 and 1 and reflect the proportion of responses that are estimated (based on the data) to be random instead of in line with model predictions. Including the noise parameter greatly improved the model fit to the data (see SI Appendix for details). We did not preregister the inclusion of a noise parameter for experiment 3 but did so for all subsequent experiments.

In experiment 3, we collected new behavioral data for each of the eight conditions.

Experimentally, we combined the procedures of experiment 1 and 2. The test setup was identical to experiment 1, however, before making a request, the speaker interacted with the objects so that some of them were preferred by or new to them (Fig. 1C). We discuss and visualize the results as the proportion with which participants chose the more informative object (i.e., the object that would be the more informative referent when only general expectations are considered). Before comparing responses to model predictions we fit a generalized linear mixed model (GLMM) to the data to test if participant's responses varied across conditions. Participants distinguished between congruent and incongruent trials when the speaker remained the same (model term: `alignment x speaker`; β = -2.90, se = 0.56, p < .001).

Model predictions were highly correlated with the average response in each condition (Fig. 3A and B). We compared the fit of the pragmatics model described above to two alternative models: the *flat prior* model ignored the speaker specific expectation and the *prior only* model ignored the informativeness inference. This analysis tests whether the trade-off between expectations in the pragmatics model captured the structure in the data better compared to models that assume that participants only consider one type of expectation at a time. Model fit was much better for the pragmatic compared to the flat prior (Bayes Factor (BF) = 4.2e+53) or the prior only model (BF = 2.5e+34), suggesting that participants considered and integrated both types of expectations. The estimated proportion of random responses according to the pragmatics model was 0.30 (95% Highest Density Interval (HDI): 0.23 - 0.36). This value was considerably lower for the pragmatics model, compared to the alternative models (see SI Appendix), lending additional support to the conclusion that the pragmatics model better captured the behavioral data.

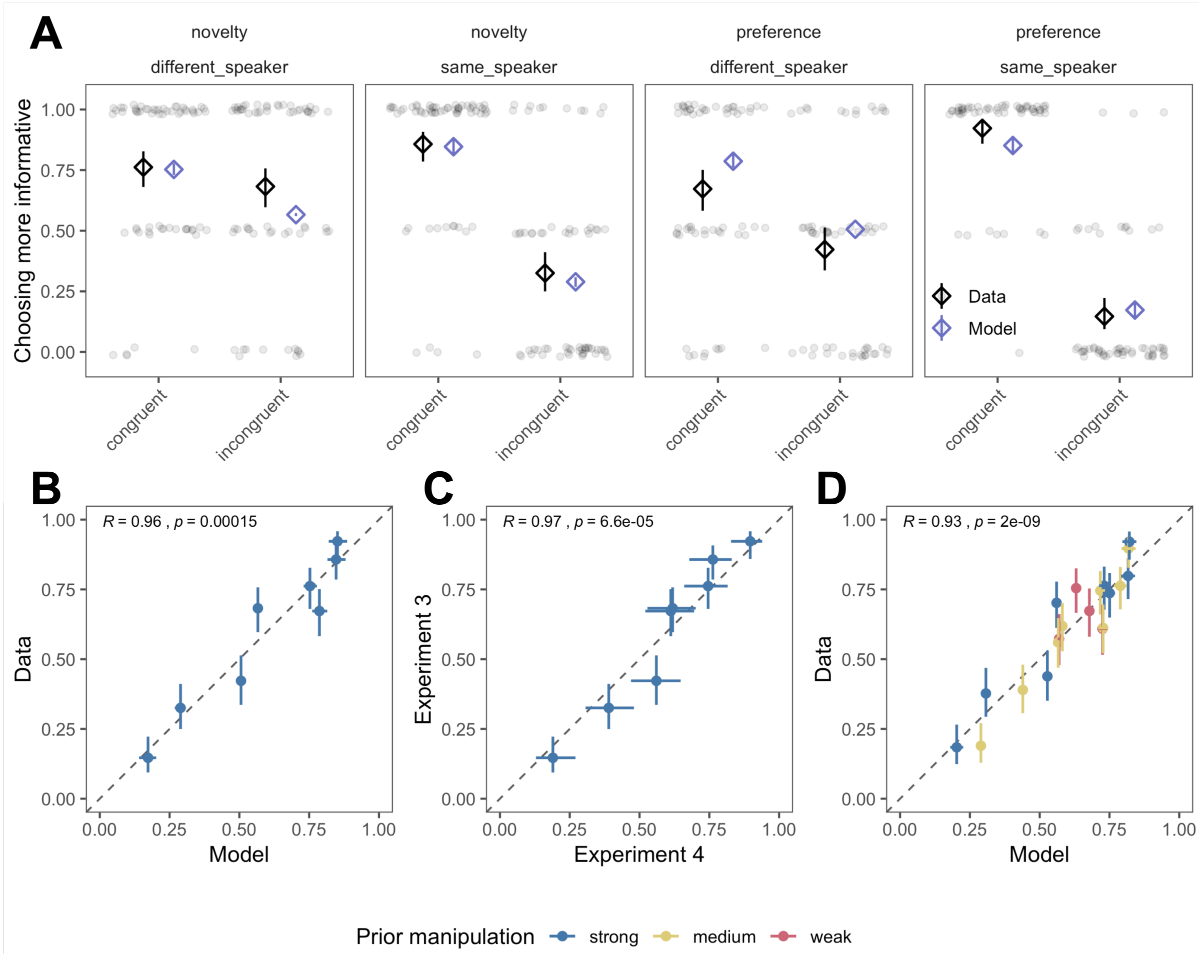

Figure 3. Results from experiment 3 and 4 for adults. Data and model predictions by condition for experiment 3 (A). Transparent dots show data from individual participants, diamonds represent condition means. Corrrelation between model predictions and data in experiment 3 (B), between data in experiment 3 and data for the strong prior manipulation in experiment 4 (direct replication - C) and between model predictions and data in experiment 4 (D). Coefficients and p-values are based on Pearson correlation statistics. Error bars represent 95% HDIs.

Replication and extension to different levels of speaker specific expectation (Experiment 4).

The goal of experiment 4 was twofold: First, to replicate the results of experiment 3 and second, to test the generalizability of the pragmatics model to a larger range of conditions. Compared to experiment 3, we manipulated the strength of the common ground expectations. For both, preference and novelty, we planned to have a strong, a medium and a weak manipulation. Expectation strength was manipulated by changing the way the speaker interacted with the objects prior to the request and measured as the proportion with which participants chose the preferred/novel object in the same speaker condition (see SI Appendix for results). For novelty, we succeeded in finding three qualitatively different manipulations. For preference, we piloted a number of manipulations but did not find a way to indicate only weak preference. Experiment 4 therefore included strong and medium manipulations for novelty and preference and a weak manipulation for novelty. The general expectation was the same as in experiment 3. For each level we crossed general and common ground expectations in the same way as in experiment 3, resulting in a total of 8 (strong) + 8 (medium) + 4 (weak) = 20 conditions. The strong manipulation was a direct replication of experiment 3. Model predictions were generated in the same way as in experiment 3.

Again, before evaluating model predictions, we fit GLMMs to the data for each level of common ground manipulation. Participants distinguished between congruent and incongruent trials when the speaker remained the same for the strong (β = -1.60, se = 0.52, p = .002)[1] and medium (β = -2, se = 0.51, p < .001), but not the weak manipulation (β = -0.68, se = 0.55, p = .219, model term: `alignment x speaker` in all cases). This pattern validates our manipulations. If the speaker specific manipulation is weak, changing the speaker has no differential impact on the informativeness inference. Results from the strong manipulation in experiment 4 further replicated findings from experiment 3 (Fig. 3C).

Model predictions from the pragmatics model were again highly correlated with the average response in each condition (Fig. 3D). We evaluated model fit for the same models as in experiment 3 and found again that the pragmatics model fit the data much better compared to the flat prior (BF = 4.7e+71) or the prior only model (BF = 8.9e+82). The

[1] Here, this effect differed between novelty and preference: model term manipulation x alignment x speaker; β = -2.96, se = 0.73, p < .001

inferred level of noise based on the data for the pragmatics model was 0.36 (95% HDI: 0.31 - 0.41), which was again lower compared to the alternative models (see SI Appendix).

Summarizing experiment 3 and 4, we saw that adults flexibly integrated different expectations about a speaker in a word learning scenario. This process was best described by a model that treated information integration as a form of probabilistic inference. Next, we took a developmental perspective and studied if the same model could be used to describe children's word learning.

Children

Assessment of general and speaker specific expectations (Experiment 1 and 2)

We approached children's word learning in the same way as adults'. First, we measured the strength of general (informativeness) and common ground (preference) expectations and used these measurements to set model parameters. Then we generated model predictions for how expectations should be integrated. Finally, we compared model predictions to new behavioral data. In this, we took a truly developmental perspective. We sampled children continuously across the entire age range and generated model predictions for the average developmental trajectory in each condition. In all the following experiments, we tested children between 3.0 and 4.9 years of age. For experiment 1 and 2, we also tested two-year-olds but did not find sufficient evidence that they make general and/or common ground inferences in the tasks we used.

Children were only tested in the test condition in experiment 1. To compare children's performance to chance level, we binned age by year. Four-year-olds selected the more informative object (i.e. the object that was unique to the location the speaker turned to) above chance ($t(29) = 2.80$, $p = .009$). Three-year-olds, on the other hand, did not ($t(31) = -1.31$, $p = .198$). Consequently, when we fit a GLMM to the data with age as a continuous predictor, performance increased with age ($\beta = 0.38$, se = 0.11, $p < .001$, see Fig. 4). Thus, children's expectation that speakers communicate in an informative way increased with age.

In experiment 2, we assessed whether children use common ground information to identify the referent of a novel word. We tested children with the novelty as well as the preference manipulation but found little evidence that children distinguished between requests made by the same speaker or a different speaker in the case of novelty. We therefore dropped this manipulation and focused on preference for the remainder of the study.

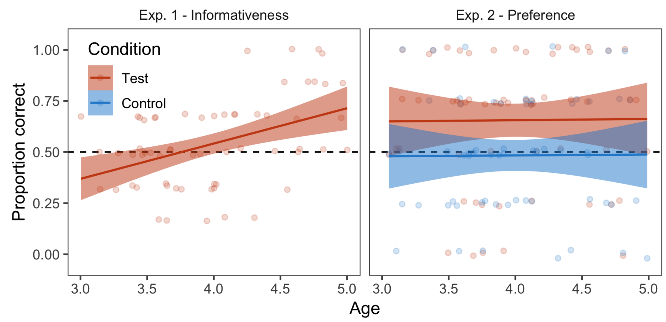

Figure 4. Results from experiment 1 and 2 for children For preference, control refers to to the different speaker condition (see Fig. 1B). Transparent dots show data from individual participants, regression lines show fitted linear models with 95% CIs. Dashed line indicates performance expected by chance.

For preference, four-year-olds selected the preferred object above chance when the same speaker made the request (t(30) = 4.14, $p < .001$), whereas three-year-olds did not (t(29) = 1.62, $p = .117$). However, when we fit a GLMM to the data with age as a continuous predictor, we found an effect of speaker identity ($\beta = 0.92$, se = 0.30, $p = .002$) but no effect of age ($\beta = 0.04$, se = 0.18, $p = .83$) or interaction between speaker identity and age ($\beta =$ -0.04, se = 0.24, $p = .886$, Fig. 4). Thus, children across the age range used common ground information to infer the referent of a novel word.

Model predictions for information integration evaluated against new data (Experiment 3)

We used the measurements from experiment 1 and 2 to specify α and $P_S(r_s)$ in the pragmatics model. However, instead of setting these parameters to a single value, we used the data to infer the intercept and slope that best described the developmental trajectory for each parameter. For $P_S(r_s)$, this was done separately for the same speaker and the different speaker condition (see SI Appendix for details). These parameter settings were then used to generate age sensitive model predictions in 2 (same or different speaker) x 2 (congruent or incongruent) = 4 conditions. As for adults, all models included a noise parameter, which was estimated based on the data.

In experiment 3, we combined the procedures of experiment 1 and 2 and collected new data from children between 3.0 and 4.9 years of age in each of the four conditions (Fig. 1C). Before comparing the data to model predictions, we fit a GLMM to the data to test if children distinguished between conditions. Children's propensity to differentiate between congruent and incongruent trials for the same or a different speaker increased with age (model term: `age x alignment x speaker`; β = -1.18, se = 0.50, p = .018).

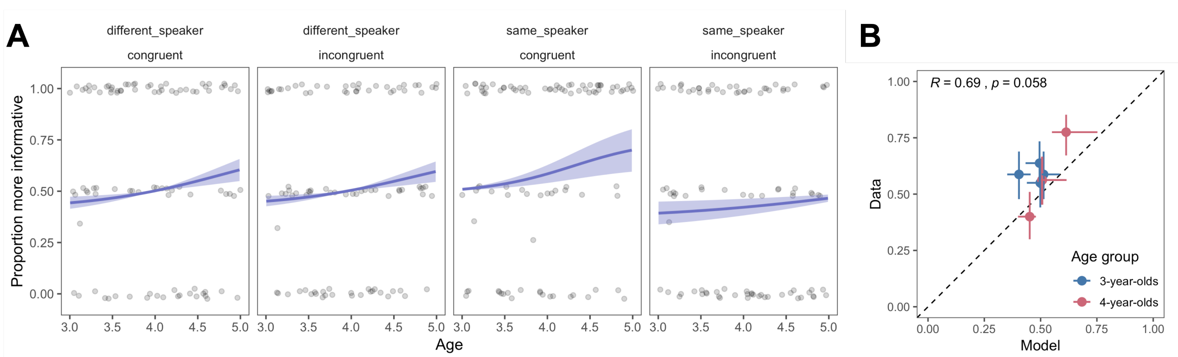

Figure 5. Results from experiment 3 for children. Purple lines show model predictions with 95% HDIs, transparent dots show data from individual participants (A). Correlation between model predictions and condition means binned by year (B). Coefficent and p-value are based on Pearson correlation statistic. Error bars represent 95% HDIs.

Next, we evaluated the model predictions (Fig 5A). For the correlational analysis, we binned model predictions and data by year. There was a substantial correlation between the predicted and measured average response in each age bin (Fig. 5B). For the model comparison, we treated age continuously. As for adults, we found a much better model fit for the pragmatics model compared to the flat prior (BF = 577) or the prior only model (BF = 10560). The inferred level of noise based on the data for the pragmatics model was 0.51 (95% HDI: 0.26 - 0.77), which was lower compared to the alternative models considered (see SI Appendix). The high level of inferred noise moved the model predictions in all conditions close to chance level (Fig. 5A). We therefore compared two additional sets of models with different parameterizations that emphasized differences between conditions in the model predictions more. This analysis was not preregistered. Parameter free models did not include a noise parameter and developmental noise models allowed the noise parameter to change with age. In each case, the pragmatics model provided a better fit compared to the alternative models (flat prior: parameter free BF = 334, developmental noise BF = 16361; prior only: parameter free BF = 20, developmental noise BF = 1e+06). Taken together, these results suggest that children's word learning is best described as a flexible integration of general and common ground expectations.

Discussion

...

Methods

All experimental procedures, sample sizes and statistical analysis were pre-registered (https://osf.io/u7kxe/). Experimental stimuli, data files and analysis scripts are freely available in an online repository (https://github.com/manuelbohn/mcc).

Participants

Adult participants were recruited via Amazon Mechanical Turk (MTurk) and received payment equivalent to an hourly wage of ~ $9. Experiment 1 and each manipulation of experiment 2 had $N = 40$ participants. Sample size in experiment 3 was $N = 121$. $N = 167$ participated in the experiments to measure the strong, medium and weak preference and novelty manipulations. Finally, experiment 4 had $N = 286$ participants.

Children were recruited from the floor of the Children's Discovery Museum in San Jose, California, USA. Parents gave informed consent and provided demographic information. We collected data from a total of 243 children between 3.0 and 4.9 years of age. We excluded 15 due to less than 75% of reported exposure to English, five because they responded incorrect on 2/2 training trials, three because of equipment malfunction and two because they quit before half of the test trials were completed. The final sample size in each experiment was as follows: $N = 62$ (41 girls, mean age = 4) in experiment 1, $N = 61$ (28 girls, mean age = 3.99) in experiment 2 and $N = 96$ (54 girls, mean age = 3.96) in experiment 3.

Experimental procedures

All experiments were framed as games in which participants would learn words from animals. They were implemented in HTML/JavaScript as a website. Adults were directed to the website via MTurk and responded by clicking objects. Children were guided through the game by an experimenter and responded by touching objects on the screen. For each animal character, we recorded a set of utterances (one native English speaker per animal) that were used to provide information and make requests. All studies started with an introduction to the animals and two training trials in which familiar objects were requested (car and ball). Subsequent test trials in each condition were presented in a random order.

The setup of experiment 1 for adults is shown in Fig. 1A. In the beginning of each trial, the animal introduced themselves (e.g. "Hi, I'm Dog") and then turned towards the table with the two objects. The same utterance was used to make a request in all adult studies ("Oh cool, there is a [non-word] on the table, how neat, can you give me the [non-word]?"). In the test condition, there was one object on the other table whereas in the control condition there were two. Participants received six trials, three per condition.

The setup for experiment 2 is shown in Fig. 1B. In the preference manipulation, the animal introduced themselves, then turned to one of the tables and expressed either that they liked ("Oh wow, I really like that one") or disliked ("Oh bleh, I really don't like that one") the object before turning to the other side and expressing the respective other attitude. Next the animal disappeared and, after a short pause, either the same or a different animal returned and requested an object while facing straight ahead. This procedure was the strong preference manipulation. In the medium version, the animal only expressed preference and did so in a more subtle way (simply saying: "Oh, wow").

In the novelty manipulation one of the tables was initially empty. The animal turned to one of the sides and commented either on the presence ("Aha, look at that") or the absence ("Hm…, nothing there") of an object before turning to the other side and commenting in a complementary way. After shortly disappearing, the same animal repeated the sequence above. When the animal left a second time, a new object appeared on the empty table. Next, either the same or a different animal returned and requested an object. This corresponded to the strong manipulation. For the medium manipulation, the animal turned to each table only once before the new object appeared. In the weak manipulation, the animal only commented on the present object once and never turned to the empty table. Participants always received six trials, three with the same and three with the different speaker.

For experiment 3 and 4 we inserted the common ground manipulation before the request in the setup of experiment 1 (Fig. 1C). For example, the animal turned to the table with one object and express that they liked object A, then turned to the other table and express that they did not like object B. Next, after quickly disappearing, the animal reappeared, turned to the table with two objects and make a request. To make it clear, which of the objects the speaker commented on while being turned to the table with the two objects during the common ground manipulation, the object was temporarily enlarged. Participants completed eight trials for one of the common ground manipulations with two trials per condition (same/different speaker x congruent/incongruent).

Experiment 1 for children was modeled after Frank and Goodman (2014). Instead of on tables, objects were presented as hanging in trees. After introducing themselves, the animal turned to the tree with two objects and said: "This is a tree with a [non-word], how neat, a tree with a [non-word]"). Next, the trees and the objects in them disappeared and new trees replaced them. The two objects from the tree the animal turned to previously were now spread across the two trees (one object per tree, position counterbalanced). While facing straight, the animal first said "Here are some more trees" and then asked the child to pick the tree with the object that corresponded to the novel word ("Which of these trees has a [non-word]"). Children received six trials in a single test condition.

Experiment 2 for children was identical to the strong preference manipulation for adults. Children received eight trials, four with the same and four with a different animal returning.

In experiment 3 for children, we again inserted the preference manipulation into the setup of experiment 1. After greeting the child, the animal turned to one of the trees, pointed to an object (object was temporarily enlarged and moved closer to the animal) and expressed liking or disliking. Then the animal turned to the other tree and expressed the other attitude for the other kind of object. Next, the animal disappeared and either the same or a different animal returned. The rest of the trial was identical to the label and request phase of experiment 1. Children received eight trials, two per condition (same/different speaker x congruent/incongruent).

Analysis

All analysis were run in R (R Core Team, 2018). GLMMs were fit via the function `glmer` from the package `lme4` (Bates, Mächler, Bolker, & Walker, 2015). Probabilistic models and model comparisons were implemented in WebPPL (Goodman & Stuhlmüller, 2014) using the r package `rwebppl` (Braginsky, Tessler, & Hawkins, 2019). The flat prior model had the same structure as the pragmatics model but used a uniform prior over objects. The prior only model predicted choice based on the prior probability for each object alone. Note that the proportions that informed the prior distribution in the pragmatics and prior only model were measured when participants chose between two objects in experiment 2. In experiment 3, however, three objects were involved. For each object we used the proportion measured in experiment 2 as the prior probability. This approach spread out the absolute probability mass but conserved the relative relation between objects. Bayes Factors for model comparisons are based on marginal likelihoods of each model given the data. The corresponding model code can be found in the associated online repository.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Braginsky, M., Tessler, M. H., & Hawkins, R. (2019). *Rwebppl: R interface to webppl*. Retrieved from https://github.com/mhtess/rwebppl

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96.

Goodman, N. D., & Stuhlmüller, A. (2014). The design and implementation of probabilistic programming languages. http://dippl.org.

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