Dissertation Proposal

Claire Bergey[1]

[1] The University of Chicago

Dissertation Proposal

An utterance can say much more about the world than its literal interpretation might suggest. For instance, if you hear a colleague say "We should hire a female professor," you might infer something about the speaker's goals, the makeup of a department, or even the biases of a field—none of which is literally stated. These inferences depend on recognition that a speaker's intended meaning can differ from the literal meaning of their utterance, and the process of deriving this intended meaning is called pragmatics. General frameworks for understanding pragmatic inference posit that speakers tend to follow general principles of conversation—for instance, that they tend to be relevant, brief, and otherwise helpfully informative (Clark, 1990; H Paul Grice, 1975; Sperber & Wilson, 1986). When a speaker deviates from these principles, a listener can reason about the alternative utterances the speaker might have said and infer some intended meaning that goes beyond the literal meaning of their utterance.

Pragmatic inference is also a potentially powerful mechanism for learning about new words and concepts. People can learn the meanings of words by tracking associations between word use and present objects alone (Yu & Smith, 2007), but reasoning about a speaker's intended meaning and not just the words they say may support more rapid and more accurate learning (Frank, Goodman, & Tenenbaum, 2009). For example, Akhtar, Carpenter, and Tomasello (1996) showed that young children can infer the meaning of a new word by using the principle that people tend to remark on things that are new and interesting to them. In this study, an experimenter leaves the room and a new toy emerges in her absence; once she comes back, the toy is familiar to the child but not to the experimenter. When she uses a novel name, "gazzer," the child can infer that the word refers to the toy that is novel to the experimenter, and not other toys the experimenter had already seen. Experiments with adults show that they too can use general principles of informativeness to infer a novel referent's name (Frank & Goodman, 2014).

One potential pragmatic tool for learning about referents is contrastive inference from description. To the extent that communicators strive to be minimal and informative, description should discriminate between the referent and some relevant contrasting set. This contrastive inference is fairly obvious from some types of description, such as some postnominal modifiers: "The door with the lock" clearly implies a contrasting door without one (Ni, 1996). The degree of contrast implied by more common descriptive forms, such as prenominal adjectives in English, is less clear: speakers do not always use prenominal adjectives minimally, often describing more than is needed to establish reference (Engelhardt, Barış Demiral, & Ferreira, 2011; Mangold & Pobel, 1988; Pechmann, 1989). Nevertheless, Sedivy, Tanenhaus, Chambers, and Carlson (1999) showed that people can use these inferences to resolve referential ambiguity in familiar contexts. When asked to "Pick up the tall cup," people directed their attention more quickly to the target when a short cup was present, and did so in the period before they heard the word "cup." Because the speaker would not have needed to specify "tall" unless it was informative, listeners were able to use the adjective to direct their attention to a tall object with a shorter counterpart. Subsequent work using similar tasks has corroborated that people can use contrastive inferences to direct their attention among familiar referents (Aparicio, Xiang, & Kennedy, 2016; Ryskin, Kurumada, & Brown-Schmidt, 2019; Sedivy, 2003).

But what if you didn't know the meaning of the key words in someone's utterance–could you use the same kind of contrastive inferences to learn about new words and categories? Suppose a friend asks you to "Pass the tall dax." Intuitively, your friend must have said the word "tall" for a reason. One possibility is that your friend wants to distinguish the dax they want from another dax they do not. In this case, you might look around the room for two similar things that vary in height, and hand the taller one to them. If, alternatively, you only see one object around whose name you don't know, you might draw a different inference: this dax might be a particularly tall dax. In this case, you might think your friend used the word "tall" for a different reason–not to distinguish the dax they want from other daxes around you, but to distinguish the dax they want from other daxes in the world. This would be consistent with data from production studies, in which people tend to describe atypical features more than they describe typical ones (Mitchell, Reiter, & Deemter, 2013; Rubio-Fernández, 2016a; Westerbeek et al., 2015a). For instance, people almost always say "blue banana" to refer to a blue banana, but almost never say "yellow banana" to refer to a yellow one.

In each of these cases, you would have used a pragmatic inference to learn something new. In the second case, you would have learned the name for a novel category "dax," and also something about the typical of size of daxes: most of them are shorter than the one you saw. In the first case, you would have resolved the referential ambiguity in the speaker's utterance. But would have you learned something about the typical size of daxes as well, beyond the daxes you observed? One possibility is that you would not: You can explain your friend's use of "tall" as being motivated by the need to distinguish between the two daxes in the room, and thus you should infer nothing about the other daxes in the world. If reference is the primary motivator of speakers' word choice, as implicitly assumed in much research (e.g., Pechmann, 1989; Arts, Maes, Noordman, & Jansen, 2011; Engelhardt et al., 2011), then people should draw no further inferences once the need for referential disambiguation can explain away a descriptor like "tall." On this reference-first view, establishing reference has priority in understanding the utterance, and any further inferences are blocked if the utterance is minimally informative with respect to reference. If, on the other hand, pragmatic reasoning weighs multiple goals simultaneously–here, reference and conveying typicality–people may integrate typicality as just one factor the speaker considers in using description, leading to graded inferences about the referent's identity and about its category's features.

This dissertation will explore the ways in which people can learn about new words and categories from contrastive inference, with an eye toward understanding how contrastive inference could help children learn language. In Chapter 1, we will establish that adults can use contrastive inferences both to learn the name of a new object and to learn its category's feature distributions. Using contrastive inference to learn useful information about a category's feature distribution relies on speakers consistently using adjectives to point out atypical features. Thus, in Chapter 2, we will show that both adults speaking to other adults and caregivers speaking to children tend to mention the atypical rather than the typical features of things. In Chapter 3, we will test whether children are able to use contrastive inferences to learn about the feature distributions of new categories.

Chapter 1

Experiment 1

In Experiment 1, we ask whether people use descriptive contrast to identify the target of an ambiguous referring expression. Our experiment was inspired by work from Sedivy et al. (1999) showing that people can use contrastive inferences to guide their attention to referents as utterances progress. In their task, participants saw displays of four objects: a target (e.g., a tall cup), a contrastive pair (e.g., a short cup), a competitor that shares the target's feature but not category (e.g., a tall pitcher), and an irrelevant distractor (e.g., a key). Participants then heard a referring expression: "Pick up the tall cup." Participants looked more quickly to the correct object when the utterance referred to an object with a same-category contrastive pair (tall cup vs. short cup) than when it referred to an object without a contrastive pair (e.g., when there was no short cup in the display).

Their results suggest that listeners expect speakers to use prenominal description when they are distinguishing between potential referents of the same type, and listeners use this inference to rapidly allocate their attention to the target as an utterance progresses. This principle does not apply equally across adjective types, however: color adjectives seem to hold less contrastive weight (Sedivy, 2003), perhaps because color adjectives are often used redundantly in English–that is, people describe objects' colors even when this description is not necessary to establish reference (Pechmann, 1989). Kreiss and Degen (2020) demonstrate that listeners' familiar referent choices closely conform to speakers' production norms, such that over-specified modifiers hold less contrastive weight. If this generalizes to novel object choice, we should find that size adjectives prompt stronger contrastive inferences than color adjectives.

In a pre-registered referential disambiguation task, we presented participants with arrays of novel fruit objects. On critical trials, participants saw a target object, a lure object that shared the target's critical feature but not its shape, and a contrastive pair that shared the target's shape but not its critical feature (Fig. 1). Participants heard an utterance, sometimes mentioning the critical feature: "Find the [blue/big] toma." In all trials, utterances used the definite determiner "the," which conveys that there is a specific referent to be identified. For the target object, which had a same-shaped counterpart, use of the adjective was necessary to establish reference. For the lure, which was unique in shape, the adjective was relatively superfluous description. If participants use contrastive inference to choose novel referents, they should choose the target object more often than the lure. To examine whether contrast occurs across adjective types, we tested participants in two conditions: color contrast and size contrast. Though we expected participants to shift toward choosing the item with a contrastive pair in both conditions, we did not expect them to treat color and size equally. Because color is often used redundantly in English while size is not, we expected size to hold more contrastive weight, encouraging a more consistent contrastive inference (Pechmann, 1989). The pre-registration of our method, recruitment plan, exclusion criteria, and analyses can be found on the Open Science Framework here: https://osf.io/pqkfy .

Method

Participants. We recruited a pre-registered sample of 300 participants through Amazon Mechanical Turk. Half of the participants were assigned to a condition in which the

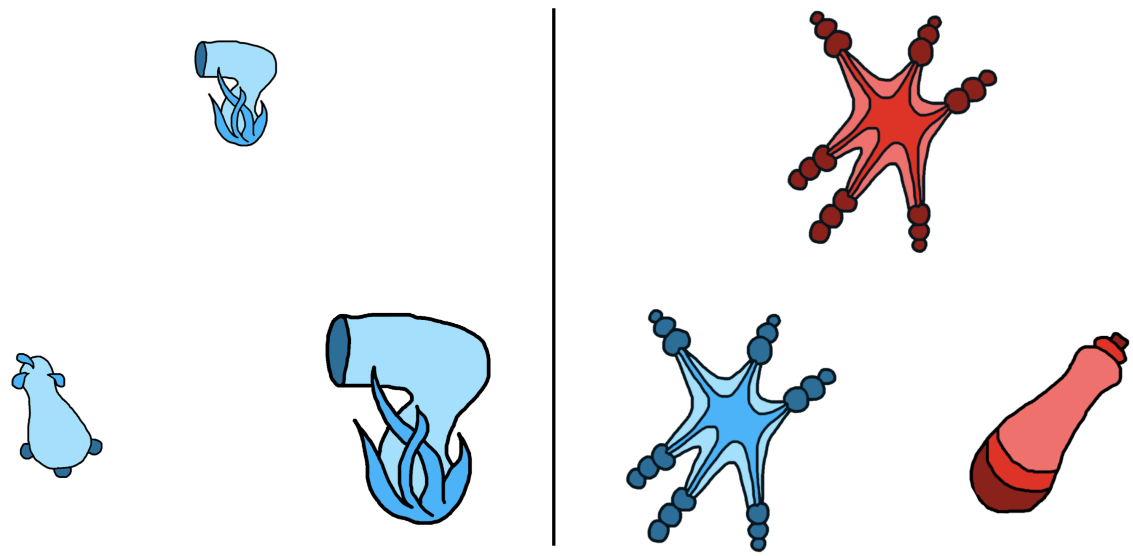

Figure 1. On the left: an example of a contrastive trial in which the critical feature is size. Here, the participant would hear the instruction "Find the small toma." On the right: an example of a contrastive trial in which the critical feature is color. Here, the participant would hear the instruction "Find the red toma." In both cases, the target is the top object.

critical feature was color (stimuli contrasted on color), and the other half were assigned to a condition in which the critical feature was size. Each participant gave informed consent and was paid $0.30 in exchange for their participation.

Stimuli. Stimulus displays were arrays of three novel fruit objects. Fruits were chosen randomly at each trial from 25 fruit kinds. Ten of the 25 fruit drawings were adapted and redrawn from Kanwisher, Woods, Iacoboni, and Mazziotta (1997); we designed the remaining 15 fruit kinds. Each fruit kind had an instance in each of four colors (red, blue, green, or purple) and two sizes (big or small). Particular target colors were assigned randomly at each trial and particular target sizes were counterbalanced across display types. There were two display types: unique target displays and contrastive displays. Unique target displays contained a target object that had a unique shape and was unique on the trial's critical feature (color or size), and two distractor objects that matched each other's (but not the target's) shape and critical feature. These unique target displays were included as a

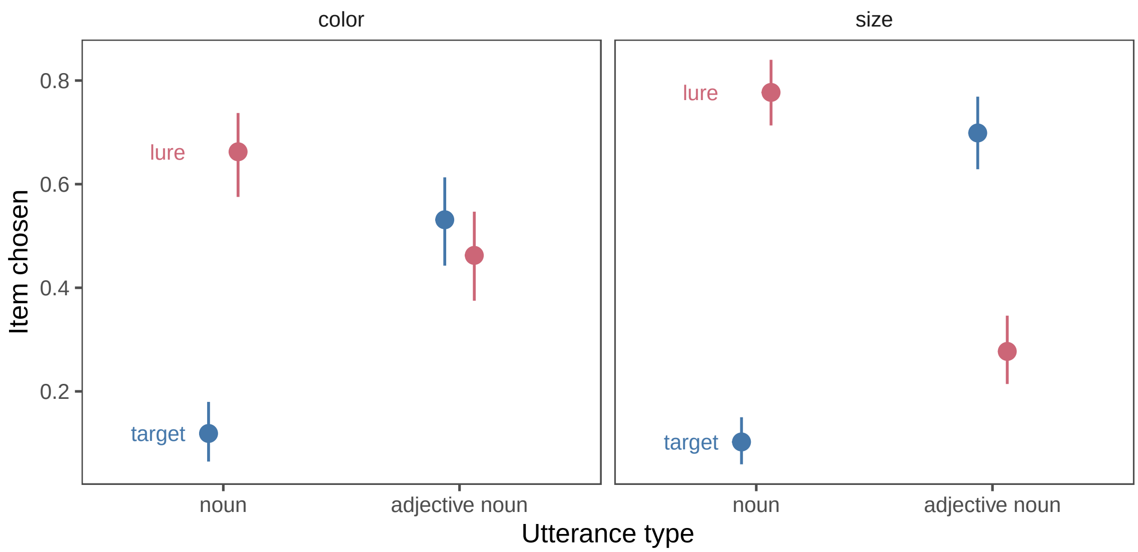

Figure 2. Proportion of times that participants chose the target and lure items as a function of condition and whether an adjective was provided. Points indicate group means; error bars indicate 95% confidence intervals computed by non-parametric bootstrapping.

check that participants were making reasonable referent choices and to space out contrastive displays to prevent participants from dialing in on the contrastive object setup during the experiment. Contrastive displays contained a target, its contrastive pair (matched the target's shape but not its critical feature), and a lure (matched the target's critical feature but not its shape; Fig. 1). The on-screen positions of the target and distractor items were randomized within a triad configuration.

Design and Procedure. Participants were told they would play a game in which they would search for strange alien fruits. Each participant saw eight trials. Half of the trials were unique target displays and half were contrastive displays. Crossed with display type, half of trials had audio instructions that described the critical feature of the target (e.g., "Find the [blue/big] toma"), and half of trials had audio instructions with no adjective description (e.g., "Find the toma"). A name was randomly chosen at each trial from a list of eight nonce names: blicket, wug, toma, gade, sprock, koba, zorp, and lomet.

After completing the study, participants were asked to select which of a set of alien words they had heard previously during the study. Four were words they had heard, and four were novel lure words. Participants were dropped from further analysis if they did not meet our pre-registered exclusion criteria of responding to at least 6 of these 8 memory check questions correctly (above chance performance as indicated by a one-tailed binomial test at the p = .05 level) and answering all four color perception check trials correctly (resulting n = 163)[1].

Results

We first confirmed that participants understood the task by analyzing performance on unique target trials, the filler trials in which there was a target unique on both shape and the relevant adjective. We asked whether participants chose the target more often than expected by chance (33%) by fitting a mixed effects logistic regression with an intercept term, a random effect of subject, and an offset of $logit(1/3)$ to set chance probability to the correct level. The intercept term was reliably different from zero for both color (β = 6.64, t = 4.10, p < .001) and size (β = 2.25 , t = 6.91, p < .001), indicating that participants consistently chose the unique object on the screen when given an instruction like "Find the (blue) toma." In addition, participants were more likely to select the target when an adjective was provided in the audio instruction in both conditions. We confirmed this effect statistically by fitting a mixed effects logistic regression predicting target selection from condition, adjective use, and their interaction with random effects of participants. Use of description in the audio increased target choice (β = 3.85, t = 3.52, p < .001), and adjective type (color vs. size) was not statistically related to target choice (β = -0.48, t = -1.10, p = .269). The two effects had a marginal interaction (β = -2.24, t = -1.95, p = .051). Participants had a general tendency to choose the target in unique target trials, which was strengthened if the audio instruction

[1] Experiments 1 and 3 were run in 2020, during the COVID-19 pandemic, when high exclusion rates on Amazon Mechanical Turk were being reported by many experimenters. Though our pre-registered criteria led to many exclusions, the check given to participants tested memory for a few novel words heard in the experiment, which we do not believe was an overly stringent requirement.

contained the relevant adjective. These effects did not significantly differ between color and size adjectives, which suggests that participants did not treat color and size differently in these baseline trials, though the marginal interaction suggests that use of an adjective may strengthen their tendency to choose the unique object more powerfully in the size condition.

Our key pre-registered analysis was whether participants would choose the target object on contrastive trials–when they heard an adjective in the referring expression. To perform this test, we compared participants' rate of choosing the target to their rate of choosing the lure, which shares the relevant critical feature with the target, when they heard the adjective. Overall, participants chose the target with a contrasting pair more often than the unique lure, indicating that they used contrastive inferences to resolve reference ($\beta = 0.53$, $t = 3.83$, $p = < .001$). To test whether the strength of the contrastive inference differed between color and size conditions, we pre-registered a version of this regression with a term for adjective type, and found that people were more likely to choose the target over the lure in the size condition than the color condition ($\beta = 0.87$, $t = 3.12$, $p = .002$).

Given this result, we tested whether people consistently chose the target over the lure on the color and size data separately, as a stricter check of whether the effect was present in both conditions (not pre-registered). Considering color and size separately, participants chose the target significantly more often than the lure in the size condition ($\beta = 0.86$, $t = 4.41$, $p = < .001$), but not in the color condition ($\beta = 0.15$, $t = 0.75$, $p = .455$). On contrastive trials in which a descriptor was not given, participants dispreferred the target, instead choosing the lure object, which matched the target on the descriptor but had a unique shape ($\beta =$ -2.65, $t =$ -5.44, $p = < .001$). Participants' choice of the target in the size condition was therefore not due to a prior preference for the target in contrastive displays, but relied on contrastive interpretation of the adjective. In the Supplemental Materials, we report an additional pre-registered analysis of all Experiment 1 data with maximal terms and random effects; those results are consistent with the more focused tests reported here.

Discussion

When faced with unfamiliar objects referred to by unfamiliar words, people can use pragmatic inference to resolve referential ambiguity and learn the meanings of these new words. In Experiment 1, we found that people have a general tendency to choose objects that are unique in shape when reference is ambiguous. However, when they hear an utterance with description (e.g., "blue toma", "small toma"), they shift away from choosing unique objects and toward choosing objects that have a similar contrasting counterpart. Furthermore, use of size adjectives–but not color adjectives–prompts people to choose the target object with a contrasting counterpart more often than the unique lure object. We found that people are able to use contrastive inferences about size to successfully resolve which unfamiliar object an unfamiliar word refers to.

Model

To formalize the inference that participants were asked to make, we developed a model in the Rational Speech Act Framework (RSA, Frank & Goodman, 2012). In this framework, pragmatic listeners (L) are modeled as drawing inferences about speakers' (S) communicative intentions in talking to a hypothetical literal listener (L_0). This literal listener makes no pragmatic inferences at all, evaluating the literal truth of a statement (e.g., it is true that a red toma can be called "toma" and "red toma" but not "blue toma"), and chooses randomly among all referents consistent with that statement. In planning their referring expressions, speakers choose utterances that are successful at accomplishing two goals: (1) making the listener as likely as possible to select the correct object, and (2) minimizing their communicative cost (i.e., producing as few words as possible). Note that though determiners are not given in the model's utterances, the assumption that the utterance refers to a specific reference is built into the model structure, consistent with the definite determiners used in the task. Pragmatic listeners use Bayes' rule to invert the speaker's utility function, essentially inferring what the speaker's intention was likely to be given the utterance they produced.

$$Literal: P_{Lit} = \delta\left(u, r\right) P\left(r\right)$$

$$Speaker: P_S\left(u|r\right) \propto \alpha\left(P_{Lit}\left(r|u\right) - C\right)$$

$$Listener: P_{Learn}\left(r|u\right) \propto P_s\left(u|r\right) P\left(r\right)$$

For this experiment, we build on a Rational Speech Act model developed by Frank and Goodman (2014) to jointly resolve reference and learn new words. The primary modification of RSA is use of a pragmatic *learner*: a pragmatic listener who has uncertainty about the meanings of words in their language, and thus cannot directly compute the speaker's utility as written. Instead, the speaker's utility is conditioned on the set of mappings, and the learner must also infer which set of mappings is correct:

$$Learner: P_L\left(r|u\right) \propto P_s\left(u|r; m\right) P\left(r\right) P\left(m\right)$$

In these experiments, we assume that the prior probability to refer to each object $(P\left(r\right))$ is equal, and similarly that all mappings $(P\left(m\right))$ are equally likely, so they cancel out in computations. We further assume that the cost of producing any word is identical, and so the cost of an utterance is equal to its length. All that remains is to specify the possible mappings, and literal meanings, and alternative utterances possible on each trial of the experiment. We describe the size condition here, but the computation for the color condition is analogous.

On the trial shown in the left panel of Figure 1 people see two objects that look something like a hair dryer and one that looks like a pear and they are asked to "Find the toma." Here, in the experiment design and the model, we take advantage of the fact that English speakers tend to assume that nouns generally correspond to differences in shape rather than other features (Landau, Smith, & Jones, 1992). Given this, the two possible mappings are $\{m_1 : hairdryer - \text{“}toma\text{”}, pear - \text{“?”}\}$ and $\{m_2 : hairdryer - \text{“?”}, pear - \text{“}toma\text{”}\}$. The literal semantics of each object allow them to be referred to by their shape label (e.g. “toma”), or by a descriptor that is true of them (e.g. “small”), but not names for other shapes or untrue descriptors.

Having heard “Find the toma,” the model must now choose a referent. If the true mapping for “toma” is the hair dryer (m_1), this utterance is ambiguous to the literal listener, as there are two referents consistent with the literal meaning toma. Consequently, whichever of the two referents the speaker intends to point out to the learner, the speaker's utility will be relatively low. Alternatively, if the true mapping for “toma” is the pear (m_2), then the utterance will be unambiguous to the literal listener, and thus the speaker's utterance will have higher utility. As a result, the model can infer that the more likely mapping is m_2 and choose the pear, simultaneously resolving reference and learning the meaning of “toma.”

If instead the speaker produced “Find the small toma,” the model will make a different inference. If the true mapping for “toma” is hair dryer (m_2), this utterance now uniquely identifies one referent for the literal listener and thus has high utility. It also uniquely identifies the target if “toma” means pear (m_1). However, if “toma” means pear, the speaker's utterance was inefficient because the single word utterance “toma” would have identified the target to the literal listener and incurred less cost. Thus, the model can infer that “toma” is more likely to mean hair dryer and choose the small hair dryer appropriately.

While these descriptions use deterministic language for clarity, the model's computation is probabilistic and thus reflects tendencies to choose those objects rather than fixed rules. Figure 3 shows model predictions alongside people's behavior for the size and color contrast conditions in Experiment 1. In line with the intuition above, the model predicts that hearing a bare noun (e.g. “toma”) should lead people to infer that the intended referent is the unique object (lure), whereas hearing a modified noun (e.g. “small toma”)

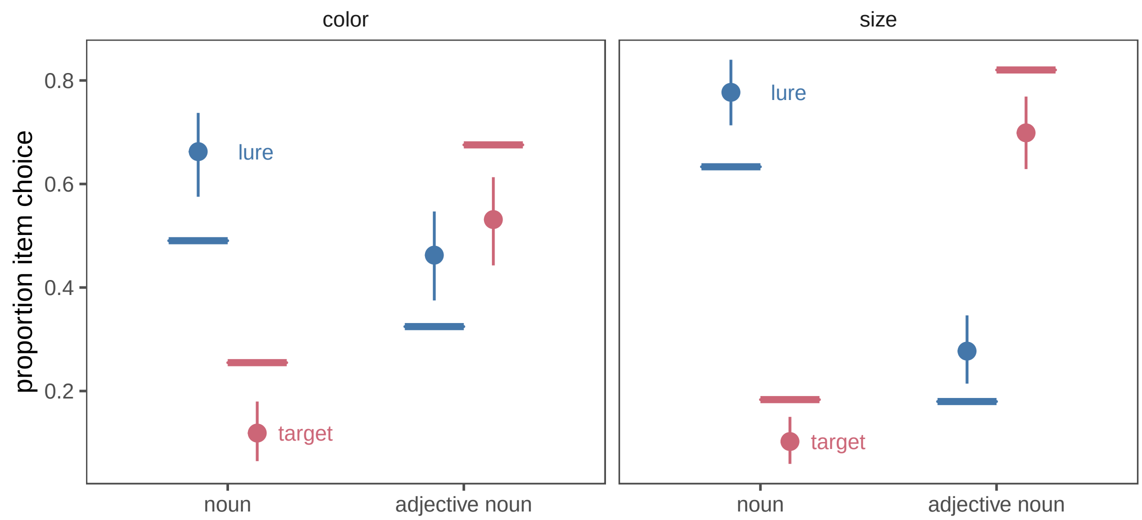

Figure 3. Proportion of times that people (and our model) chose the target and lure items as a function of adjective type and whether an adjective was provided. Points indicate empirical means; error bars indicate 95% confidence intervals computed by non-parametric bootstrapping. Solid horizontal lines indicate model predictions.

should lead people to infer that the speaker's intended referent has a same-shaped counterpart without the described feature (i.e., is the target object).

Our empirical data suggest that people treat color and size adjectives differently, making a stronger contrastive inference with size than with color. One potential explanation for this difference is that people are aware of production asymmetries between color and size. As mentioned, speakers tend to over-describe color, providing more color adjectives than necessary to establish reference, while describing size more minimally (Nadig & Sedivy, 2002; Pechmann, 1989). Listeners may be aware of this production asymmetry and discount the contrastive weight of color adjectives with respect to reference.

In the Rational Speech Act model, this kind of difference is captured neatly by a difference in the listener's beliefs about the speaker's rationality (i.e. how sensitive the speaker is to differences in utility of different utterances). To determine the value of the rationality parameter that best describes participants' behavior in each condition, we used

Bayesian data analysis, estimating the posterior probability of the observed data under each possible value of α multiplied by the prior probability of each of those values. In each condition, α was drawn from a Gamma distribution with shape and scale parameters set to 2 ($Gamma\,(2, 2)$). This prior encodes a weak preference for small values of α, but the estimates below were not sensitive to other choices of hyper-parameters.

Posterior mean estimates of rationality varied substantially across conditions. In the color condition, the rationality parameter was estimated to be 2.00 with a 95% credible interval of [1.37, 2.63]. In the size condition, rationality was estimated to be 3.98 [3.22, 4.74].

Figure 3 shows the model predictions along with the empirical data from Experiment 1. The model broadly captures the contrastive inference–when speakers produce an adjective noun combination like "red toma," the model selects the target object more often than the lure object. The extent to which the model makes this inference varies as predicted between the color and size adjective conditions in line with the different estimated rationality values. In both conditions, despite estimating the value of rationality that makes the observed data most probable, the model overpredicts the extent of the contrastive inference that people make. Intuitively, it appears that over and above the strength of their contrastive inferences, people have an especially strong tendency to choose a unique object when they hear an unmodified noun (e.g. "toma"). In an attempt to capture this uniqueness tendency, the model overpredicts the extent of the contrastive inference.

The model captures the difference between color and size in a difference in the rationality parameter, but leaves open the ultimate source of this difference in rationality. Why there is a production asymmetry in the first place? For now, we bracket this question and note that listeners in our task appropriately discount color's contrastive weight given production norms.

An alternative way to capture this preference would be to locate it in a different part of the model. One possibility is that the literal semantics of color and size work differently. A recent model from Degen, Hawkins, Graf, Kreiss, and Goodman (2020) does predict a color–size asymmetry based on different semantic exactness. In this model, literal semantics are treated as continuous rather than discrete, so "blue" is neither 100% true nor 100% false of a particular object, but can instead be 90% true. They successfully model a number of color–size asymmetries in production data by treating color as having stronger literal semantics (e.g. "blue toma" is a better description of a small blue toma than "small toma" is). However, this model predicts the opposite asymmetry of what we found. Because color has stronger semantics than size, the listener in this model shows a stronger contrast effect for color than size (see demonstration in the Supplemental Materials). Thus, though a continuous semantics can explain our asymmetry, this explanation is unlikely given that the continuous semantics that predicts other empirical color–size asymmetries does not predict our findings.

Yet another way to explain the difference between size and color adjectives is to attribute size adjectives' contrastive strength with respect to reference to the fact that size adjectives are gradable and relative. There are multiple ways to implement this possibility in the model. One way would be to specify that speakers tend to remark on relative, gradable features when making distinctions among present objects because direct comparisons for the meaning of 'small' and 'big' are at hand, whereas color adjectives are more often mentioned superfluously because they have more absolute meaning and do not need immediate comparisons. This possiblity is consistent with the model we have specified, and is just one possible reason for a production asymmetry which listeners are responding to rationally in their inferences. Another possibility is that the gradable, relative nature of size adjectives should be encoded in the pragmatic learner part of the model: a learner might need a comparison point to tell whether a novel object is small or big, but not red or purple, and thus avoid choosing a unique (shaped) object when size is specified but be willing to choose a unique object when color is specified. This possibility would require more fundamental changes to the model. Here, we make the conservative choice to encode the color-size asymmetry in the broad rationality parameter, though changing the pragmatic learner's decision process is an intriguing possibility for future work.

Overall, we found that people can use contrastive inferences from description to map an unknown word to an unknown object. This inference is captured by an extension of the Rational Speech Act model using a pragmatic learner, who is simultaneously making inferences over possible referents and possible lexicons. This model can also capture people's tendency to make stronger contrastive inferences from color description than size description through differences in the rationality parameter, though the origin of these differences cannot be pinned down with this experiment alone. Our experiment and model results suggest that people can resolve a request like "Give me the small dax" by reasoning that the speaker must have been making a useful distinction by mentioning size, and therefore looking for multiple similar objects that differ in size and choosing the smaller one. Immediately available objects are not the only ones worth making a distinction from, however. Next, we turn to another salient set of objects a speaker might want to set a referent apart from: the referent's category.

Experiment 2

When referring to a *big red dog* or a *hot-air balloon*, we often take care to describe them—even when there are no other dogs or balloons around. Speakers use more description when referring to objects with atypical features (e.g., a yellow tomato) than typical ones (e.g., a red tomato; Mitchell et al., 2013; Bergey, Morris, & Yurovsky, 2020; Rubio-Fernández, 2016a; Westerbeek et al., 2015a). This selective marking of atypical objects potentially supplies useful information to listeners: they have the opportunity to not only learn about the object at hand, but also about its broader category. Alexandra C. Horowitz and Frank (2016) demonstrated that, combined with other contrastive cues (e.g., "Wow, this one is a zib. This one is a TALL zib"), prenominal adjectives prompted adults and children to infer that the described referent was less typical than one that differed on the mentioned feature (e.g., a shorter zib).

Further, this kind of contrast may help make sense of the asymmetry between color and size adjectives we found in Experiment 1. Color adjectives that are redundant with respect to reference are not necessarily redundant in general. Rubio-Fernández (2016a) demonstrates that speakers often use 'redundant' color adjectives to describe colors when they are variable and central to the category's meaning (e.g., colorful t-shirts) or when they are atypical (e.g., a purple banana). Comprehenders, in turn, expect color adjectives to be used informatively with respect to typicality, and upon hearing color adjectives tend to look to referents for which the adjective describes a less-typical feature (e.g., "Choose the yellow…" prompts people to look to a yellow shirt over a yellow banana; Rohde & Rubio-Fernandez, 2021). Therefore, while size may hold more contrastive weight with respect to reference, color and size may hold similar contrastive weight with respect to the category's feature distribution. In Experiment 2, we test whether listeners use descriptive contrast with a novel object's category to learn about the category's feature distribution.

If listeners do make contrastive inferences about typicality, it may not be as simple as judging that an over-described referent is atypical. Description can serve many purposes. In Experiment 1, we investigated its use in contrasting between present objects. If a descriptor was needed to distinguish between two present objects, it may not have been used to mark atypicality. For instance, in the context of a bin of heirloom tomatoes, a speaker who wanted a red one in particular might specify that they want a "red tomato" rather than just asking for a "tomato." In this case, the adjective "red" is being used contrastively with respect to reference (as in Experiment 1), and not to mark atypicality. Thus, a listener who does not know much about tomatoes may attribute the use of "red" to referential disambiguation given the context and not infer that red is an unusual color for tomatoes.

In Experiment 2, we used an artificial language task to set up just this kind of learning situation. We manipulated the contexts in which listeners hear adjectives modifying novel names of novel referents. These contexts varied in how useful the adjective was to identify the referent: in one context the adjective was necessary, in another it was helpful, and in a third it was entirely redundant. On a reference-first view, use of an adjective that was necessary for reference can be explained away and should not prompt further inferences about typicality–an atypicality inference would be blocked. If, on the other hand, people take into account speakers' multiple reasons for using adjectives without giving priority to reference, they may alter their inferences about typicality across these contexts in a graded way: if an adjective was necessary for reference, it may prompt slightly weaker inferences of atypicality; if an adjective was redundant with respect to reference, it may be inferred to mark atypicality more strongly. Further, these contexts may also prompt distinct inferences when no adjective is used: for instance, when an adjective is necessary to identify the referent but elided, people may infer that the elided feature is particularly typical. To account for the multiple ways context effects might emerge, we analyze both of these possibilities. Overall, we asked whether listeners infer that these adjectives identify atypical features of the named objects, and whether the strength of this inference depends on the referential ambiguity of the context in which adjectives are used.

Method

Participants. 240 participants were recruited from Amazon Mechanical Turk. Half of the participants were assigned to a condition in which the critical feature was color (red, blue, purple, or green), and the other half of participants were assigned to a condition in which the critical feature was size (small or big).

Stimuli & Procedure. Stimulus displays showed two alien interlocutors, one on the left side (Alien A) and one on the right side (Alien B) of the screen, each with two novel fruit objects beneath them (Figure 4). Alien A, in a speech bubble, asked Alien B for one of its fruits (e.g., "Hey, pass me the big toma"). Alien B replied, "Here you go!" and the

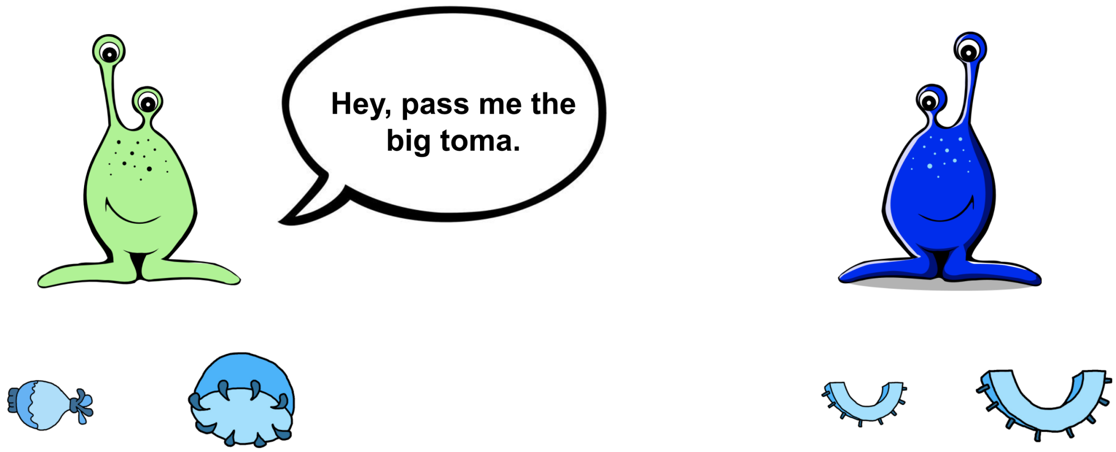

Figure 4. Experiment 2 stimuli. In the above example, the critical feature is size and the object context is a within-category contrast: the alien on the right has two same-shaped objects that differ in size.

referent disappeared from Alien B's side and reappeared on Alien A's side.

We manipulated the critical feature type (color or size) between subjects. Two factors (presence of the critical adjective in the referring expression and object context) were fully crossed within subjects. Object context had three levels: within-category contrast, between-category contrast, and same feature (Figure 5). In the within-category contrast condition, Alien B possessed the target object and another object of the same shape, but with a different value of the critical feature (e.g., a big toma and a small toma). In the between-category contrast condition, Alien B possessed the target object and another object of a different shape, and with a different value of the critical feature (e.g., a big toma and a small blicket). In the same feature condition, Alien B possessed the target object and another object of a different shape but with the same value of the critical feature as the target (e.g., a big toma and a big dax). Thus, in the within-category contrast condition, the descriptor was necessary to distinguish the referent; in the between-category contrast condition it was unnecessary but potentially helpful; and in the same feature condition it was unnecessary and unhelpful.

Note that in all context conditions, the set of objects on screen was the same in terms of the experiment design: there was a target (e.g., big toma), an object with the same shape as the target and a different critical feature (e.g., small toma), an object with a different shape from the target and the same critical feature (e.g., big dax), and an object with a different shape from the target and a different critical feature (e.g., small blicket). Context was manipulated by rearranging these objects such that the relevant referents (the objects under Alien B) differed and the remaining objects were under Alien A. Thus, in each case, participants saw the target object and one other object that shared the target object's shape but not its critical feature–they observed the same kind of feature distribution of the target object's category in each trial type. The particular values of the features were chosen randomly for each trial.

Participants completed six trials. After each exchange between the alien interlocutors, they made a judgment about the prevalence of the target's critical feature in the target object's category. For instance, after seeing a red blicket being exchanged, participants would be asked, "On this planet, what percentage of blickets do you think are red?" They would answer on a sliding scale between zero and 100. In the size condition, participants were asked, "On this planet, what percentage of blickets do you think are the size shown below?" with an image of the target object they just saw available on the screen.

After completing the study, participants were asked to select which of a set of alien words they had seen previously during the study. Four were words they had seen, and four were novel lure words. Participants were dropped from further analysis if they did not respond to at least 6 of these 8 correctly (above chance performance as indicated by a one-tailed binomial test at the $p = .05$ level). This resulted in excluding 47 participants, leaving 193 for further analysis.

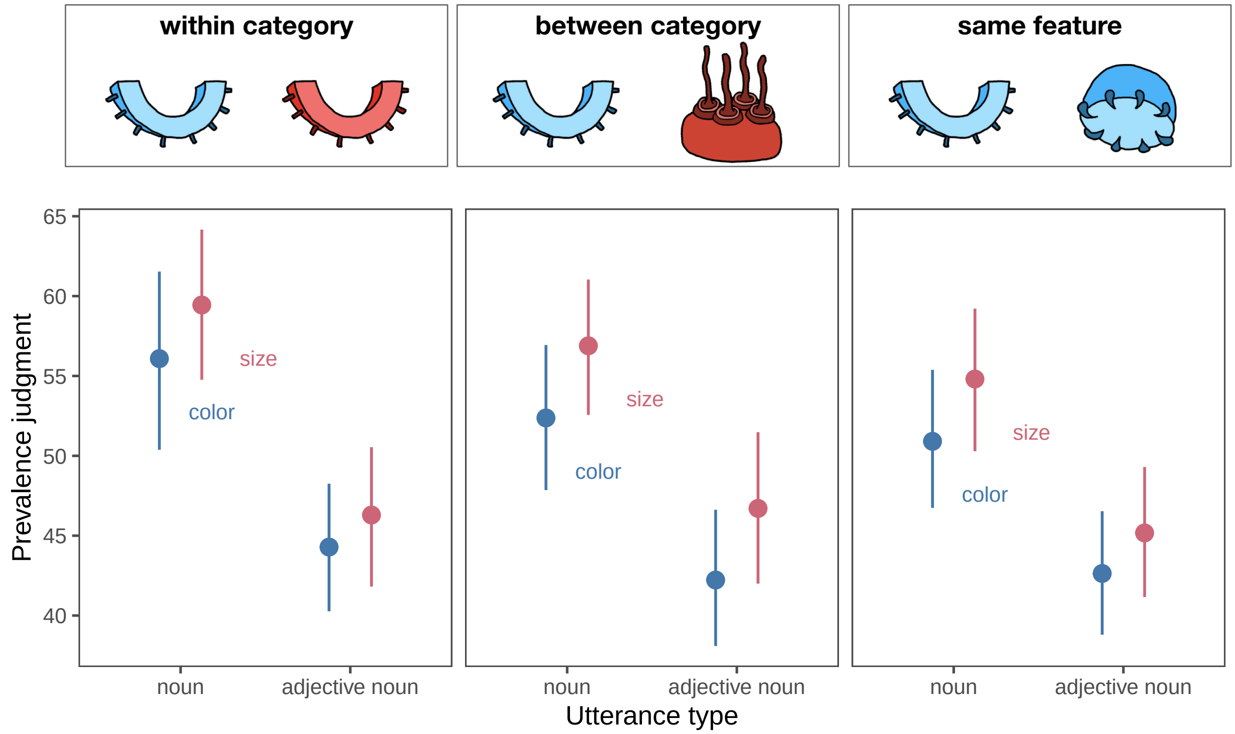

Figure 5. Prevalence judgments from Experiment 2. Participants consistently judged the target object as less typical of its category when the referent was described with an adjective (e.g., "Pass me the blue toma") than when it was not (e.g., "Pass me the toma"). This inference was not significantly modulated by object context (examples shown above each figure panel).

Results

Our key test is whether participants infer that a mentioned feature is less typical than one that is not mentioned. In addition, we tested whether inferences of atypicality are modulated by context. One way to test this is to analyze the interaction between utterance type and context, seeing if the difference between adjective and no adjective utterances is larger when the adjective was highly redundant or smaller when the adjective was necessary for reference.

We analyzed participants' judgments of the prevalence of the target object's critical feature in its category. We began by fitting a maximum mixed-effects linear model with effects of utterance type (adjective or no adjective), context type (within category, between category, or same feature, with between category as the reference level), and critical feature (color or size) as well as all interactions and random slopes of utterance type and context type nested within subject. Random effects were removed until the model converged. The final model included the effects of utterance type, context type, and critical feature and their interactions, and a random slope of utterance type by subject. This model revealed a significant effect of utterance type ($\beta_{adjective}$ = -10.22, t = -3.37, p = .001), such that prevalence judgments were lower when an adjective was used than when it was not. Participants' inferences did not significantly differ between color and size adjective conditions (β_{size} = 4.73, t = 1.46, p = .146). Participants' inferences did not significantly vary by context type (β_{within} = 3.92, t = 1.63, p = .104; β_{same} = -1.48, t = -0.62, p = .537). There was not a significant interaction between context and presence of an adjective in the utterance ($\beta_{within*adjective}$ = -1.58, t = -0.46, p = .644; $\beta_{same*adjective}$ = 2.13, t = 0.63, p = .532). That is, participants did not significantly adjust their inferences based on object context, nor did they make differential inferences based on the combination of context and adjective use. However, they robustly inferred that mentioned features were less prevalent in the target's category than unmentioned features.

This lack of a context effect may be because people do not take context into account, or because they make distinct inferences when an adjective is *not* used: for instance, when an adjective is necessary for reference but elided, people may infer that the unmentioned feature is very typical. This inference would lead to a difference between the adjective and no adjective utterances in the within-category context, but not because people are failing to attribute the adjective to reference. To account for this possibility, we additionally test for differences in the context conditions among only the utterances with adjectives. We fit a model with effects of context type and critical feature as well as their interaction and random slopes by subject. Participants did not significantly adjust their inferences by context among only the adjective utterances (β_{within} = 2.43, t = 1.16, p = .247; β_{same} = 0.67,

$t = 0.32$, $p = .750$). Thus, even by this more specific test, participants did not adjust their inferences based on the referential context.

Discussion

Description is often used not to distinguish among present objects, but to pick out an object's feature as atypical of its category. In Experiment 2, we asked whether people would infer that a described feature is atypical of a novel category after hearing it mentioned in an exchange. We found that people robustly inferred that a mentioned feature was atypical of its category, across both size and color description. Further, participants did not use object context to substantially explain away description. That is, even when description was necessary to distinguish among present objects (e.g., there were two same-shaped objects that differed only in the mentioned feature), participants still inferred that the feature was atypical of its category. This suggests that, in the case of hearing someone ask for a "red tomato" from a bin of many-colored heirloom tomatoes, a person naive about tomatoes would infer that tomatoes are relatively unlikely to be red.

Unlike Experiment 1, in which people made stronger contrastive inferences for size than color, there were not substantial differences between people's inferences about color and size in Experiment 2. If an account based on production norms is correct, this suggests that people track both how often people use color compared to size description and also for what purpose–contrasting with present objects or with the referent's category. That is, color description may be more likely to be used superfluously with respect to present objects but informatively with respect to the category. Indeed, color description that seems overdescriptive with respect to object context often occurs when the category has many-colored members (e.g., t-shirts) or when the object's color is atypical (Rubio-Fernández, 2016a). However, our results are consistent with several potential explanations of the color-size asymmetry (or lack thereof). Future work addressing the source of the color-size asymmetry will need to explain differences in its extent when distinguishing among present objects compared to the referent's category.

Model

To allow the Rational Speech Act Framework to capture inferences about typicality, we modified the Speaker's utility function to have an additional term: the listener's expected processing difficulty. Speakers may be motivated to help listeners to select the correct referent not just eventually but as quickly as possible. People are both slower and less accurate at identifying atypical members of a category as members of that category (Dale, Kehoe, & Spivey, 2007; Rosch, Simpson, & Miller, 1976). If speakers account for listeners' processing difficulties, they should be unlikely to produce bare nouns to refer to low typicality exemplars (e.g. unlikely to call a purple carrot "carrot"). This is roughly the kind of inference encoded in Degen et al. (2020)'s continuous semantics Rational Speech Act model.

We model the speaker as reasoning about the listener's label verification process. Because the speed of verification scales with the typicality of a referent, a natural way of modeling it is as a process of searching for that particular referent in the set of all exemplars of the named category, or alternatively of sampling that particular referent from the set of all exemplars in that category, $P\left(r|Cat\right)$. On this account, speakers want to provide a modifying adjective for atypical referents because the probability of sampling them from their category is low, but the probability of sampling them from the modified category is much higher (a generalization of the size principle (Xu & Tenenbaum, 2007)). Typicality is just one term in the speaker's utility, and thus is directly weighed with the literal listener's judgment and against cost.

If speakers use this utility function, a listener who does not know the feature distribution for a category can use a speaker's utterance to infer it. Intuitively, a speaker should prefer not to modify nouns with adjectives because they incur a cost for producing an extra word. If they did use an adjective, it must be because they thought the learner would have a difficult time finding the referent from a bare noun alone because of typicality, competing referents, or both. To infer the true prevalence of the target feature in the category, learners combine the speaker's utterance with their prior beliefs about the feature distribution. We model the learner's prior about the prevalance of features in any category as a Beta distribution with two parameters α and β that encode the number of hypothesized prior psuedo-exemplars with the feature and without feature that the learner has previously observed (e.g., one red dax and one blue dax). We assume that the learner believes they have previously observed one hypothetical psuedo-examplar of each type, which is a weak symmetric prior indicating that the learner expects features to occur in half of all members of a category on average, but would find many levels of prevalence unsurprising. To model the learner's direct experience with the category, we add the observed instances in the experiment to these hypothesized prior instances. After observing one member of the target category with the relevant feature and one without, the listener's prior is thus updated to be Beta (2, 2).

As in Experiment 1, we used Bayesian data analysis and the same prior to estimate posterior mean rationality parameter that participants are using to draw inferences about speakers in both the color and size conditions. In contrast to Experiment 1, the absolute values of these parameters are driven largely by the number of pseudo-exemplars assumed by the listener prior to exposure. Thus, the rationality parameters inferred in the two experiments are not directly comparable. However, differences between color and size within each model are interpretable. As in Experiment 1, we found that listeners inferred speakers to be more rational when using size adjectives (0.89 [0.63, 1.13]) than color adjectives (0.60 [0.37, 0.83]), but the two inferred confidence intervals were overlapping, suggesting that people treated size and color adjectives similarly when making inferences about typicality.

Figure 6 shows the predictions of our Rational Speech Act model compared to empirical data from participants. The model captures the trends in the data correctly,

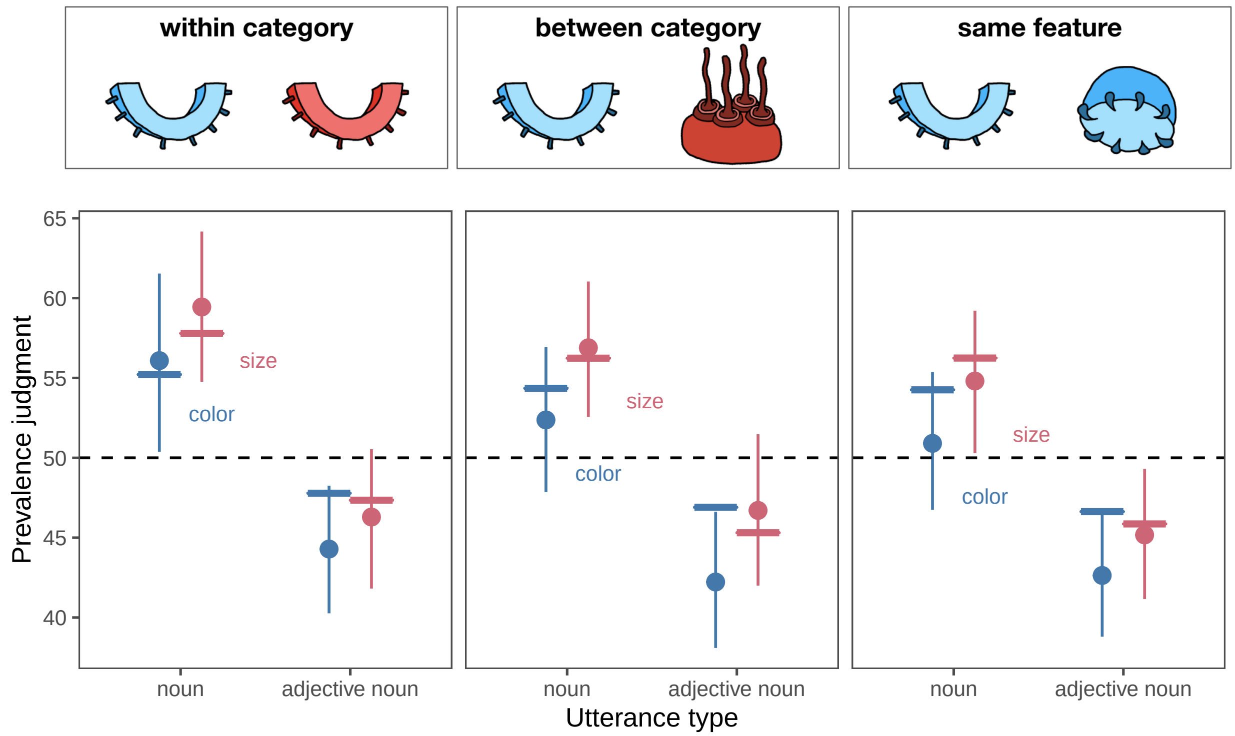

Figure 6. Participants' prevalence judgments from Experiment 2, compared to model predictions (horizontal lines).

inferring that the critical feature was less prevalent in the category when it was mentioned (e.g., "red dax") than when it was not mentioned (e.g., "dax"). The model also infers the prevalence of the critical feature to be numerically higher in the within-category condition, like people do. That is, in the within-category condition when an adjective is used to distinguish between referents, the model thinks that the target color is slightly less atypical. When an adjective would be useful to distinguish between two objects of the same shape but one is not used, the model infers that the color of the target object is slightly more typical.

Overall, our model captures the inference people make: when the speaker mentions a feature (e.g., "the blue dax"), that feature is inferred to be less typical of the category (daxes are less likely to be blue in general). It further captures that when the object context requires an adjective for successful reference, people weaken this atypicality inference only slightly, if at all. In contrast to a reference-first view, which predicts that these two kinds of inferences would trade off strongly–that is, using an adjective that is necessary for reference blocks the inference that it is marking atypicality–the model captures the graded way in which people consider these two communicative goals.

Experiment 3

In Experiments 1 and 2, we established that people can use contrastive inferences to resolve referential ambiguity and to make inferences about the feature distribution of a novel category. Additionally, in Experiment 2, we found that these two inferences do not seem to trade off substantially: even if an adjective is necessary to establish reference, people infer that it also marks atypicality. We also found that inferences of atypicality about color and size adjectives pattern very similarly, though their baseline typicality is shifted, while color and size are not equally contrastive with respect to referential disambiguation (Experiment 1).

To strengthen our findings in a way that would allow us to better detect potential trade-offs between these two types of inference, in Experiment 3 we conducted a pre-registered replication of Experiment 2 with a larger sample of participants. In addition, we tested how people's prevalence judgments from utterances with and without an adjective compare to their null inference about feature prevalence by adding a control utterance condition: an alien utterance, which the participants could not understand. This also tests the model assumption we made in Experiment 2: that after seeing two exemplars of the target object with two values of the feature (e.g., one green and one blue), people's prevalence judgments would be around 50%. In addition to validating this model assumption, we more strongly tested the model here by comparing predictions from same model, with parameters inferred from the Experiment 2 data, to data from Experiment 3. Our pre-registration of the method, recruitment plan, exclusion criteria, and analyses can be found on the Open Science Framework: https://osf.io/s8gre (note that this experiment is labeled Experiment 2 in the OSF repository but is Experiment 3 in the paper).

Method

Participants. A pre-registered sample of 400 participants was recruited from Amazon Mechanical Turk. Half of the participants were assigned to a condition in which the critical feature was color (red, blue, purple, or green), and half of the participants were assigned to a condition in which the critical feature was size (small or big).

Stimuli & Procedure. The stimuli and procedure were identical to those of Experiment 2, with the following modifications. Two factors, utterance type and object context, were fully crossed within subjects. Object context had two levels: within-category contrast and between-category contrast. In the within-category context condition, Alien B possessed the target object and another object of the same shape, but with a different value of the critical feature (color or size). In the between-category contrast condition, Alien B possessed the target object and another object of a different shape, and with a different value of the critical feature. Thus, in the within-category contrast condition, an adjective is necessary to distinguish the referent; in the between-category contrast condition it is unnecessary but potentially helpful. There were three utterance types: adjective, no adjective, and alien utterance. In the two alien utterance trials, the aliens spoke using completely unfamiliar utterances (e.g., "Zem, noba bi yix blicket"). Participants were told in the task instructions that sometimes the aliens would talk in a completely alien language, and sometimes their language will be partly translated into English. To keep participants from making inferences about the content of the alien utterances using the utterance content of other trials, both alien language trials were first; other than this constraint, trial order was random. We manipulated the critical feature type (color or size) between subjects.

After completing the study, participants were asked to select which of a set of alien words they had seen previously during the study. Four were words they had seen, and four were novel lure words. Participants were dropped from further analysis if they did not meet our pre-registered criteria of responding to at least 6 of these 8 correctly (above chance performance as indicated by a one-tailed binomial test at the p = .05 level) and answering all four color perception check questions correctly. Additionally, six participants were excluded because their trial conditions were not balanced due to an error in the run of the experiment. This resulted in excluding 203 participants, leaving 197 for further analysis. In our pre-registration, we noted that we anticipated high exclusion rates, estimating that approximately 150 people per condition would be sufficient to test our hypotheses.

Results

We began by fitting a pre-registered maximum mixed-effects linear model with effects of utterance type (alien utterance, adjective, or no adjective; alien utterance as reference level), context type (within category or between category), and critical feature (color or size) as well as all interactions and random slopes of utterance type and context type nested within subject. Random effects were removed until the model converged, which resulted in a model with all fixed effects, all interactions and a random slope of utterance type by subject. The final model revealed a significant effect of the no adjective utterance type compared to the alien utterance type (β = 7.48, t = 2.80, p = .005) and no significant effect of the adjective utterance type compared to the alien utterance type (β = -0.64, t = -0.24, p = .808). The effects of context type (within-category or between-category) and adjective type (color or size) were not significant (β_{within} = -2.70, t_{within} = -1.23, p_{within} = .220; β_{size} = 4.44, t_{size} = 1.33, p_{size} = .185). There were marginal interactions between the adjective utterance type and the size condition (β = -6.56, t = -1.72, p = .086), the adjective utterance type and the within-category context (β = 5.77, t = 1.86, p = .064), and the no adjective utterance type and the within-category context (β = 5.57, t = 1.79, p = .073). No other effects were significant or marginally significant. Thus, participants inferred that an object referred to in an intelligible utterance with no description was more typical of its category on the target feature than an object referred to with an alien utterance. Participants did not substantially adjust their inferences based on the object context. The marginal interactions between the within-category context and both the adjective and no adjective utterance types suggest that people might have judged the target feature as slightly more prevalent in the within-category context when intelligible utterances (with a bare noun or with an adjective) were used compared to the alien utterance. If people are discounting their atypicality inferences when the adjective is necessary for reference, we should expect them to have slightly higher typicality judgments in the within-category context when an adjective is used, and this marginal interaction suggests that this may be the case. However, since typicality judgments in the no adjective utterance type are also marginally greater in the within-category context, and because judgments in the alien utterance conditions (the reference category) also directionally move between the two context conditions, it is hard to interpret whether this interaction supports the idea that people are discounting their typicality judgments based on context.

Given that interpretation of these results with respect to the alien utterance condition can be difficult, we pre-registered a version of the same full model excluding alien utterance trials with the no adjective utterance type as the reference level. This model revealed a significant effect of utterance type: participants' prevalence judgments were lower when an adjective was used than when it was not (β = -8.12, t = -3.46, p = .001). No other effects were significant. This replicates the main effect of interest in Experiment 2: when an adjective is used in referring to the object, participants infer that the described feature is less typical of that object's category than when the feature goes unmentioned. It also shows that the possibility that people may discount their typicality judgments based on context (suggested by the marginal interaction described above) is not supported when we compare the adjective and no adjective utterance types directly. In the Supplemental Materials, we report two more pre-registered tests of the effect of utterance type alone on prevalence judgments whose results are consistent with the fuller models reported here.

As in Experiment 2, our test of whether participants' inferences are modulated by context is potentially complicated by people making distinct inferences when an adjective is necessary but *not* used. Thus, we additionally tested whether participants' inferences varied by context among only utterances with an adjective by fitting a model with effects of context and adjective type and their interaction, as well as random slopes by subject (not pre-registered). Participants' inferences did not significantly differ by context ($\beta_{within} = 3.07$, $t_{within} = 1.70$, $p_{within} = .091$). Thus, participants' inferences did not significantly differ between contexts, whether tested by the interaction between utterance type and contexts or by the effect of context among only utterances with an adjective.

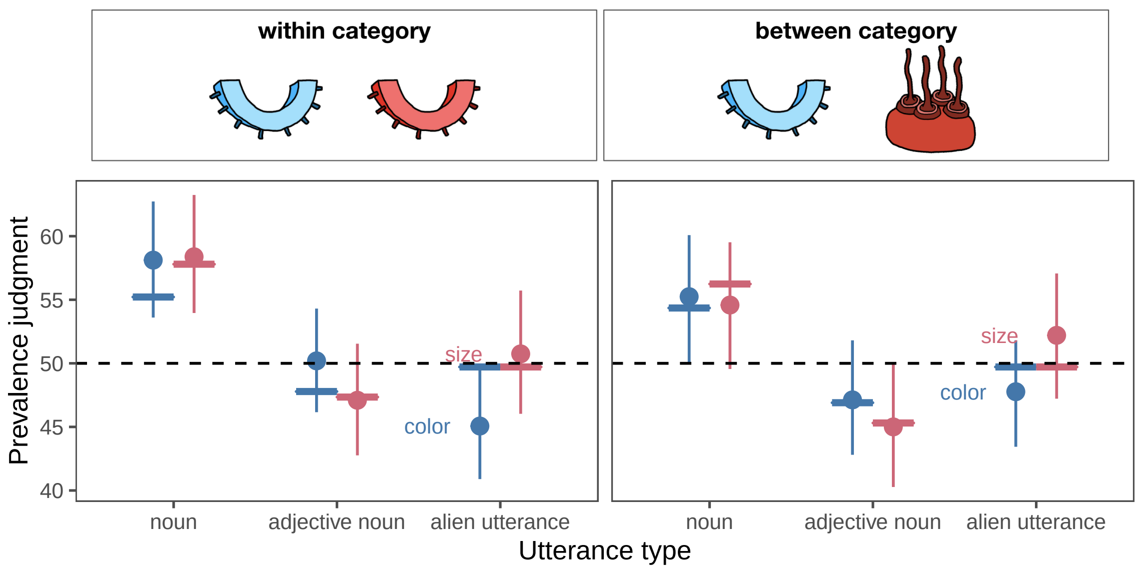

Figure 7. Participants' prevalence judgments in Experiment 3, with model predictions using the parameters estimated in Experiment 2 (horizontal lines).

Model

To validate the model we developed for Experiment 2, we compared its estimates using the previously fit parameters to the new data for Experiment 3. As shown in Figure 7, the model predictions were well aligned with people's prevalence judgments. In addition, in Experiment 2, we fixed the model's prior beliefs about the prevalence of the target object's color or size to be centered at 50% because the model had seen one pseudo-exemplar of the target color/size, and one psuedo-exemplar of the non-target color/size. In Experiment 3, we aimed to estimate this prior empirically in the alien utterance condition, reasoning that people could only use their prior to make a prevalence judgment (as we asked the model to do). In both the color and size conditions, people's judgments indeed varied around 50%, although in the color condition they were directionally lower. This small effect may arise from the fact that size varies on a scale with fewer nameable points (e.g., objects can be big, medium-sized or small) whereas color has many nameable alternatives (e.g., red, blue, green, etc.). Thus, the results of Experiment 3 confirm the modeling assumptions we made in estimating people's prior beliefs, and further validate the model we developed as a good candidate model for how people simultaneously draw inferences about speakers' intended referents and the typicality of these referents. That is, when people think about why a speaker chose their referring expression, they consider the context of not only present objects, but also the broader category to which the referent belongs.

Discussion

In Experiment 3, we replicated the main finding of interest in Experiment 2: when a novel object's feature is described, people infer that the feature is rarer of its category than when it goes unmentioned. Again, this effect was consistent across both size and color adjectives, and people did not substantially adjust this inference based on how necessary the description was to distinguish among potential referents. We also added an alien language condition, in which the entire referring expression was unintelligible to participants, to probe people's priors on feature typicality. We found that in the alien language condition, people judged features to be roughly between the adjective utterance and no adjective utterance conditions, and significantly different from the no adjective utterance condition. In the alien language condition, people's prevalence judgments were roughly around our model's prevalence judgments (50%) after observing the objects on each trial and before any inferences about the utterance.

The similarity of people's prevalence judgments in the alien language condition and the adjective condition raises the question: is this effect driven by an atypicality inference in the adjective conditions, or a *typicality* inference when the feature is unmentioned? Our results suggest that it is a bit of both. When someone mentions an object without extra description, the listener can infer that its features are likely more typical than their prior; when they use description, they can infer that its features are likely less typical. Because using an extra word–an adjective–is generally not thought of as the default way to refer to something, this effect is still best described as a contrastive inference of *atypicality* when people use description. However, the fact that people infer high typicality when an object is referred to without description suggests that, in some sense, there is no neutral way to refer: people will make broader inferences about a category from even simple mentions of an object.

Chapter 2

Children learn a tremendous amount about the structure of the world around them in just a few short years, from the rules that govern the movement of physical objects to the hierarchical structure of natural categories and even relational structures among social and cultural groups (Baillargeon, 1994; Legare & Harris, 2016; Rogers & McClelland, 2004). Where does the information driving this rapid acquisition come from? Undoubtedly, a sizeable component comes from direct experience observing and interacting with the world (Sloutsky & Fisher, 2004; Stahl & Feigenson, 2015). But another important source of information comes from the language people use to talk about the world (Landauer & Dumais, 1997; Rhodes, Leslie, & Tworek, 2012). How similar is the information available from children's direct experience to the information available in the language children hear?

Two lines of work suggest that they may be surprisingly similar. One compelling area of work is the comparison of semantic structures learned by congenitally blind children to those of their sighted peers. In several domains that would at first blush rely heavily on visual information, such as verbs of visual perception (e.g., *look*, *see*), blind children and adults make semantic similarity judgments that mirror their sighted peers (Bedny, Koster-Hale, Elli, Yazzolino, & Saxe, 2019; Landau, Gleitman, & Landau, 2009). A second line of evidence supporting the similarity of information in perception and language is the broad success of statistical models trained on language alone in approximating human judgments across a variety of domains (Landauer & Dumais, 1997; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Even more compellingly, models trained on both language usage and perceptual features for some words can infer the perceptual features of linguistically related words entirely from the covariation of language and perception (Johns & Jones, 2012).

Still, there is reason to believe that some semantic features may be harder to learn from language than these data suggest. This is because we rarely use language merely to provide running commentary on the world around us; instead, we use language to talk about things that diverge from our expectations or those of our conversational partner (Herbert P Grice, 1975). People tend to avoid being over- or under-informative when they speak. In particular, when referring to objects, people are informative with respect to both the referential context and the typical features of the referent (Rubio-Fernández, 2016b; Westerbeek et al., 2015b). People tend to refer to an object that is typical of its category with a bare noun (e.g., calling an orange carrot "a carrot"), but often specify when an object has an atypical feature (e.g, "a purple carrot"). Given these communicative pressures, naturalistic language statistics may provide surprisingly little evidence about what is typical (Willits, Sussman, & Amato, 2008).

If parents speak to children in this minimally informative way, children may be faced with input that emphasizes atypicality in relation to world knowledge they do not yet have. For things like carrots–which children learn about both from perception and from language–this issue may be resolved by integrating both sources of information. Likely almost all of the carrots children see are orange, and hearing an atypical exemplar noted as a "purple carrot" may make little difference in their inferences about the category of carrots more broadly. But for things to which they lack perceptual access–such as rare objects, unfamiliar social groups, or inaccessible features like the roundness of the Earth–much of the information must come from language (Harris & Koenig, 2006). If language predominantly notes atypical features rather than typical ones, children may overrepresent atypical features as they learn the way things in the world tend to be.

On the other hand, parents may speak to children far differently from the way they speak to other adults. Parents' speech may reflect typical features of the world more veridically, or even emphasize typical features in order to teach children about the world. Parents alter their speech to children along a number of structural dimensions, using simpler syntax and more reduplications (Snow, 1972). Their use of description may reflect similar alignment to children's growing knowledge.

We examine the typicality of adjectives in a large, diverse corpus of parent-child interactions recorded in children's homes to ask whether parents talking to their children tend to use adjectives predominantly to mark atypical features. We find that they do: Parents and children overwhelmingly choose to mention atypical rather than typical features. We also find that parents use adjectives differently over the course of children's development, noting typical features more often to younger children. We then ask whether the co-occurrence structure of language nonetheless captures typicality information by training vector space models on child-directed speech. We find that relatively little typical feature information is represented in these semantic spaces.

Adjective typicality

In order to determine whether parents use adjectives mostly to mark atypical features of categories, we analyzed caregiver speech from a large corpus of parent-child interactions. We extracted a subset of adjective-noun combinations that co-occurred, and asked a sample of Amazon Mechanical Turkers to judge how typical the property described by each adjective was for the noun it modified. We then examined both the broad features of this typicality distribution and the way it changes over development. Our theoretical hypotheses, statistical models, sample size, and exclusion criteria were pre-registered on the Open Science Framework (https://osf.io/ypdzv/).

Corpus

We used data from the Language Development Project, a large-scale, longitudinal corpus of parent-child interactions recorded in children's homes. Families were recruited to be representative of the Chicagoland area in both socio-economic and racial composition (Goldin-Meadow et al., 2014). Recordings were taken in the home every 4 months from when the child was 14 months old until they were 58 months old, resulting in 12 timepoints. Each recording was of a 90-minute session in which parents and children were free to behave and interact as they liked.

Our sample consisted of 64 typically-developing children and their caregivers with data from at least 4 timepoints ($mean = 11.3$ timepoints). Together, this resulted in a total of 641,402 distinct parent utterances.

Stimulus Selection

From these utterances, we extracted all of the nouns (using human-coded part of speech tags) resulting in a set of 8,150 total nouns. Because of our interest in change over development, we considered only nouns that appeared at least once every 3 sessions (i.e. per developmental year). This yielded a set of some 1,829 potential target nouns used over 198,014 distinct utterances.

We selected from the corpus all 35,761 distinct utterances containing any of these nouns and any word tagged as an adjective. We considered for analysis all adjective-noun pairs that occurred in any utterance (e.g., utterances with one noun and three adjectives were coded as three pairs) for a total of 18,050 distinct pairs. This set contained a number of high-frequency idiomatic pairs whose typicality was difficult to classify (e.g., "good"–"job"; "little"–"bit"). To resolve this issue, we used human judgments of words' concreteness to identify and exclude candidate idioms (Brysbaert, Warriner, & Kuperman, 2014). We retained for analysis only pairs in which both the adjective and noun were in the top 25% of the concreteness ratings (e.g., "dirty" – "dish"; "green" – "fish") restricting our set to 2,477. Finally, human coders in the lab judged whether each pair was "incoherent or unrelated" and we excluded a final 576 pairs from the sample (e.g., incoherent pairs such as "flat" – "honey").

Thus, our final sample included 1,901 unique adjective-noun pairs drawn from 3,749 distinct utterances. The pairs were combinations of 637 distinct concrete nouns and 111 distinct concrete adjectives. We compiled these pairs and collected human judgments on Amazon Mechanical Turk for each pair, as described below. Table 2 contains example utterances from the final set and typicality judgments from our human raters.

Participants

Each participant rated 20 adjective-noun pairs, and each pair was rated by four participants; we used Dallinger, a tool for automating complex recruitment on Amazon Mechanical Turk, to balance recruitment. Overall, we recruited 444 participants to rate our final sample of adjective–noun pairs. After exclusions using an attention check that asked participants to simply choose a specific number on the scale, we retained 8,580 judgments, with each adjective–noun pair retaining at least two judgments.

Design and Procedure

To evaluate the typicality of the adjective–noun pairs that appeared in parents' speech, we asked participants on Amazon Mechanical Turk to rate each pair. Participants were presented with a question of the form "How common is it for a cow to be a brown cow?" and asked to provide a rating on a seven-point scale: (1) never, (2) rarely, (3) sometimes, (4) about half the time, (5) often, (6) almost always, (7) always.

Results

The human typicality ratings were combined with usage data from our corpus analysis to let us determine the extent to which parents use language to describe typical and atypical features. In our analyses, we token-weighted these judgments, giving higher weight to pairs that occurred more frequently in children's inputs. However, results are qualitatively identical and all significant effects remain significant without these re-weightings.

If caregivers speak informatively to convey what is atypical or surprising in relation to their own sophisticated world knowledge, we should see that caregiver description is dominated by modifiers that are sometimes or rarely true of the noun they modify. If instead child-directed speech privileges redundant information, perhaps to align to young children's limited world knowledge, caregiver description should yield a distinct distribution dominated by highly typical modifiers. As predicted in our pre-registration, we find that parents' description predominantly focuses on features that are atypical (Figure **??**).

To confirm this effect statistically, we centered the ratings (i.e. "about half" was coded as 0), and then predicted the rating on each trial with a mixed effect model with only an intercept and a random effect of noun (`typicality ~ 1 + (1|noun)`). The intercept was reliably negative, indicating that adjectives tend to refer to atypical features of objects ($\beta =$ -0.77, $t =$ -19.72, $p < .001$). We then re-estimated these models separately for each age in the corpus, and found a reliably negative intercept for every age group (smallest effect $\beta_{14} =$ -0.50, $t =$ -4.45, $p = < .001$). These data suggest that even when talking with very young children, caregiver speech is structured according to adult communicative pressures observed in the lab.

For comparison, we performed the same analyses but with typicality judgments weighted not by the frequency of each adjective-noun pair's occurrence in the Language Development Project, but instead by their frequency of occurrence in the Corpus of

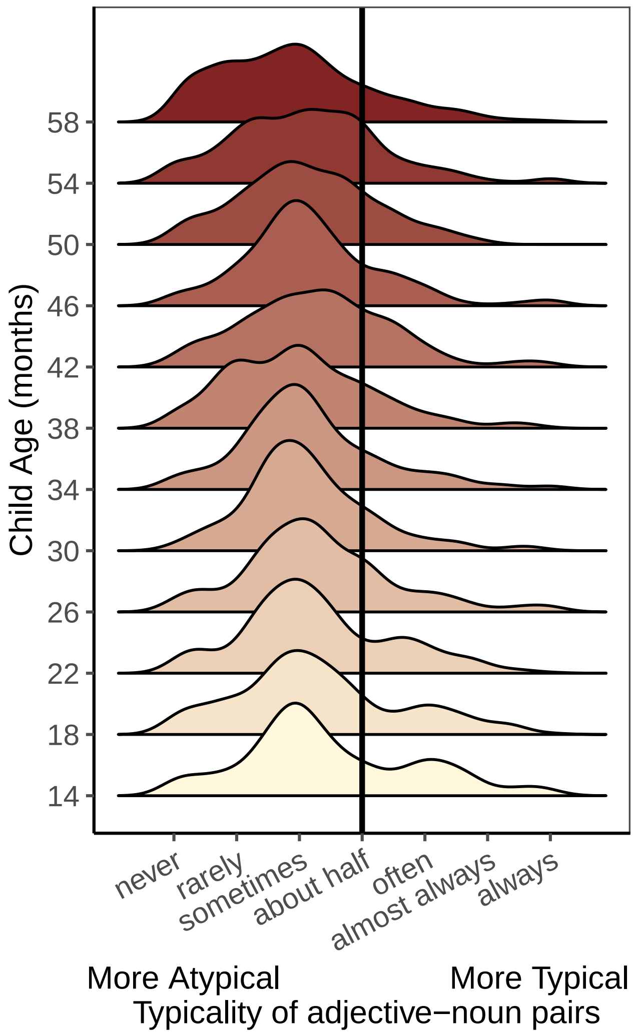

Figure 8. Density plots showing usage at each timepoint based on the typicality of the adjective-noun pair.

Contemporary American English (COCA; Davies, 2008). While this estimate of adult usage is imperfect—the adjective-nouns pairs produced by parents in our corpus may not be a representative sample of adjectives and nouns spoken by the adults in COCA—it provides a first approximation to adult usage. When we fit the same mixed-effects model to the data, we found that the intercept was reliably negative, indicating that adult-to-adult speech is likely also biased toward description of atypical features (β = -0.30, t = -19.72, p < .001).

Returning to caregiver speech, while descriptions at every age tended to point out

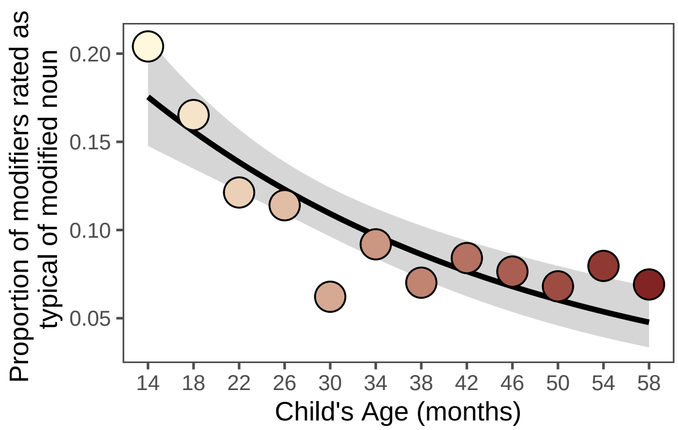

Figure 9. Proportion of caregiver description that is about typically-true features, as a function of age.

atypical features (as in adult-to-adult speech), this effect changed in strength over development. As predicted, an age effect added to the previous model was reliably negative, indicating that parents of older children are relatively more likely to focus on atypical features (β = -0.11, t = -3.47, p = .001). In line with the idea that caregivers adapt their speech to their children's knowledge, it seems that caregivers are more likely to provide description of typical features for their young children, compared with older children. As a second test of this idea, we defined adjectives as highly typical if Turkers judged them to be 'often', 'almost always', or 'always' true. We predicted whether each judgment was highly typical from a mixed-effects logistic regression with a fixed effect of age (log-scaled) and a random effect of noun. Age was a highly reliable predictor (β = -0.94, t = -5.01, p = $<$ .001). While children at all ages hear more talk about what is atypically true (Figure **??**), younger children hear relatively more talk about what is typically true than older children do (Figure **??**).

Child Speech. Given the striking consistency in adult-to-adult speech and caregiver speech across ages, we next briefly consider what kind of information is contained in children's speech. By analyzing children's own utterances, we can determine when children come to use description in a way that looks like adult speech. Are children mirroring adult-like uses of description even from a young age, or are they choosing to describe more typical features of the world?

The Language Development Corpus contains 368,348 child utterances. Using the set of adjective-noun pairs for which we have judgments from our analysis of caregiver speech, we repeat our analysis on usage data for a set of 460 distinct adjective-noun pairs which also appeared in children's productions. While preliminary, a mixed effects model predicting typicality had a highly-reliable negative intercept ($\beta =$ -0.72, $t =$ -10.14, $p = < .001$), but adding an age term did not improve model fit. Thus, children's speech is also biased towards atypical descriptions, and this bias does not change reliably over the first 5 years.

Discussion

In sum, we find robust evidence that language is used to discuss atypical, rather than typical, features of the world. Description in caregiver speech seems to largely mirror the usage patterns that we observed in adult-to-adult speech, suggesting that these patterns arise from general communicative pressures. Indeed, even children's own productions show a similar usage pattern, with more description of atypical features of the world even at the youngest ages.

It should be noted that children's utterances come from naturalistic conversations with caregivers, and their use of atypical description may be prompted by parent-led discourse. That is, if a caregiver chooses to describe the *purpleness* of a cat in book, the child may well respond by asking about that same feature. Further, atypical descriptors may actually be more likely to elicit imitation from child speakers, compared with typical descriptors (Bannard, Rosner, & Matthews, 2017). Future analyses would need to better disentangle the extent to which children's productions are imitative of caregivers.

Interestingly, the descriptions children hear change over development, becoming increasingly focused on atypical features. The higher prevalence of typical descriptors in early development may help young learners learn what is typical; however, even at the earliest point we measured, the bulk of language input describes atypical features.

This usage pattern aligns with the idea that language is used informatively in relation to background knowledge about the world. It may pose a problem, however, for young language learners with still-developing world knowledge. If language does not transparently convey the typical features of objects, and instead (perhaps misleadingly) notes the atypical ones, how might children come to learn what objects are typically like? One possibility is that information about typical features is captured in regularities across many utterances. If this is true, language may still be an important source of information about typicality as children may be able to extract more accurate typicality information by tracking second-order co-occurrence.

Extracting Typicality from Language Structure

Much information can be gleaned from language that does not seem available at first glance. From language alone, simple distributional learning models can recover enough information to perform comparably to non-native college applicants on the Test of English as a Foreign Language (Landauer & Dumais, 1997). Recently, Lewis, Zettersten, and Lupyan (2019) demonstrated that even nuanced feature information may be learnable through distributional semantics alone, without any complex inferential machinery. We take a similar approach to ask whether a distributional semantics model trained on the language children hear can capture typical feature information.

Method

To test this possibility, we trained word2vec–a distributional semantics model–on the same corpus of child-directed speech used in our first set of analyses. Word2vec is a neural network model that learns to predict words from the contexts in which they appear. This leads word2vec to learn representations in which words that appear in similar contexts become similar to each-other (Firth, 1957).

We used the continuous-bag-of-words (CBOW) implementation of word2vec in the `gensim` package (Řehůřek & Sojka, 2010). We trained the model using a surrounding context of 5 words on either side of the target word and 100 dimensions (weights in the hidden layer) to represent each word. After training, we extracted the hidden layer representation of each word in the model's vocabulary–these are the vectors used to represent these words.

If the model captures information about the typical features of objects, we should see that the model's noun-adjective word pair similarities are correlated with the typicality ratings we elicited from human raters. For a second comparison, we also used an off-the-shelf implementation of word2vec trained on Wikipedia (Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2018). While the Language Development Project corpus likely underestimates the amount of structure in children's linguistic input, Wikipedia likely overestimates it.

Results

We find that similarities in the model trained on the Language Development Project corpus have near zero correlation with human adjective–noun typicality ratings ($r = 0.03$, $p = .208$). However, our model does capture other meaningful information about the structure of language, such as similarity. Comparing with pre-existing large-scale human similarity judgements for word pairs, our model shows significant correlations (correlation with wordsim353 similarities of noun pairs, 0.28; correlation with simlex similarities of noun, adjective, and verb pairs, 0.16). This suggests that statistical patterns in child-directed speech are likely insufficient to encode information about the typical features of objects, despite encoding at least some information about word meaning more broadly.

However, the corpus on which we trained this model was small; perhaps our model did not get enough language to draw out the patterns that would reflect the typical features of objects. To test this possibility, we asked whether word vectors trained on a much larger corpus—English Wikipedia—strongly correlate with typicality ratings. This model's similarities were significantly correlated with human judgments, although the strength of the correlation was still fairly weak ($r = 0.25$, $p < .001$). Interestingly, similarities from the two models correlated more highly to each other than either model correlated with human judgments ($r = 0.29$, $p < .001$). This suggests that these models are picking up on some systematic associations between nouns and adjectives, but not the typical features of things.

One possible confound in these analyses is that the similarity judgments produced by our models reflect many dimensions of similarity, but our human judgments reflect only typicality. To accommodate this, we performed a second analysis in which we considered only the subset of 73 nouns that had both a typical (rated as at least "often") and an atypical (rated as at most "sometimes") adjective. We then asked whether the models rated the typical adjective as more similar to the noun it modified than the atypical adjective. The LDP model correctly classified 38 out of 73 (0.52), which was not better than chance ($p =$ .815). The Wikipedia model correctly classified 56 out of 73 (0.77), which was better than chance according to a binomial test, but still fairly poor performance ($p = < .001$). Fig 10 shows the ratings from Turkers and the two models for the 73 nouns. Table 1 gives the six cases in which word2vec similarities are worst at predicting human typicality judgments, judging the low-typicality adjective to be *more* similar to the noun than the high-typicality adjective.

General Discussion

Language provides children a rich source of information about the world. However, this information is not always transparently available: because language is used to comment on the atypical, it does not perfectly mirror the world. Among adult conversational partners whose world knowledge is well-aligned, this allows people to converse informatively and avoid redundancy. But between a child and caregiver whose world knowledge is asymmetric, this pressure competes with other demands: what is minimally informative to an adult may be

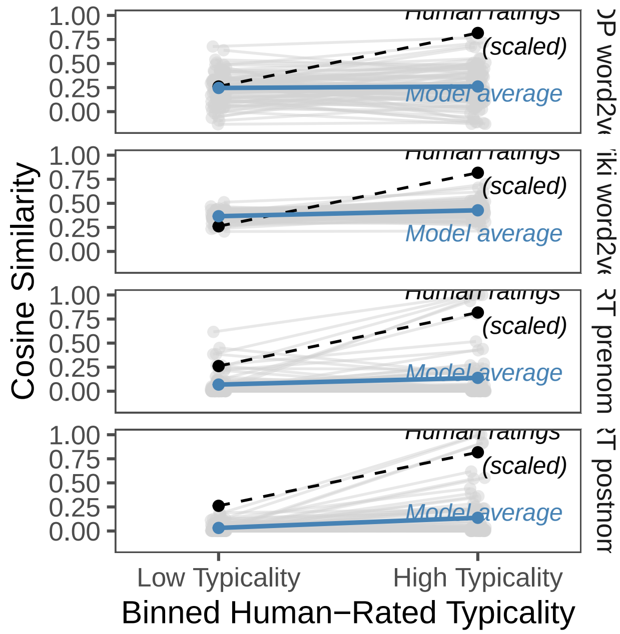

Figure 10. Plots of word2vec noun-adjective similarities for nouns for which there was at least one atypical adjective (rated at most "sometimes"), and at least one typical adjective (rated at least "often").

misleading to a child. Our results show that this pressure structures language to create a peculiar learning environment, one in which caregivers predominantly point out the atypical features of things.

How, then, do children learn about the typical features of things? While younger children may gain an important foothold from hearing more description of typical features, they still face language dominated by atypical description. When we looked at more nuanced ways of extracting information from language (which may or may not be available to the developing learner), we found that models of distributional semantics capture little typical feature information.

Of course, perceptual information from the world may simplify this problem. In many cases, perceptual information may swamp information from language; children likely see enough orange carrots in the world to outweigh hearing “purple carrot.” It remains unclear,

noun	typical adjective	atypical adjective
puzzle	flat	giant
apple	red	brown
bird	outside	purple
elephant	fat	pink
whale	wet	red
frog	green	purple

Table 1

The top six cases in which Wikipedia-trained word2vec similarities were worst at predicting human typicality judgments. In each case, word2vec judged the low-typicality adjective to be more similar to the noun than the high-typicality adjective.

however, how children learn about categories for which they have scarcer evidence. Indeed, language information likely swamps perceptual information for many other categories, such as abstract concepts or those that cannot be learned about by direct experience. If such concepts pattern similarly to the concrete objects analyzed here, children are in a particularly difficult bind.

It is also possible that other cues from language and interaction provide young learners with clues to what is typical or atypical, and these cues are uncaptured by our measure of usage statistics. Caregivers may highlight when a feature is typical by using certain syntactic constructions, such as generics (e.g., “tomatoes are red”). Caregivers may also mark the atypicality of a feature, for example demonstrating surprise. Such cues from language and the interaction may provide key information in some cases; however, given the sheer frequency of atypical descriptors, it seems unlikely that they are consistently well-marked.

Another possibility is that children expect language to be used informatively at a young age. Under this hypothesis, their language environment is not misleading at all, even without additional cues from caregivers. Children as young as two years old tend to use words to comment on what is new rather than what is known or assumed (Baker & Greenfield, 1988). Children may therefore expect adjectives to comment on surprising features of objects. If young children expect adjectives to mark atypical features (Alexandra C Horowitz & Frank, 2016), they can use description and the lack thereof to learn more about the world. Indeed, this idea is consistent with our finding that even young children largely choose to describe atypical features. Though this effect can be explained by simpler means such as mimicry, it suggests that caregivers and children may be usefully aligned in the aspects of the world they choose to talk about.

Whether adult-directed, child-directed, or a child's own speech, language is used with remarkable consistency: people talk about the atypical. Though parents might reasonably be broadly over-informative in order to teach their children about the world, this is not the case. This presents a potential puzzle for young learners who have limited world knowledge and limited pragmatic inferential abilities. Perceptual information and nascent pragmatic abilities may help fill in the gaps, but much remains to be explored to link these explanations to actual learning. Communication pressures are pervasive forces structuring the language children hear, and future work must disentangle whether children capitalize on them or are misled by them in learning about the world.

Stimuli, data, and analysis code available at https://osf.io/ypdzv/

Acknowledgements

This research was funded by a James S. McDonnell Foundation Scholar Award to DY.

References

Acknowledgements

This research was funded by James S. McDonnell Foundation Scholar Award in Understanding Human Cognition #220020506 to Dan Yurovsky. The funding body had no involvement in the conceptualization, data collection, or analysis of this project.

Each chapter in this proposal represents collaborative work. The collaborators on each chapter are as follows: Chapter 1, Dan Yurovsky; Chapter 2, Benjamin Morris and Dan Yurovsky; Chapter 3, Rachel King and Dan Yurovsky.

Thank you to Ming Xiang, Benjamin Morris, Ashley Leung, Michael C. Frank, Judith Degen, Stephan Meylan, and Ruthe Foushee for feedback on portions of this manuscript. Portions of this work were published in the proceedings of Experiments in Linguistic Meaning (2020) and the proceedings of the 42nd annual meeting of the Cognitive Science Society. The authors are grateful for feedback from reviewers and attendees of Experiments in Linguistic Meaning, the meeting of the Cognitive Science Society, the meeting of the Society for Research in Child Development, the Midwestern Cognitive Science Conference, and the Dubrovnik Conference on Cognitive Science.

References

Akhtar, N., Carpenter, M., & Tomasello, M. (1996). The Role of Discourse Novelty in Early Word Learning. *Child Development*, *67*(2), 635–645. https://doi.org/10.1111/j.1467-8624.1996.tb01756.x

Aparicio, H., Xiang, M., & Kennedy, C. (2016). Processing gradable adjectives in context: A visual world study. In *Semantics and linguistic theory* (Vol. 25, pp. 413–432).

Arts, A., Maes, A., Noordman, L. G. M., & Jansen, C. (2011). Overspecification in written instruction. *Linguistics*, *49*(3), 555–574.

Baillargeon, R. (1994). How do infants learn about the physical world? *Current Directions in Psychological Science*, *3*(5), 133–140.

Baker, N. D., & Greenfield, P. M. (1988). The development of new and old information in young children's early language. *Language Sciences*, *10*(1), 3–34.

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

Bedny, M., Koster-Hale, J., Elli, G., Yazzolino, L., & Saxe, R. (2019). There's more to "sparkle" than meets the eye: Knowledge of vision and light verbs among congenitally blind and sighted individuals. *Cognition*, *189*, 105–115.

Bergey, C., Morris, B., & Yurovsky, D. (2020). *Children hear more about what is atypical than what is typical*. PsyArXiv. https://doi.org/10.31234/osf.io/5wvu8

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Clark, E. V. (1990). On the pragmatics of contrast. *Journal of Child Language*, *17*(2), 417–431. https://doi.org/10.1017/S0305000900013842

Dale, R., Kehoe, C., & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory & Cognition*, *35*(1), 15–28.

Davies, M. (2008). The corpus of contemporary american english (coca): 520 million words, 1990-present.

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review*, *127*, 591–621.

Engelhardt, P. E., Barış Demiral, Ş., & Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, *77*(2), 304–314. https://doi.org/10.1016/j.bandc.2011.07.004

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*.

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

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

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585.

Goldin-Meadow, S., Levine, S. C., Hedges, L. V., Huttenlocher, J., Raudenbush, S. W., & Small, S. L. (2014). New evidence about language and cognitive development based on a longitudinal study: Hypotheses for intervention. *American Psychologist*, *69*(6), 588.

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Harris, P. L., & Koenig, M. A. (2006). Trust in testimony: How children learn about science and religion. *Child Development*, *77*(3), 505–524.

Horowitz, A. C., & Frank, M. C. (2016). Children's Pragmatic Inferences as a Route for Learning About the World. *Child Development*, *87*(3), 807–819.

Horowitz, A. C., & Frank, M. C. (2016). Children's pragmatic inferences as a route for learning about the world. *Child Development*, *87*(3), 807–819.

Johns, B. T., & Jones, M. N. (2012). Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, *4*(1), 103–120.

Kanwisher, N., Woods, R. P., Iacoboni, M., & Mazziotta, J. C. (1997). A locus in human extrastriate cortex for visual shape analysis. *Journal of Cognitive Neuroscience*, *9*(1), 133–142.

Kreiss, E., & Degen, J. (2020). Production expectations modulate contrastive inference. In.

Landau, B., Gleitman, L. R., & Landau, B. (2009). *Language and experience: Evidence from the blind child* (Vol. 8). Harvard University Press.

Landau, B., Smith, L. B., & Jones, S. (1992). Syntactic context and the shape bias in children's and adults' lexical learning. *Journal of Memory and Language*, *31*(6), 807–825.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.

Psychological Review, *104*(2), 211.

Legare, C. H., & Harris, P. L. (2016). The ontogeny of cultural learning. *Child Development*, *87*(3), 633–642.

Lewis, M., Zettersten, M., & Lupyan, G. (2019). Distributional semantics as a source of visual knowledge. *Proceedings of the National Academy of Sciences*, *116*(39), 19237–19238.

Mangold, R., & Pobel, R. (1988). Informativeness and Instrumentality in Referential Communication. *Journal of Language and Social Psychology*, *7*(3-4), 181–191.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the international conference on language resources and evaluation (lrec 2018).*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mitchell, M., Reiter, E., & Deemter, K. van. (2013). Typicality and Object Reference, 7.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of Perspective-Taking Constraints in Children's On-Line Reference Resolution. *Psychological Science*, *13*(4), 329–336.

Ni, W. (1996). Sidestepping garden paths: Assessing the contributions of syntax, semantics and plausibility in resolving ambiguities. *Language and Cognitive Processes*, *11*(3), 283–334.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*(1), 89–110.

Rhodes, M., Leslie, S.-J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, *109*(34), 13526–13531.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

Rohde, H., & Rubio-Fernandez, P. (2021). Color interpretation is guided by informativity expectations, not by world knowledge about colors.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 491.

Rubio-Fernández, P. (2016a). How Redundant Are Redundant Color Adjectives? An Efficiency-Based Analysis of Color Overspecification. *Frontiers in Psychology*, *7*.

Rubio-Fernández, P. (2016b). How Redundant Are Redundant Color Adjectives? An Efficiency-Based Analysis of Color Overspecification. *Frontiers in Psychology*, *7*.

Ryskin, R., Kurumada, C., & Brown-Schmidt, S. (2019). Information integration in modulation of pragmatic inferences during online language comprehension. *Cognitive Science*, *43*(8), e12769.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.

Sedivy, J. C. (2003). Pragmatic Versus Form-Based Accounts of Referential Contrast: Evidence for Effects of Informativity Expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147.

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General*, *133*(2), 166.

Snow, C. E. (1972). Mothers' speech to children learning language. *Child Development*, 549–565.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Citeseer.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.

Westerbeek, H., Koolen, R., & Maes, A. (2015a). Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00935

Westerbeek, H., Koolen, R., & Maes, A. (2015b). Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in Psychology*, *6*.

Willits, J. A., Sussman, R. S., & Amato, M. S. (2008). Event knowledge vs. Verb knowledge. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 2227–2232).

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review*, *114*(2), 245.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.

utterance	pair	rating 1	rating 2	rating 3	rating 4	mean typicality
especially with wooden shoes.	wooden-shoe	2	2	3	2	2.25
you like red onions?	red-onion	3	5	3	4	3.75
the garbage is dirty.	dirty-garbage	7	7	5	7	6.50

Table 2

Sample typicality ratings from 4 human coders for three adjective-noun pairs drawn from the corpus.