Predicting pragmatic cue integration in adults' and children's inferences about novel word meanings

Manuel Bohn[1,2], Michael Henry Tessler[3], Megan Merrick[1], & 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

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

Abstract

Language is learned in complex social settings where listeners must reconstruct speakers' intentend meanings from context. To navigate this challenge, children can use pragmatic reasoning to learn the meaning of unfamiliar words. One important challenge for pragmatic reasoning is that it requires integrating multiple information sources. Here we study this integration process. We isolate two sources of pragmatic information and, using a probabilistic model of conversational reasoning, formalize both how they should be combined and how this process might develop. We use this model to generate quantitative predictions, which we test against new behavioral data from three- to five-year-old children and adults in a series of pre-registered experiments. Results show close numerical alignment between model predictions and data. This work integrates distinct sets of findings regarding early language and suggests that pragmatic reasoning models can provide a quantitative framework for understanding developmental changes in language learning.

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

Predicting pragmatic cue integration in adults' and children's inferences about novel word meanings

What someone means by an utterance is oftentimes not reducible to the words they used. It takes pragmatic inference – context-sensitive reasoning about the speaker's intentions - to recover the intended meaning[1–3]. Contextual information comes in many forms. On the one hand, there is information provided by the utterance[1] itself. Competent language users expect each other to communicate in a cooperative way such that speakers produce utterances that are relevant and informative. Thus, semantic ambiguity can be resolved by reasoning about why the speaker produced this particular utterance[1,3–5]. On the other hand, there is information provided by common ground (the body of knowledge and beliefs shared between interlocutors)[4,6,7]. Because utterances are embedded in common ground, pragmatic reasoning in context always requires information integration. But how does integration proceed? And how does it develop? Verbal theories assume that information is integrated and that this process develops but do not specify how. We bridge this gap by formalizing information integration and development in a probabilistic model of pragmatic reasoning.

Children learning their first language make inferences about intended meanings based on utterance-level and common-ground information both for language understanding and language learning[5,8,9]. Starting very early, infants expect adults to produce utterances in a cooperative way[10], and expect language to be carrying information[11]. By age two, children are sensitive to the informativeness of communication[12]. By age three children can use this expectation to make pragmatic inferences[13,14] and to infer novel word meanings[15]. And

[1]We use the terms utterance, utterance-level information or utterance-level cues to capture all cues that the speaker provides for their intended meaning. This includes direct referential information in the form of actions such as pointing or gazing, semantic information in the form of conventional word meanings as well as pragmatic inferences that are licenced by the particular choice of words or actions.

although older children continue to struggle with some complex pragmatic inferences until age five and beyond[16], an emerging consensus identifies these difficulties as stemming from difficulties reasoning about linguistic alternatives rather than pragmatic deficits[17–19]. Thus, children's ability to reason about utterance-level pragmatics is present at least by ages three to five, and possibly substantially younger.

Evidence for the use of common ground information by young children is even stronger: Common ground information guides how infants produce non-verbal gestures and interpret ambiguous utterances[20,21]. For slightly older children, common ground – in the form of knowledge about discourse novelty, preferences, and even discourse expectations – also facilitates word learning[22–25].

All of these examples, however, highlight children's use of a single pragmatic information source or cue. Harnessing multiple – potentially competing – cues poses a separate challenge. One aspect of this integration problem is how to balance common ground information that is built up over the course of an interaction against information gleaned from the current utterance. Much less is known about whether and how children – or even adults – combine these types of information. While many theories of pragmatic reasoning presuppose that both information sources are integrated, the nature of their relationship has typically not been specified.

Recent innovations in probabilistic models of pragmatic reasoning provide a quantitative method for addressing the problem of integrating multiple sources of contextual information. This class of computational models, which are referred to as Rational Speech Act (RSA) models[26,27] formalize the problem of language understanding as a special case of Bayesian social reasoning. A listener interprets an utterance by assuming it was produced by a cooperative speaker who had the goal to be informative. Being informative is defined as providing a message that would increase the probability of the listener recovering the speaker's intended meaning in context. This notion of contextual informativeness captures the Gricean idea of cooperation between speaker and listener, and provides a first approximation to what we have described above as utterance-level pragmatic information.

Listeners and speakers also enter into a conversation with assumptions about what is likely to be talked about, a reflection of the common ground shared between them. RSA models capture common ground information as a shared prior distribution over possible intended meanings. Thus, a natural locus for information integration within probabilistic models of pragmatic reasoning is the trade off between the prior probability of a meaning and the informativeness of the utterance. This trade off between contextual factors during word learning is a unique aspect of the word learning problem that is not addressed by other computational models of word learning, which have focused on learning from cross-situational, co-occurrence statistics[28,29] or describing generalizations about word meaning[30].

We make use of this framework to study pragmatic cue integration across development. To this end, we adapt a method used in perceptual cue integration studies[31] predictions about conditions in which they either coincide or conflict. Finally, we pre-register these quantitative predictions and test them against new data from adults and children.

We start by replicating previous findings with adults showing that listeners make pragmatic inferences based on non-linguistic properties of utterances in isolation (Experiment 1). In separate experiments, we then show that adults make inferences based on common ground information (Experiment 2A and 2B). We use data from these experiments as parameters to generate a priori predictions from RSA models about how utterance and common ground information should be integrated. We consider three models that make different assumptions about the integration process: In the pragmatics model, the two information sources are integrated with one another; according to the flat prior model, participants focus only on the utterance information and in the prior only model, only common ground information is considered. We compare predictions from these models to new empirical data from experiments in which utterance and common ground information are manipulated simultaneously (Experiment 3 and 4).

After successfully validating this approach with adults, we apply the same model-driven experimental procedure to children: We first show that they make pragmatic inferences based on utterance and common ground information separately (Experiment 5 and 6). Then we generate a priori model predictions and compare them to data from an experiment – parallel to Experiment 3 – in which both information sources have to be integrated (Experiment 7).

Taken together, this work makes two primary contributions: first, it shows that both adults and children integrate utterance-level (Gricean) and common-ground information flexibly. Second, it uses Bayesian data analysis within the RSA framework to provide a model for understanding the multiple loci for developmental change in complex behaviors like contextual communication.

Results

How do adults integrate contextual sources of information?

Inferences based on utterance and common ground information (Experiments 1 and 2).

In Experiment 1, participants could learn which object a novel word referred to by assuming that the speaker communicated in an informative way[15]. 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 and pointed to the table with the two objects (A and B) and used a 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 had wanted to refer to object A, they would have pointed to the table with the single object (this being the least ambiguous way to refer to that object). In the control condition, both tables contained both objects and no inference could be made

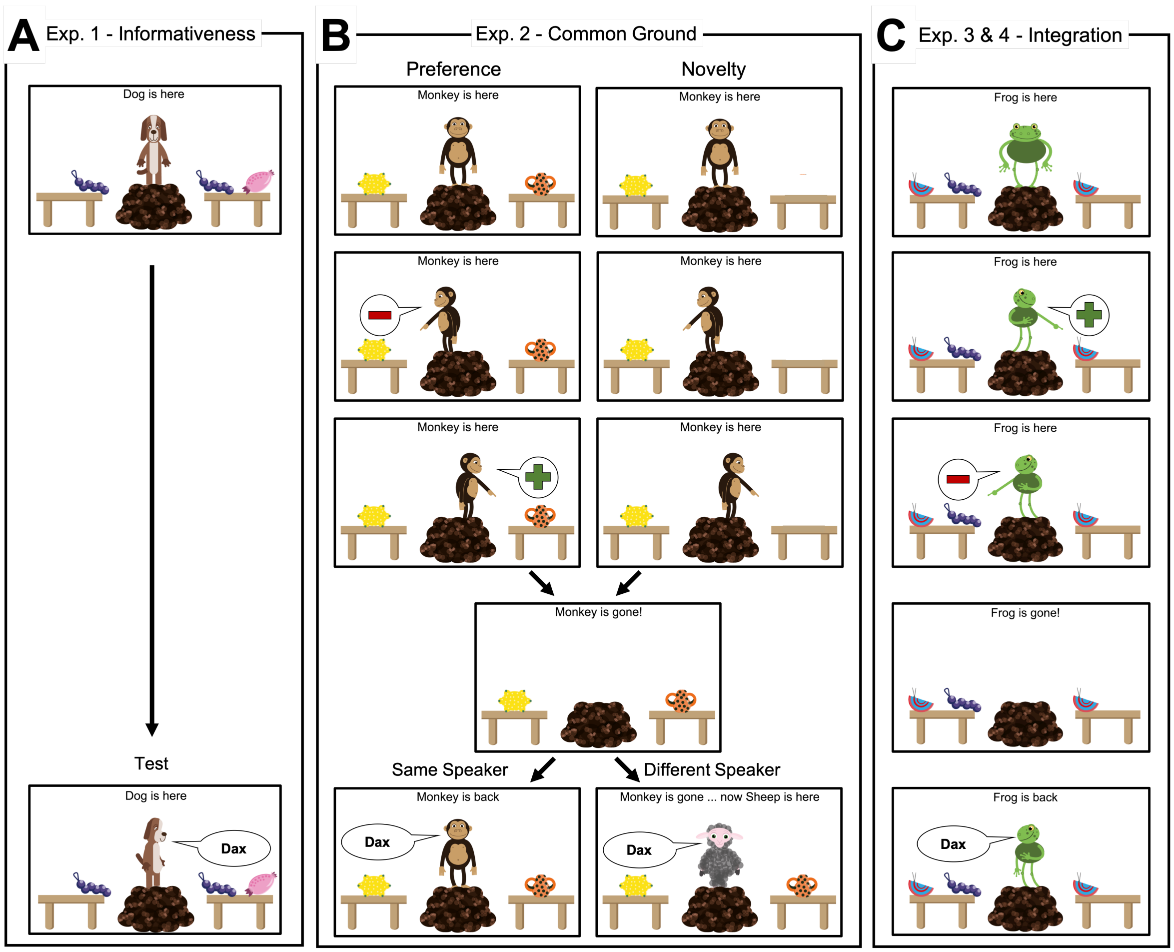

Figure 1. Schematic experimental procedure with screenshots from the adult experiments. In all conditions, at test (bottom), the speaker ambiguously requested an object using a non-word (e.g. "dax"). Participants clicked on the object they thought the speaker referred to. Informativeness (Experiment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

based on the speaker's behavior. Participants selected object B above chance in the test condition (mean = 0.74, 95% CI of mean = [0.65; 0.83], $t(39) = 5.51$, $p < .001$, $d = 0.87$)

and more often compared to the control condition ($\beta = 1.28$, se = 0.29, $p < .001$, see Fig 2).

In Experiments 2A and 2B, we tested if participants use common ground information that is specific to a speaker to identify the referent of a novel word[22,24]. In Experiment 2A, the speaker expressed a preference for one of two objects (Fig 1B, left). Later, the speaker used a novel word to request an object. Adults selected the preferred object above chance (mean = 0.97, 95% CI of mean = [0.93; 1], t(39) = 29.14, $p < .001$, d = 4.61) and more so than in a control condition, where a different speaker, whose preferences were unknown, made the request ($\beta = 2.92$, se = 0.57, $p < .001$). In Experiment 2B, common ground information came in the form of novelty (Fig 1B, right). First, the speaker encountered one object on one of the tables. Later, a second object appeared. When the same speaker then used a novel word to request an object, participants selected the new object above chance (mean = 0.83, 95% CI of mean = [0.73; 0.93], t(39) = 6.77, $p < .001$, d = 1.07) , and more often compared to when a different speaker (to whom both objects were equally new) made the request ($\beta = 6.27$, se = 1.96, $p = .001$, see Fig 2). Taken together, Experiments 1 and 2 confirmed that adults make pragmatic inferences based on information provided by the utterance as well as by common ground and provided quantitative estimates of the strength of these inferences for use in our model.

Model predictions for information integration evaluated against new data (Experiment 3). We modeled the integration of utterance informativity and common ground as a process of socially-guided probabilistic inference, using the results of Experiments 1 and 2 to inform key parameters of a computational model. The Rational Speech Act (RSA) model architecture introduced by[26] encodes conversational reasoning through the perspective of a listener ("he" pronoun) who is trying to decide on the intended meaning of the utterance he heard from the speaker ("she" pronoun). The basic idea is that the listener combines his uncertainty about the speaker's intended meaning - a prior distribution over referents P(r) - with his generative model of how the utterance was

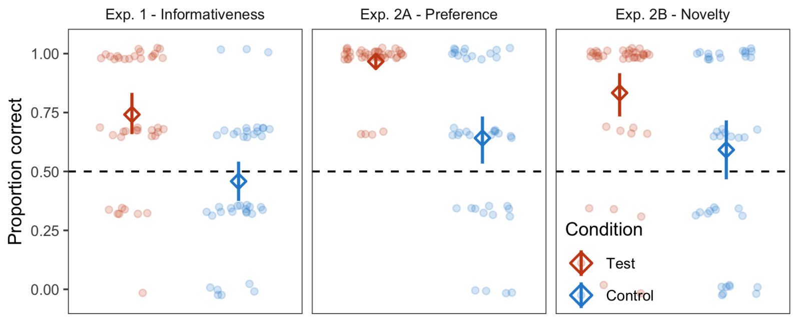

Figure 2. Results from Experiments 1, 2A, and 2B for adults. For preference and novelty, control refers to a different speaker (see Fig 1B). Transparent dots show data from individual participants, diamonds represent condition means, error bars are 95% CIs. Dashed line indicates performance expected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

produced: a speaker trying to convey information to him. To adapt this model to the word learning context, we enrich this basic architecture with a mechanism for expressing uncertainty about the meanings of words (lexical uncertainty) - a prior distribution over lexica P(L)[32].

$$P_L(r, \mathcal{L}|u) \propto P_S(u|r, \mathcal{L}) \cdot P(\mathcal{L}) \cdot P(r)$$

In the above equation, the listener is trying to jointly resolve the speaker's intended referent r and the meaning of words (thus learning the lexicon $\mathcal{L}$). He does this by imagining what a rational speaker would say, given the referent they are trying to communicate and a lexicon. The speaker is an approximately rational Bayesian actor (with degree of rationality alpha), who produces utterances as a function of their informativity. The space of utterances the speaker could produce depends upon the lexicon $P(u|\mathcal{L})$; simply put, the speaker labels objects with the true labels under a given lexicon L (see supplementary information for details):

$$P_S(u|r, \mathcal{L}) \propto Informativity(u; r)^\alpha \cdot P(u|\mathcal{L})$$

The informativity of an utterance for a referent is taken to be the probability with which a naive listener, who only interprets utterances according to their literal semantics, would select a particular referent given an utterance.

$$Informativity(u; r) = P(r|u) \propto P(r) \cdot \mathcal{L}_{point}$$

The speaker's possible utterances are pairs of linguistic and non-linguistic signals, namely labels and points. Because the listener does not know the lexicon, the informativity of an utterance comes from the speaker's point, the meaning of which is encoded in $\mathcal{L}_{point}$ and is simply a truth-function checking whether or not the referent is at the location picked out by the speaker's point. Though the speaker makes their communicative decision assuming the listener does not know the meaning of the labels, we assume that in addition to a point, the speaker produces a label consistent with their own lexicon $\mathcal{L}$, described by $P(u|\mathcal{L})$ (see supplementary information for modeling details).

This computational model provides a natural avenue to formalize quantitatively how informativeness and common ground trade-off during word learning. As mentioned above, the common ground shared between speaker and listener plays the role of the listener's prior distribution over meanings, or types of referents, that the speaker might be referring to and which we posit depends on prior interactions around the referents in the present context (e.g., preference or novelty; Experiment 2A and B). We use the results from Experiment 2 to specify this distribution. The in-the-moment, contextual informativeness of the utterance is captured in the likelihood term, whose value depends on the rationality parameter α. Assumptions about rationality may change depending on context and we therefore used the data from Experiment 1 to specify α (see supplementary information for details about these parameters).

The model generates predictions for situations in which utterance and common ground expectations are jointly manipulated (Fig 1C - see supplementary information for additional details and a worked example of how predictions were generated). In addition to the parameters fit to the data from previous experiments, we include an additional noise parameter to account for responses better explained by a process of random guessing than by pragmatics; we estimate this parameter from the observed data (Experiment 3). Including the noise parameter greatly improved the model fit to the data (see supplementary information for details). We did not pre-register the inclusion of a noise parameter for Experiment 3 but did so for all subsequent experiments.

In Experiment 3, we combined the procedures of Experiment 1 and 2A or 2B. 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 utterance information is considered). Participants distinguished between congruent and incongruent trials when the speaker remained the same, as evidenced by the fit of a generalized linear mixed effects model (model term: `alignment x speaker`; β = -2.64, se = 0.48, $p < .001$).

Participants' average responses were highly correlated with the model's predictions in each condition (Fig 3B). To test whether participants in fact balanced both information sources, we compared the pragmatics model to two alternative models: the *flat prior model*, which ignores common ground information and the *prior only model*, which ignores utterance information. Model fit was considerably better for the pragmatics model compared to the flat prior model (Bayes Factor (BF) = 4.2e+53) or the prior only model (BF = 2.5e+34), suggesting that participants considered and integrated both sources of information. The

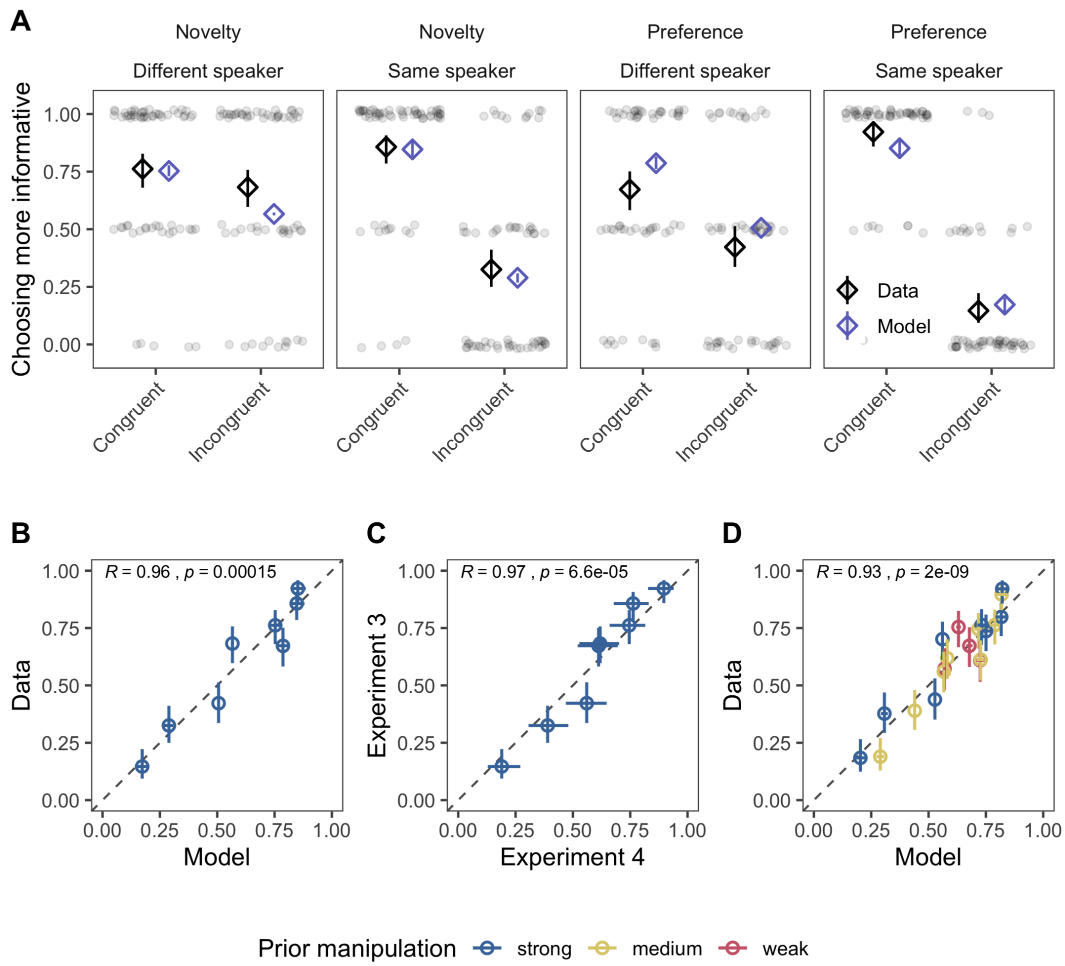

Figure 3. Results from Experiment 3 and 4 for adults. (A) Data and model predictions by condition for Experiment 3. Transparent dots show data from individual participants, diamonds represent condition means. (B) Correlation between model predictions and data in Experiment 3, (C) between data in Experiment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

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 substantially lower for the pragmatics model compared to the alternative models (see supplementary information), lending additional support to the conclusion that the pragmatics model better captured the behavioral data. Rather than explaining systematic structure in the data, the alternative models achieved their best fit only by assuming a very high level of noise.

Replication and extension to different levels of common ground information (Experiment 4). To test if our model makes accurate predictions for different combinations, we first replicated and then extended the results of Experiment 3 to a broader range of experimental conditions. Specifically, we manipulated the strength of the common ground information (strong, medium and weak manipulation) by changing the way the speaker interacted with the objects prior to the request We ran a total of 20 conditions, including a direct replication of Experiment 3 (see Fig 3C).

Model predictions from the pragmatics model were again highly correlated with the average response in each condition (see 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 inferred level of noise based on the data for the pragmatics model was 0.36 (95% HDI: 0.31 - 0.41), which was similar to Experiment 3 and again lower compared to the alternative models (see supplementary information).

Do children integrate contextual information?

The previous section showed that competent language users flexibly integrate information during pragmatic word learning. Do children make use of multiple information sources during word learning as well? When does this integration emerge developmentally? While many verbal theories of language learning imply that this integration takes place, the actual process has neither been described in detail nor tested. Here we provide an explanation in the form of our pragmatics model and test if it is able to capture children's word learning. Embedded in the assumptions of the model is the idea that developmental change is change in the strength of the individual inferences, leading to a change in the strength of the integrated inference. As a starting point, our model assumes developmental continuity in the integration process itself, though this assumption could be called into question by a poor model fit.

Inferences based on utterance and common ground information (Experiment 5 and 6). The study for children followed the same general pattern as the one for adults. We generated model predictions for how information should be integrated by first measuring children's ability to use utterance (informativeness) and common ground (preference) information in isolation when making pragmatic inferences. We then adapted our model to study developmental change: We sampled children continuously between 3.0 and 5.0 years of age – a time in which children have been found to make the kind of pragmatic inferences we studied here [8; frank2014inferring] - and generated model predictions for the average developmental trajectory in each condition[2].

Experiment 5 was analogous to Experiment 1 for adults. 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 (mean = 0.62, 95% CI of mean = [0.53; 0.71], t(29) = 2.80, p = .009, d = 0.51). Three-year-olds, on the other hand, did not (mean = 0.46, 95% CI of mean = [0.41; 0.52], t(31) = -1.31, p = .198, d = 0.23). 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 ability to use utterance information in a word learning context increased with age.

In Experiment 6, we assessed whether children use common ground information to

[2]For Experiment 5 and 6, we also tested two-year-olds but did not find sufficient evidence that they use utterance and/or common ground information in the tasks we used to justify investigating their ability to integrate the two.

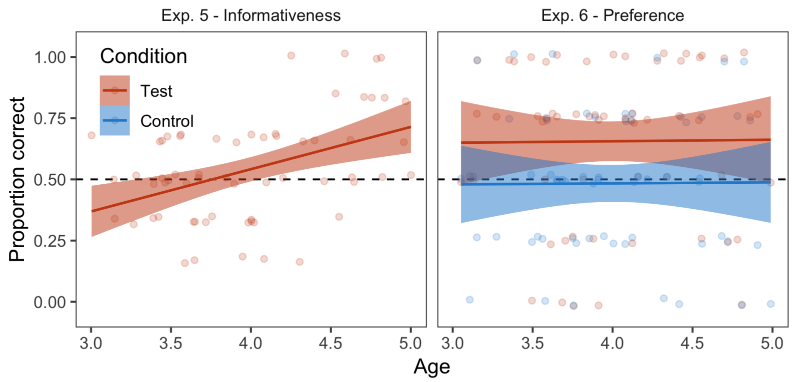

Figure 4. Results from Experiment 5 and 6 for children. For preference, control refers to to the different speaker condition (see Fig. 1B). Transparent dots show data from individual participants, regression lines show fitted linear models with 95% CIs. Dashed line indicates performance expected by chance.

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. Since our focus was on how children selectively integrate the two sources of information, we therefore dropped this manipulation and focused on preference for the remainder of the study.

For preference, four-year-olds selected the preferred object above chance when the same speaker made the request (mean = 0.71, 95% CI of mean = [0.61; 0.81], t(30) = 4.14, $p <$ .001, d = 0.74), whereas three-year-olds did not (mean = 0.60, 95% CI of mean = [0.47; 0.73], t(29) = 1.62, p = .117, d = 0.30). However, when we fit a GLMM to the data with age as a continuous predictor, we found an effect of speaker identity (β = 0.89, se = 0.24, $p <$ .001) but no effect of age (β = 0.02, se = 0.16, p = .92) or interaction between speaker identity and age (β = -0.01, se = 0.23, p = .97, see Fig 4). Thus, children across the age range used common ground information to infer the referent of a novel word.

Developmental model predictions evaluated against new data (Experiment 7). We used the measurements from Experiment 5 and 6 to specify the strength of informativity, α, and common ground in the pragmatics model. Instead of inferring a single value we inferred the intercept and slope for each parameter that best described the developmental trajectory in the data of Experiment 5 and 6. 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 7, we combined the procedures of Experiment 5 and 6 and collected new data from children between 3.0 and 5.0 years of age in each of the four conditions (Fig 1C). 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 = -0.89, se = 0.36, p = .013).

Our modeling results suggest that children flexibly integrate both common ground and informativity information, and that this integration process is accurately captured by the pragmatics model at least for four-year-olds. For the correlational analysis, we binned model predictions and data by year. There was a substantial correlation between the predicted and measured average response for four-year-olds, but less so for three-year-olds (Fig 5B). One of the reasons for the latter was the low variation between conditions. For the model comparison, we treated age continuously. As with 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 but numerically higher than that of adults (see supplementary information).

The high level of inferred noise moved the model predictions for children in all conditions close to chance level. We therefore compared two additional sets of models with different parameterizations that emphasized differences between conditions in the model predictions more (see supplementary information, see Fig 5A). This analysis was not pre-registered. 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).

Discussion

Integrating multiple sources of information is an integral part of human communication[33]. To infer the intended meaning of an utterance, listeners must combine their knowledge of communicative conventions (semantics and syntax) with social expectations about their interlocutor. This integration is especially vital in early language learning, and the different varieties of pragmatic information are among the most important sources[8]. But how are different cues integrated during word learning? Here we used a Bayesian pragmatics model to formalize this integration process. We studied how utterance-level (Gricean expectations) about informative communication are integrated with common ground information that follows from prior interactions with the speaker. Adults' and children's learning was best predicted by a model in which both sources of information traded-off flexibly. Alternative models that considered only one source of information made substantially worse predictions.

All of the models we compared here integrated some explicit structure, rather than (for example) simply weighing expectations by some ratio. We made this decision because we wanted to make predictions within a framework in which the models were models of the task, rather than simply models of the data. That is, inferences are not computed separately by the modeler and specified as inputs to a regression model, but instead are the results of an integrated process that operates over a (schematic) representation of the experimental stimuli. Further, our models are variants derived from the broader RSA framework, which

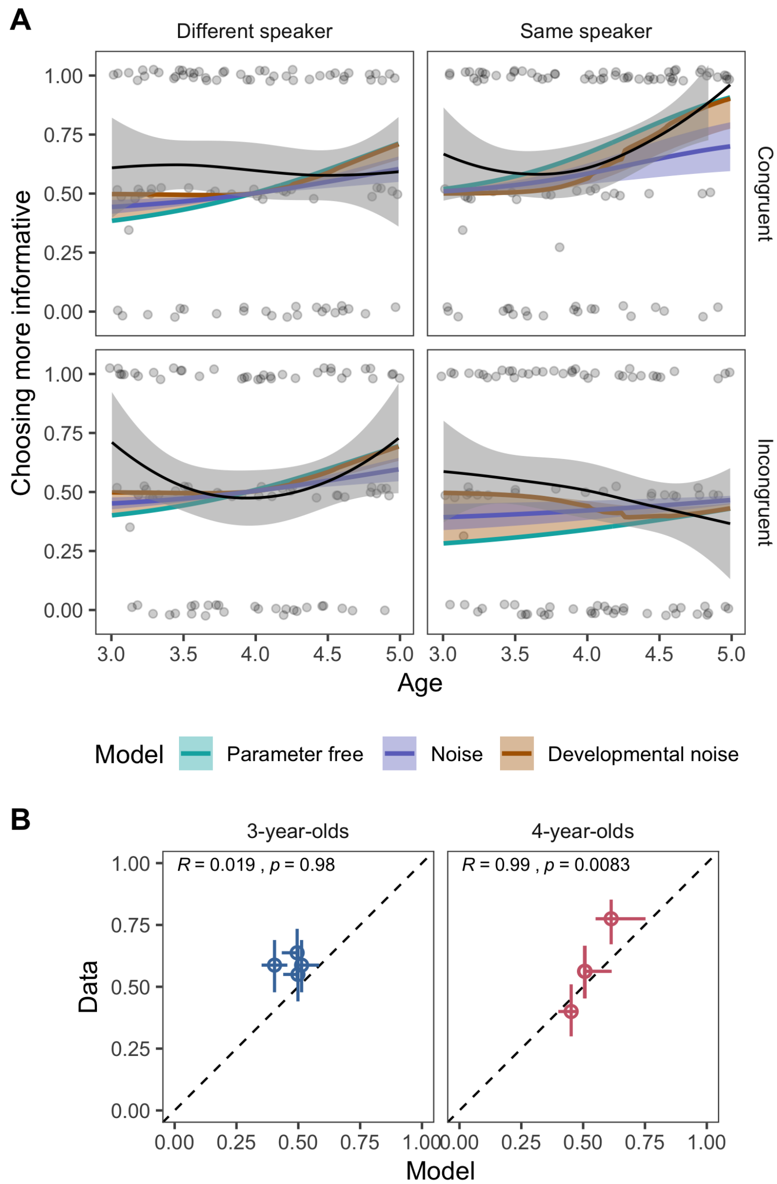

Figure 5. Results from Experiment 7 for children. (A) Model predictions and data across age in the four conditions. Colored lines show predictions from the pragmatics model with different noise parameters. Transparent black dots show data from individual participants and black lines show conditional means of the data. (B) Correlation between model predictions (with noise parameter) and condition means binned by year. Coefficients and p-values are based on Pearson correlation statistic. Error bars and shaded regions represent 95% HDIs.

has been integrated into larger systems for language learning in context[34].

We conceptualized developmental change as age related changes in the propensity to make the individual inferences. That is, while the degree to which listeners expect speakers to be informative or follow common ground changes with age, the process by which expectations are integrated remains the same. However, other developmental models are also worth exploring in future work; one possible candidate would be a model in which the integration process itself changes with age. Our model did not successfully describe three-year-olds' inferences; thus, it is possible that they were not able to integrate information sources. But our findings are also consistent with a simpler explanation, namely that the overall weaker responses we observed in the independent measurement experiments (Experiments 5 and 6), combined with some noise in responding, led the younger children to appear relatively random in their responses.

Studying how multiple types of pragmatic cues are balanced contributes to a more comprehensive understanding of word learning. In the current study, participants inferred the referent by integrating non-linguistic cues (speakers pointing to a table) with assumptions about speaker informativeness and common ground information, going beyond previous experimental work in measuring how these information sources were combined. The real learning environment is far richer than what we captured in our experimental design, however. For example, in addition to multiple layers of social information, children can rely on semantic and syntactic features of the utterances as cues to meaning[35–37]. Across development, children learn to recruit these different sources of information and integrate them. RSA models allow for the inclusion of semantic information as part of the utterance[32] and it will be a fruitful avenue for future research to model the integration of linguistic and pragmatic information across development.

More broadly, our work here shows how computational models of language comprehension can be used as powerful tools to explicate and test hypotheses about information integration. Furthermore, we took a first step towards integrating developmental change into this theoretical framework.

Methods

The study was approved by the Stanford Institutional Review Board (protocol no. 19960). 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. Each participant contributed data to only one experiment. 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. Each child contributed data to only one experiment. We collected data from a total of 243 children between 3.0 and 5.0 years of age. We excluded 15 children due to less than 75% of reported exposure to English, five because they responded incorrectly 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 5, $N = 61$ (28 girls, mean age = 3.99) in Experiment 6 and $N = 96$ (54 girls, mean age = 3.96) in Experiment 7.

Materials

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 of an iPad tablet[38]. 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 experiments 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. In the control condition, no inference was possible based on the speaker's turning. The "correct" object in the control condition was randomly chosen from the two objects on the table. Technically, this condition did not control for any alternative explanations and we therefore did not run it for children (see below). 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 expressed 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 5 for children was modeled after 15. Instead of on tables, objects were presented as hanging in trees (to facilitate showing points to distinct locations). 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) in a randomized order.

Analysis

All analyses were run in R[39]. All p-values are based on two sided analysis. Cohen's d (computed via the function `cohensD`) was used as effect size for t-tests. Frequentist GLMMs were fit via the function `glmer` from the package `lme4`[40] and had a maximal random effect structure conditional on model convergence. Probabilistic models and model comparisons were implemented in WebPPL[41] using the r package `rwebppl`[42]. Bayes Factors for model comparisons were based on marginal likelihoods of each model given the data. Details on models, including information about priors for parameter estimation and Markov chain Monte Carlo settings can be found in the supplementary information and the online repository.

References

1. Grice, H. P. *Studies in the way of words.* (Cambridge, MA: Harvard University Press, 1991).

2. Levinson, S. C. *Presumptive meanings: The theory of generalized conversational implicature.* (Cambridge, MA: MIT press, 2000).

3. Sperber, D. & Wilson, D. *Relevance: Communication and cognition.* (Cambridge, MA: Blackwell Publishers, 2001).

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

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

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

7. Clark, E. V. Common ground. in *The handbook of language emergence* (eds. MacWhinney, B. & O'Grady, W.) **87**, 328–353 (John Wiley & Sons, 2015).

8. Bohn, M. & Frank, M. C. The pervasive role of pragmatics in early language. *Annual Review of Developmental Psychology* (in press).

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

10. Behne, T., Carpenter, M. & Tomasello, M. One-year-olds comprehend the communicative intentions behind gestures in a hiding game. *Developmental science* **8**, 492–499 (2005).

11. Vouloumanos, A., Onishi, K. H. & Pogue, A. Twelve-month-old infants recognize that speech can communicate unobservable intentions. *Proceedings of the National*

Academy of Sciences **109**, 12933–12937 (2012).

12. O'Neill, D. K. & Topolovec, J. C. Two-year-old children's sensitivity to the referential (in) efficacy of their own pointing gestures. *Journal of Child Language* **28**, 1–28 (2001).

13. Stiller, A. J., Goodman, N. D. & Frank, M. C. Ad-hoc implicature in preschool children. *Language Learning and Development* **11**, 176–190 (2015).

14. Yoon, E. J. & Frank, M. C. The role of salience in young children's processing of ad hoc implicatures. *Journal of Experimental Child Psychology* **186**, 99–116 (2019).

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

16. Noveck, I. A. When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition* **78**, 165–188 (2001).

17. Skordos, D. & Papafragou, A. Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition* **153**, 6–18 (2016).

18. Horowitz, A. C., Schneider, R. M. & Frank, M. C. The trouble with quantifiers: Exploring children's deficits in scalar implicature. *Child Development* **89**, e572–e593 (2018).

19. Barner, D., Brooks, N. & Bale, A. Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition* **118**, 84–93 (2011).

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

21. Saylor, M. M., Ganea, P. A. & Vázquez, M. D. What's mine is mine: Twelve-month-olds use possessive pronouns to identify referents. *Developmental Science* **14**, 859–864

(2011).

22. Akhtar, N., Carpenter, M. & Tomasello, M. The role of discourse novelty in early word learning. *Child Development* **67**, 635–645 (1996).

23. Diesendruck, G., Markson, L., Akhtar, N. & Reudor, A. Two-year-olds' sensitivity to speakers' intent: An alternative account of samuelson and smith. *Developmental Science* **7**, 33–41 (2004).

24. Saylor, M. M., Sabbagh, M. A., Fortuna, A. & Troseth, G. Preschoolers use speakers' preferences to learn words. *Cognitive Development* **24**, 125–132 (2009).

25. Sullivan, J., Boucher, J., Kiefer, R. J., Williams, K. & Barner, D. Discourse coherence as a cue to reference in word learning: Evidence for discourse bootstrapping. *Cognitive Science* **43**, e12702 (2019).

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

27. Goodman, N. D. & Frank, M. C. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences* **20**, 818–829 (2016).

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

29. Frank, M. C., Goodman, N. D. & Tenenbaum, J. B. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science* **20**, 578–585 (2009).

30. Xu, F. & Tenenbaum, J. B. Word learning as bayesian inference. *Psychological review*

114, 245 (2007).

31. Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429 (2002).

32. Bergen, L., Levy, R. & Goodman, N. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics* **9**, (2016).

33. Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M. & Sedivy, J. C. Integration of visual and linguistic information in spoken language comprehension. *Science* **268**, 1632–1634 (1995).

34. Wang, S., Liang, P. & Manning, C. D. Learning language games through interaction. in *54th annual meeting of the association for computational linguistics, acl 2016* 2368–2378 (Association for Computational Linguistics (ACL), 2016).

35. Clark, E. V. What's in a word? On the child's acquisition of semantics in his first language. in *Cognitive development and acquisition of language* (ed. Moore, T.) 65–110 (Academic Press, 1973).

36. Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S. & Steedman, M. Bootstrapping language acquisition. *Cognition* **164**, 116–143 (2017).

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

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

39. R Core Team. *R: A language and environment for statistical computing.* (R Foundation

for Statistical Computing, 2018).

40. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**, 1–48 (2015).

41. Goodman, N. D. & Stuhlmüller, A. The design and implementation of probabilistic programming languages. (2014).

42. Braginsky, M., Tessler, M. H. & Hawkins, R. *Rwebppl: R interface to webppl.* (2019).

Acknowledgements

MB received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 749229. MCF was supported by a Jacobs Foundation Advanced Research Fellowship and the Zhou Fund for Language and Cognition. We thank Jacqueline Quirke and Sabina Zacco for help with the data collection and Bria Long and Gregor Kachel for comments on an earlier version of the paper.

Author Contributions

MB and MCF conceptualized the study, MM collected the data, MB and MHT analyzed the data, MB, MHT and MCF wrote the manuscript, all authors approved the final version of the manuscript.

Competing Interests

The authors declare no competing interests.