Using contrastive inferences to learn about new words and categories

Claire Augusta Bergey[1] & Daniel Yurovsky[2]

[1] The University of Chicago

[2] Carnegie Mellon University

Author Note

All data and code for analyses are available at https://github.com/cbergey/contrast.

Correspondence concerning this article should be addressed to Claire Augusta Bergey, 5848 S. University Avenue, Chicago, IL 60637. E-mail: cbergey@uchicago.edu

Abstract

In the face of unfamiliar language or objects, description is one cue people can use to learn about both. Beyond narrowing potential referents to those that match a descriptor (e.g., "tall"), people could infer that a described object is one that contrasts with other relevant objects of the same type (e.g., "the tall cup" contrasts with another, shorter cup). This contrast may be in relation to other objects present in the environment (this cup is tall among present cups) or to the referent's category (this cup is tall for a cup in general). In three experiments, we investigate whether people use such contrastive inferences from description to learn new word-referent mappings and learn about new categories' feature distributions. People use contrastive inferences to guide their referent choice, though size—and not color—adjectives prompt them to consistently choose the contrastive target over alternatives (Experiment 1). People also use color and size description to infer that a novel object is atypical of its category (Experiments 2 and 3). However, these two inferences do not trade off substantially: people infer a described referent is atypical even when the descriptor was necessary to establish reference. We model these experiments in the Rational Speech Act (RSA) framework and find that it predicts both of these inferences, and a very small trade-off between them—consistent with the non-significant trade-off we observe in people's inferences. Overall, people are able to use contrastive inferences from description to resolve reference and make inferences about a novel object's category, allowing them to learn more about new things than literal meaning alone allows.

Keywords: concept learning; contrastive inference; word learning; pragmatics; communication; computational modeling

Word count: 11021

Using contrastive inferences to learn about new words and categories

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. 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; 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 language. People can learn the meanings of words by tracking statistical properties of their literal meaning alone (Yu & Smith, 2007), but reasoning about a speaker's intended meaning–and not just the words they say–may support more rapid and 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, 2016; Westerbeek, Koolen, & Maes, 2015). 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 you have 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. On this *probabilistic weighing view*, people can use description to make graded inferences about the referent's identity and about its category's features, and the fact that an adjective would have helped identify the referent does not completely block an inference about atypicality.

In this paper, we present a series of experiments that test two ways in which people could use pragmatic inference to learn about novel categories. First, we examine whether listeners use contrastive inference to resolve referential ambiguity. In a reference game, participants saw groups of novel objects and were asked to pick one with a referring expression, e.g., “Find the small toma.” If people interpret description contrastively, they should infer that the description was necessary to identify the referent–that the small toma contrasts with some different-sized toma on the screen. We show that people can use contrastive inference–even with unfamiliar objects–to resolve reference and thus to learn the meaning of the new word “toma.”

Second, we test whether people use contrastive inference to learn about a novel category’s feature distribution. Participants were presented with two interlocutors who exchange objects using referring expressions, such as “Pass me the blue toma.” If people interpret description as contrasting with an object’s category, they should infer that in general, few tomas are blue. Crucially, we vary the object contexts such that in some contexts, the adjective is necessary to establish reference, and in others, it is superfluous. Overall, we show that people can use contrastive inferences both to establish reference and to make inferences about novel categories’ feature distributions, and that they do not trade off strongly between these two inferences. We extend a version of the Rational Speech Act model (Frank & Goodman, 2014) that captures how listeners’ reasoning about speakers reflects a graded integration of informativity with respect to both reference and typicality.

In order to determine whether people can use contrastive inferences to disambiguate referents and learn about categories’ feature distributions, we use reference games with novel objects. Novel objects provide both a useful experimental tool and an especially interesting testing ground for contrastive inferences. These objects have unknown names and feature distributions, creating the ambiguity that is necessary to test referential disambiguation and category learning. Testing pragmatic inference in novel, ambiguous situations lays the groundwork to determine the role of pragmatic inference in learning language. Much work has focused on how pragmatic inference enriches literal meaning when the literal meaning is known–when the words and referents in play are familiar. Here, we ask: can people use pragmatic inferences from description to learn about unfamiliar things in the world?

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 reference resolution 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. (In fact, using an adjective to refer to the lure could even introduce ambiguity, as the adjective applies to both the target and lure and delays the onset of the noun, which would have unambiguously identified the lure.) 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. Each participant gave informed consent and was paid $0.30 in exchange for their participation. Participants were told the task was estimated to take 3 minutes and on average they took 44 seconds to complete the trials (not including reading the consent form).

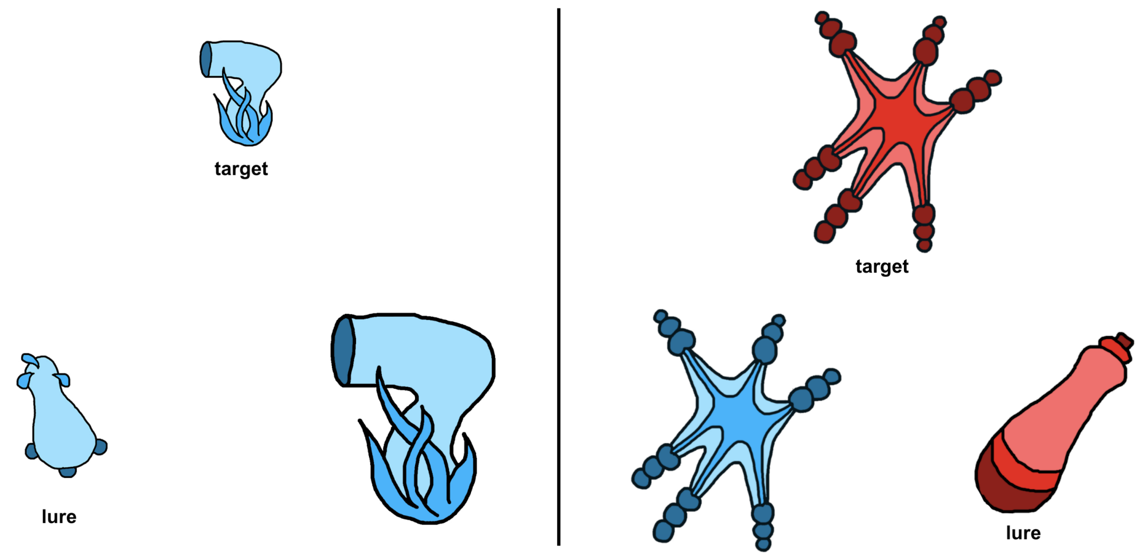

Figure 1. On the left: an example of a contrastive display trial in which the critical feature is size. Here, the participant would hear the instruction “Find the toma” or “Find the small toma.” The target is the small hairdryer-shaped object. On the right: an example of a contrastive display trial in which the critical feature is color. Here, the participant would hear the instruction “Find the toma” or “Find the red toma.” The target is the red star-shaped object. In each case, the lure shares the target’s critical feature (small on the left, red on the right) but not its shape. The contrastive pair shares the target’s shape but not its critical feature. Labels of the target and lure are provided for clarity and were not shown to participants.

Stimuli. Stimulus displays were arrays of three novel fruit objects. We chose alien fruits as stimuli because fruits are a superordinate category that can vary considerably in shape, color, and size. Fruits were selected randomly at each trial from 20 fruit kinds. Ten of the 20 fruit drawings were adapted and redrawn from Kanwisher, Woods, Iacoboni, and Mazziotta (1997); we designed the remaining 10 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. The on-screen positions of the target and distractor items were randomized within a triad configuration.

There were two display types: contrastive displays and unique target displays. 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). Contrastive displays are the display type of interest, as the presence of a contrastive pair allows for a contrastive inference.

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 filler trials, to space out contrastive displays to prevent participants from dialing in on the contrastive object setup during the experiment. Further details about these trials, and the analysis of participants' choices in them, can be found in the Supplemental Materials. All discussion of the results in the main text include only the contrastive displays.

In summary, we manipulated three factors: utterance type (adjective or no adjective), critical feature type (color or size), and display type (contrastive display or unique target display). Utterance type and display type were manipulated within subjects, as utterance type is the central manipulation of interest and variation in display type was included as filler to prevent participants from cluing into the intended inference. Critical feature (color or size) was manipulated between subjects because generating enough unique stimuli to cross every factor within participants (such that participants never saw the same object twice) was time prohibitive.

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 contrastive displays and half were unique target displays (filler trials). Crossed with display type, half of trials had audio instructions with an adjective that described the critical feature of the target (e.g., "Find the blue toma" or "Find the big toma"), and half of trials

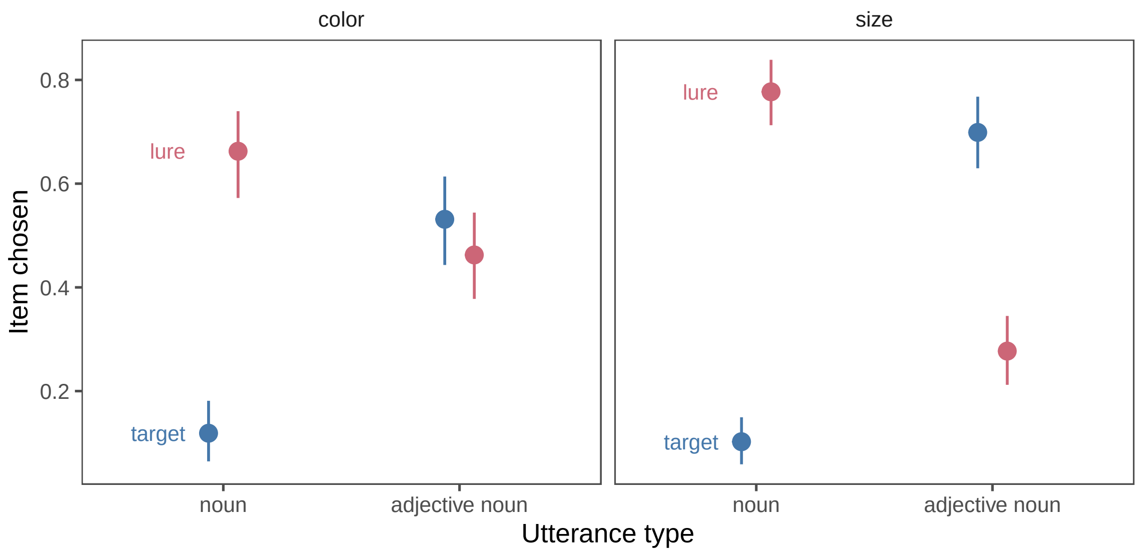

Figure 2. Proportion of times that participants chose the target and lure items as a function of condition and whether an adjective was provided. Points indicate group means; error bars indicate 95% confidence intervals computed by non-parametric bootstrapping.

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 novel names: blicket, wug, toma, gade, sprock, koba, zorp, and lomet.

After completing the study, as a check of their attention to the task, 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].

[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.

Results

Our key pre-registered analysis was whether participants would choose the target object on contrastive display trials when they heard an adjective in the referring expression. That is, when they saw the stimuli on the left side of 1 and heard "Find the small toma," would they choose the target (small hairdryer) over the lure (small pear)? 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: when they see a display like those in 1 and hear "Find the toma," they tend to choose the lure. However, when they hear an utterance with description (e.g., "Find the blue toma", "Find the small toma"), they shift away from choosing the unique lure and toward choosing the target, which has 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

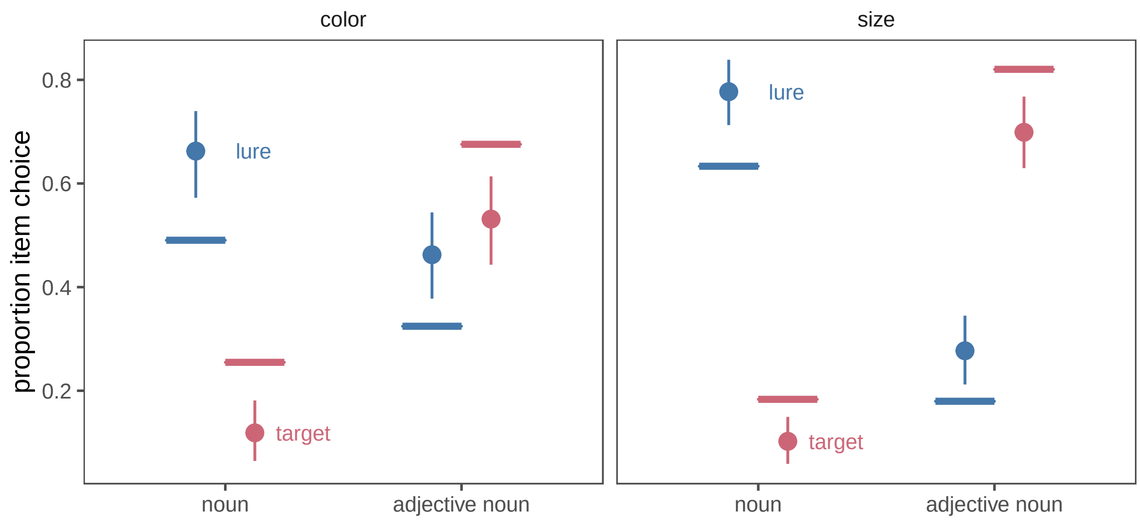

Figure 3. Proportion of times that people (and our model) chose the target and lure items as a function of adjective type and whether an adjective was provided. Points indicate empirical means; error bars indicate 95% confidence intervals computed by non-parametric bootstrapping. Solid horizontal lines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

referent is the unique object (lure), whereas hearing a modified noun (e.g. “small toma”) 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). We estimated the rationality parameter separately for color and size, reflecting that listeners may believe speakers are more attentive to differences in utility for some feature descriptions than others. (Note that the rationality parameter is sometimes used to explain *individual differences* in speaker rationality, and estimated on a person level; that is not how we are using it here.) 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 do people make stronger pragmatic inferences about size than color when determining reference? Our model implements this difference in a relatively agnostic way, and our results cannot arbitrate between particular explanations, but we spell out a few possibilities and modeling alternatives here.

One way to capture this asymmetry would be to locate it in a different part of the model: in the literal semantics of color and size. 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, implementing semantic inexactness alone in our 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.

Another possibility is that people attend to the production probabilities of different adjective types and attenuate their inferences accordingly. As discussed, speakers mention color more often than size, and listeners may keep track of these probabilities and discount the weight of color description in identifying referents. Experiments with familiar objects show that people make stronger contrastive inferences with respect to size than color, and Kreiss and Degen (2020) demonstrate that it is possible to explain differential inferences among color adjectives using production norms. Where do these production differences come from?

[XXXXXX talk about production norms predicted by inexactness]

Another difference between size and color adjectives is that size adjectives are relative gradable adjectives: their meaning is judged relative to a comparison class (e.g., "He is a tall basketball player" may have a meaning akin to "He is tall for a basketball player") (Kennedy, 2007). Because this comparison class is sensitive to context (it can even change within a sentence, e.g., "He is tall, but not tall for a basketball player"), there is active disagreement about whether this aspect of gradable adjective meaning is properly considered semantics or pragmatics (Xiang, Kennedy, Xu, & Leffel, 2022). Thus, a possible explanation is that the presence of a comparison class is necessary to judge size but not color, and this accounts for the asymmetry. That is, in a trial such as the one on the left in 1, a participant sees two hairdryer-shaped objects (one big and one small) and one small pear-shaped object. When they hear "Find the small toma," they choose the only object that is small and has a potential comparison class: the small hairdryer, which has a larger hairdryer counterpart. On the other hand, color adjectives are not relative gradable adjectives, and so a comparison class is not necessary to interpret them: they have more absolute meaning. Thus, it is possible to explain the color–size asymmetry by the necessity of a comparison class for judging size, and this may be attributed either to semantics or pragmatics.

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, 2016; Westerbeek et al., 2015). 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. 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). This work provided a useful demonstration that adjective use can contribute to inferences about feature typicality, though it did not isolate the effect of adjectives specifically. Their experiments used several contrastive cues, such as prosody (contrastive stress on the adjective: "TALL zib"), demonstrative phrases that may have marked the object as unique ("this one") and expressions of surprise at the object ("wow"), and participants may have inferred the object was atypical primarily from these cues and not from the adjective. Thus, in this experiment, we first set out to ask whether adjective use alone prompts an inference of atypicality: when you hear "purple toma," do you infer that *fewer* tomas in general are purple?

We will also test how this inference differs (or does not) between size and color adjectives. The fact that people use adjectives to draw a contrast between an object and its category 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 (2016) 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; Kreiss & Degen, 2020). 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 a task with novel objects 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.

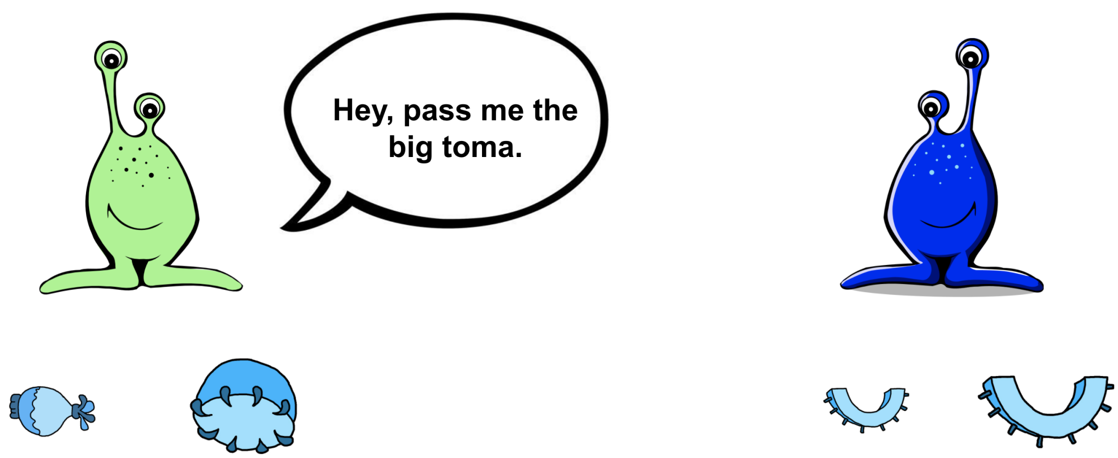

Figure 4. Experiment 2 stimuli. In the above example, the critical feature is size and the object context is a within-category contrast: the alien on the right has two same-shaped objects that differ in size.

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). Participants were paid \$0.30. Participants were told the task was estimated to take 3 minutes and on average took 118 seconds to complete the trials (not including reading the consent form).

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 referent disappeared from Alien B's side and reappeared on Alien A's side. Note that the participants do not make a referent choice in this experiment; the measure of interest is their typicality judgments of the objects' features, described below.

We manipulated three factors: utterance type, critical feature type, and context type. As in Experiment 1, we prioritized utterance type as a within-subjects manipulation because it was the central manipulation of interest. We also prioritized context type because another central question was whether context would alter the effect of utterance. We manipulated the critical feature type (color or size) between subjects.

Utterance type and context type were fully crossed within subjects. Utterance type had two levels: adjective (e.g., "Hey, pass me the big toma" or "Hey, pass me the blue toma") or no adjective (e.g., "Hey, pass me the toma"). Context type 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. This prevalence judgment, on a 0-100% scale, is our measure of interest. 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, as a check of whether they attended to the task. 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.

Results

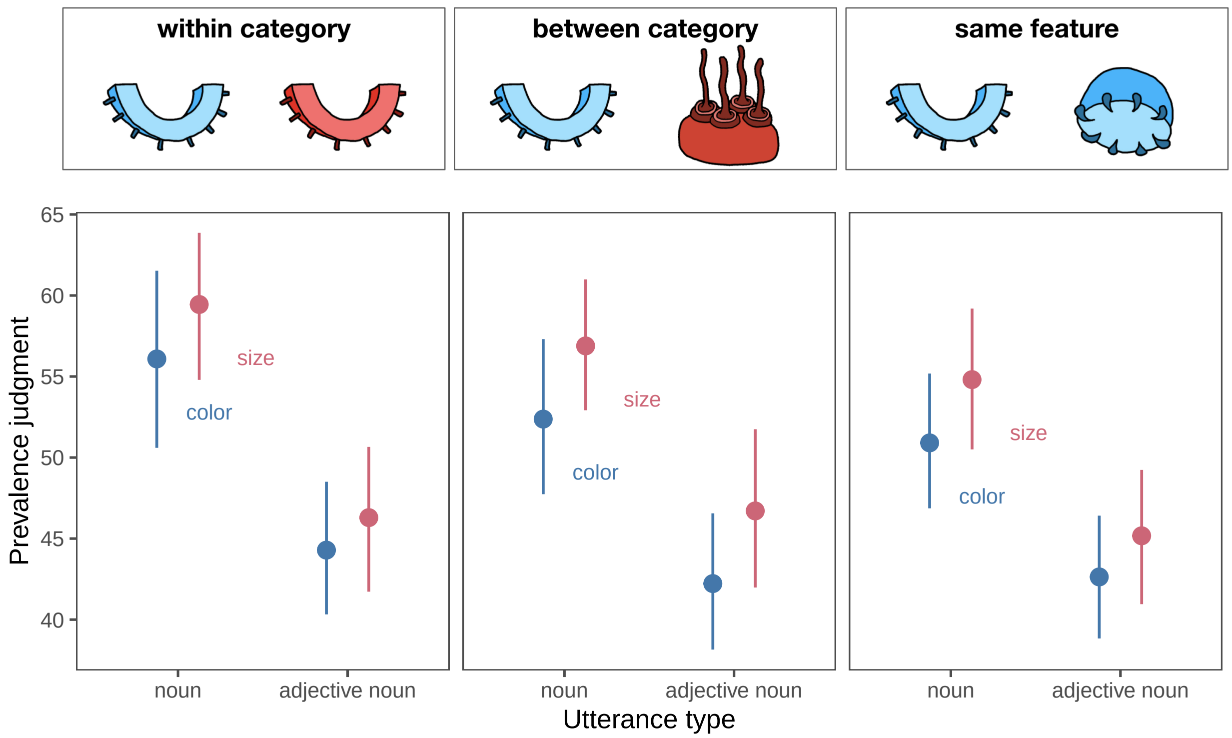

Figure 5. Prevalence judgments from Experiment 2. Participants consistently judged the target object as less typical of its category when the referent was described with an adjective (e.g., "Pass me the blue toma") than when it was not (e.g., "Pass me the toma"). This inference was not significantly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

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 separately tested whether there are effects of context among just the trials with adjectives and just the trials without adjectives. In each case, 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 noun utterances (β_{within} = 3.94, t = 1.47, p = .143; β_{same} = -1.46, t = -0.54, p = .587). That is, we did not find evidence here that people were inferring a feature to be highly typical because it went unmentioned when it was necessary for reference. Participants also did not significantly adjust their inferences by context among only the adjective noun utterances (β_{within} = 2.43, t = 1.16, p = .247; β_{same} = 0.67, t = 0.32, p = .750). That is, we did not find evidence that people modulated their typicality inferences based on the referential context among trials where this inference could not have been driven by omission either. Overall, we did not find evidence that participants significantly adjusted their inferences based on 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, 2016). 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.

Another interpretation of people's inferences in the size condition is that they are due to size adjectives being relative gradable adjectives. That is, the phrases "big toma" and "small toma" may inherently carry the meaning "big for a toma" and "small for a toma" (which can be interpreted as an aspect of the adjective's semantics, not pragmatics; see discussion in Experiment 1). It is possible to attribute people's atypicality inferences in the size condition to the relative gradable nature of size adjectives. However, people also made these inferences about color adjectives, which are not relative gradable adjectives. This kind of explanation also might predict that people's inferences about color and size would be different—for instance, that people would make larger atypicality inferences about size than color—which we do not find. Though we find it parsimonious here to explain the color and size inferences by the same mechanism, we acknowledge the possibility that the semantics of size adjectives contribute to people's inferences of atypicality.

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, $P(r|Cat)$. 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 prevalence 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 or one big dax and one small 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 $\text{Beta}\,(2, 2)$. Thus, we model learners as believing the feature prevalence is roughly 50% based on their initial priors and direct observation in the trial; they then combine this knowledge of the feature distribution with their pragmatic inference about the utterance to arrive at a final prevalence judgment.

As in Experiment 1, we encoded potential differences between people's inferences about color and size in feature rationality parameters, which we estimated separately for Experiment 2. 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 feature 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, inferring that the critical feature was less prevalent in the category when it was mentioned

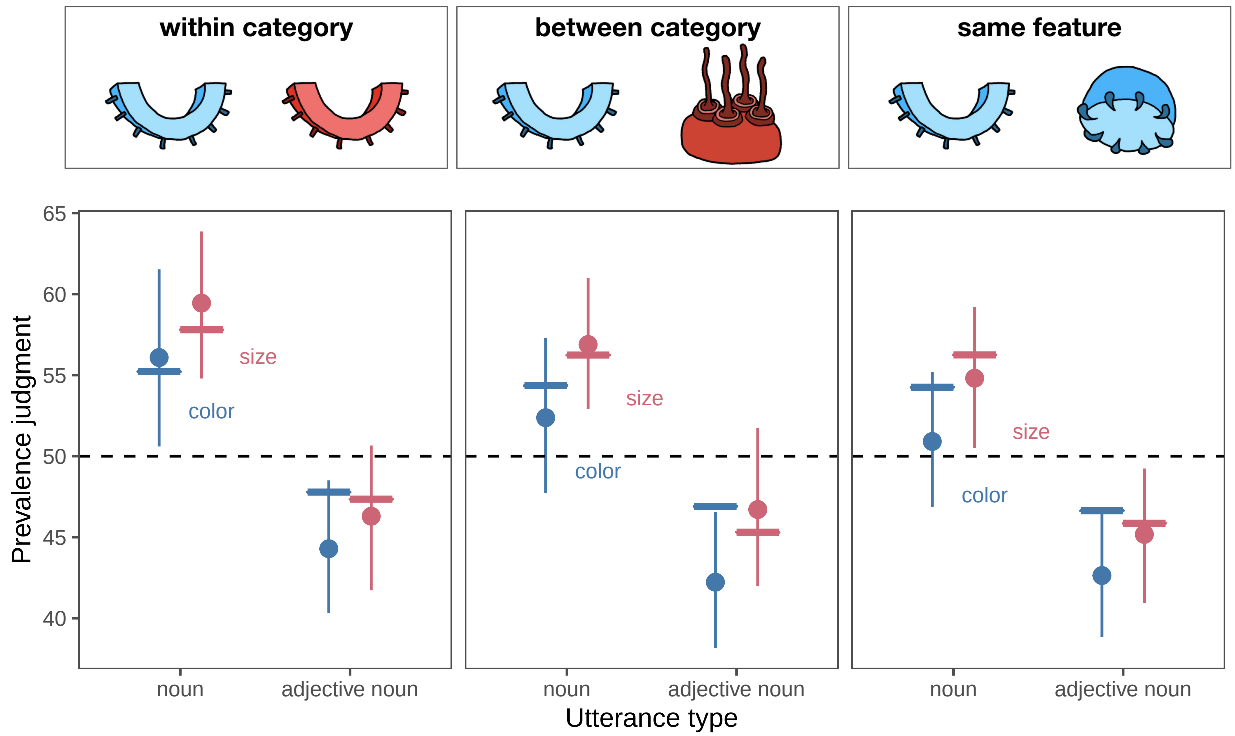

Figure 6. Participants' prevalence judgments from Experiment 2, compared to model predictions (horizontal lines).

(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). Participants were paid $0.30. Participants were told the task was estimated to take 3 minutes and on average they took 135 seconds to complete the trials (not including reading the consent form).

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, as a check of whether they attended to the task. 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 trials without an adjective and only trials with an adjective, separately. Testing only trials without an adjective checks directly whether people make higher typicality judgments when an adjective is necessary but not used, compared to when it is not necessary and not used. To check this, we fit a model on only trials with an adjective, with effects of context and feature type and their interaction, as well as random slopes by subject (not pre-registered). Participants' inferences among only utterances without an adjective did not significantly differ by context ($\beta_{within} = 0.09$, $t_{within} = 0.05$, $p_{within} = .964$). In the same way, we tested whether people's inferences varied by context among only trials with an adjective: this is a test of context effects that could not have been caused (or masked) by people's inferences about adjective omission. Participants' inferences among only utterances with an adjective 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 or without an adjective.

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

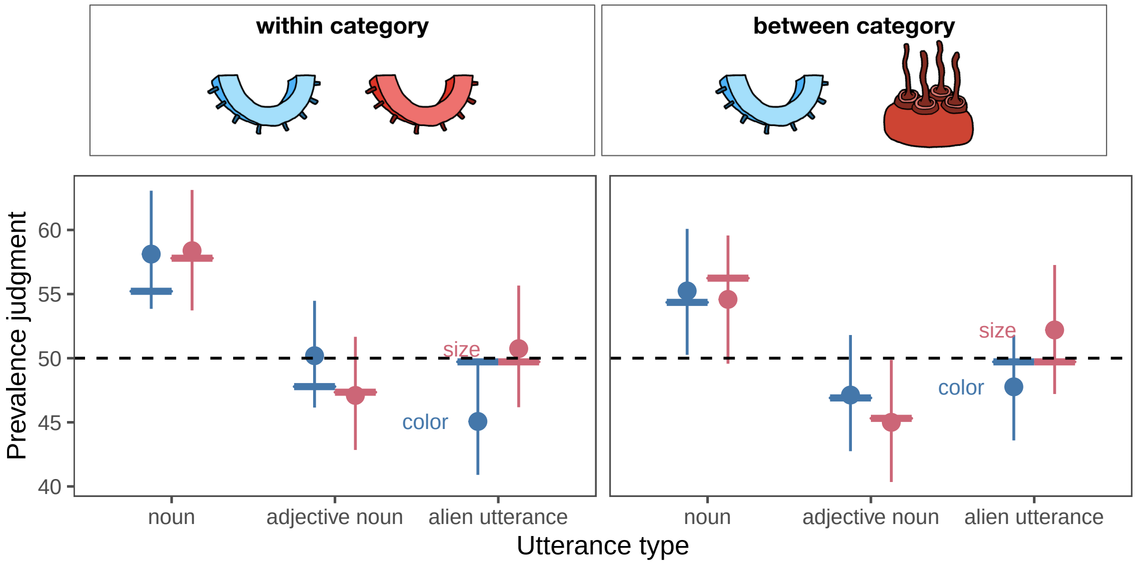

Figure 7. Participants' prevalence judgments in Experiment 3, with model predictions using the parameters estimated in Experiment 2 (horizontal lines).

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.

General Discussion

When we think about what someone is trying to communicate to us, we go far beyond the literal meanings of the words they say: we make pragmatic inferences about why they chose those particular words rather than other words they could have used instead. In most work on pragmatic reasoning, speakers and listeners share the same knowledge of language, and the question of interest is whether listeners can use their knowledge of language to learn something about the unknown state of the world. Here we focus on an even more challenging problem: Can pragmatic inference be used to learn about language and the world simultaneously?

In three studies we showed that people can use pragmatic inference to (1) learn the meaning of a novel word, (2) learn the typical features of the category described by this novel word, and (3) rationally integrate these two kinds of reasoning processes. In Experiment 1, we show that people can use descriptive contrast implied by adjectives like "big" or "blue" to resolve referential ambiguity to learn a new word; in the case of color, they shift substantially in the direction of the correct mapping, and in the case of size, they choose the correct mapping significantly more often than the incorrect one. In Experiments 2 and 3, we show that people infer that a noted feature is atypical of the object being referred to. Critically, people infer that the described feature is atypical even when the descriptor is helpful for referential disambiguation–although the size of the atypicality inference is numerically reduced.

Why do people think that the mentioned feature is atypical even when its mention is helpful for referential disambiguation? If people use language for multiple goals–for example, both for reference and for description– then listeners should reason jointly about all of the possible reasons why speakers could have used a word. To determine what rational listeners would do in this circumstance, we developed an extension of the Rational Speech Act Framework that reasons both about reference and about the typical features of categories to which objects belong. The behavior of this model was closely aligned to the behavior we observed in people. Because rational inference is probabilistic rather than deterministic, descriptors still lead to atypicality inferences even when they are helpful for referential disambiguation. This work thus adds to the growing body of work extending the Rational Speech Act framework from reasoning about just reference to reasoning about other goals as well, such as inferring that speech is hyperbolic, inferring when speakers are being polite rather than truthful, and learning new words in ambiguous contexts (Frank & Goodman, 2014; Goodman & Frank, 2016; Kao, Wu, Bergen, & Goodman, 2014; Yoon, Tessler, Goodman, & Frank, 2020).

In considering how people may integrate inferences about typicality and about reference, we raised two broad possibilities: (1) a *reference-first view*, whereby if an adjective was necessary for reference it would block an inference of atypicality completely, and (2) a *probabilistic weighing view*, whereby the goals of being informative with respect to reference and with respect to the category would trade off in a graded way. That is, we aimed to test whether there was a strong trade-off or a weak trade-off. Our model implements the latter view and fits the data well, but we do not find significant evidence of a trade-off in the regressions on people's responses: the data are also compatible with there being no trade-off whatsoever.

[XXXXXXXX]

Though the participants in our experiments were adults, the ability to disambiguate novel referents using contrast most obviously serves budding language learners—children. Contrastive use of adjectives is a pragmatic regularity in language that children could potentially exploit to establish word–referent mappings. Use of adjectives has been shown to allow children to make contrastive inferences among familiar present objects (Davies, Lingwood, Ivanova, & Arunachalam, 2021; Huang & Snedeker, 2008). When paired with other contrastive cues such as prosody, preschoolers can make inferences about novel object typicality (Horowitz & Frank, 2016), and can use novel adjectives and nouns to restrict reference (Diesendruck, Hall, & Graham, 2006; Gelman & Markman, 1985). Future work should explore whether adjective contrast that is less scaffolded by other cues is a viable way for children to learn about novel concepts.

The core computation in pragmatic inference is reasoning about alternatives–things the speaker could have said and did not. Given that others are reasoning about these alternatives, no choice is neutral. In the studies in this paper, for instance, using an adjective in referring to an object led people to infer that the feature described by that adjective was less typical than if it had not been mentioned. But, conversely, *not* using an adjective led them to think that the feature was more typical than if they could not understand the meaning of the utterance at all–all communicative choices leak one's beliefs about the world. This has implications not only for learning about novel concrete objects, as people did here, but for learning about less directly accessible entities such as abstract concepts and social groups. These inferences can be framed positively, as ways for learners to extract additional knowledge that was not directly conveyed, but can also spread beliefs that the speaker does not intend. A core challenge will be to understand how people reason about the many potential meanings a speaker might convey in naturalistic contexts to learn about others' words for and beliefs about the world.

Acknowledgements

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

The authors thank Ming Xiang and Susan Goldin-Meadow for guidance on early versions of this work and Benjamin Morris, Ashley Leung, Michael C. Frank, Ruthe Foushee, Judith Degen, and Robert Hawkins for feedback on the manuscript. Portions of this work were published in the proceedings of Experiments in Linguistic Meaning. The authors are grateful for feedback from reviewers and attendees of Experiments in Linguistic Meaning, the meeting of the Cognitive Science Society, and the Midwestern Cognitive Science Conference.

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.

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

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, C., Lingwood, J., Ivanova, B., & Arunachalam, S. (2021). Three-year-olds' comprehension of contrastive and descriptive adjectives: Evidence for contrastive inference. *Cognition*, *212*, 104707. https://doi.org/10.1016/j.cognition.2021.104707

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.

Diesendruck, G., Hall, D. G., & Graham, S. A. (2006). Children's Use of Syntactic and Pragmatic Knowledge in the Interpretation of Novel Adjectives. *Child Development*, *77*(1), 16–30.

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

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.

Gelman, S. A., & Markman, E. M. (1985). Implicit contrast in adjectives vs. Nouns: Implications for word-learning in preschoolers*. *Journal of Child Language*, *12*(1), 125–143.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

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

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.

Huang, Y. T., & Snedeker, J. (2008). Use of referential context in children's language processing. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society.*

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.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, *30*(1), 1–45. https://doi.org/10.1007/s10988-006-9008-0

Kreiss, E., & Degen, J. (2020). Production expectations modulate contrastive inference. In *Proceedings of the annual meeting of the cognitive science society.*

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.

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

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.

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. (2016). 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.

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.

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

Westerbeek, H., Koolen, R., & Maes, A. (2015). 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

Xiang, M., Kennedy, C., Xu, W., & Leffel, T. (2022). Pragmatic reasoning and semantic convention: A case study on gradable adjectives. *Semantics and Pragmatics*, *15*, 9:EA–9:EA. https://doi.org/10.3765/sp.15.9

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

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2020). Polite speech emerges from competing social goals. *Open Mind*, *4*, 71–87.

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