Using contrastive inferences to learn about new words and categories

Claire Bergey[1] & Dan Yurovsky[2]

[1] The University of Chicago

[2] Carnegie Mellon University

Author Note

All data and code for these analyses are available at https://osf.io/3f8hy/?view_only=9a196db0444c4867bc899cc70a7a1e9c.

Correspondence concerning this article should be addressed to Claire 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, listeners 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 present objects in the environment or to the referent's category. In three experiments, we investigate whether listeners use descriptive contrast to learn new word-referent mappings and learn about novel categories' feature distributions. People use size adjectives contrastively to guide referent choice, though they do not do so using color adjectives (Experiment 1). People also use description to infer that a novel object is atypical of its category (Experiment 2). 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 it predicts both of these inferences. Overall, people are able to use descriptive contrast 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: parent-child interaction; language development; communication

Word count: 1385

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, the utterance “We should hire a female professor” may convey much about the speaker’s goals, the makeup of a department, or even the biases of a field—none of which is literally stated. These pragmatic inferences are pervasive in everyday conversation: by reasoning about what someone says in relation to the context and what they might have said otherwise, we can glean more of their intended meaning. For example, [some empirical thing with familiar referents akin to our example above].

But what if you didn’t know the meaning of the key words in someone’s utterance, could you use the same kind of pragmatic inferences to learn about new words and categories? Suppose a friend asked you to “Pass the tall dax.” You might look around the room for two similar things that vary in height, and hand the taller one to them. Intuitively, your friend must have said the word tall for a reason (Grice, 1975). One possibility is that your friend wanted to distinguish the dax they wanted from the dax they did not. People appear to make these kinds of inferences quite rapidly for objects they know, for instance they already begin looking to a tall familiar object with a short competitor nearby–even if there are other tall objects around–as soon as they hear the word “tall” (Sedivy, K. Tanenhaus, Chambers, & Carlson, 1999).

If you only saw one object around whose name you didn’t know, you might draw a different inference: This dax might be a particularly tall dax. In this case, you might think your friend This is because your friend may have used the word “tall” for a different reason–not to distinguish the dax they wanted and 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 whih 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 also learned a new word, but should have you learned something about the typical size of daxes as well, beyond the two daxes you observed? One possibility is that you should 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 [cites], then people should draw no further inferences once the need for referential disambiguation can explain away a descriptor like "tall."

On the other hand, if pragmatic reasoning is probabilistic, people may integrate typicality as just one factor the speaker weighs in using description, leading to graded inferences about the referent's identity and about its category.

In this paper, we present a series of experiments to test whether and how listeners make inferences about novel referents using descriptive contrast. First, we examine whether listeners use descriptive contrast to resolve referential ambiguity. In a reference game, participants see groups of novel objects and are asked to pick one with a referring expression, e.g., "Find the blue toma." If participants interpret description contrastively, they should infer that the description was necessary to identify the referent–that the blue toma contrasts with some other-colored toma on the screen. Using this contrastive inference, they can resolve referential ambiguity, choosing a blue object with a similar non-blue counterpart rather than a blue object with no similar counterpart nearby. Second, we test whether listeners use descriptive contrast to make inferences about a novel object's category. Participants are presented with two interlocutors who exchange objects using referring expressions, such as “Pass me the blue toma.” If participants 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 to show that 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 prenominal adjective contrast to disambiguate referents, and how those inferences are affected by adjective type, we use a reference game 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. They have unknown names and feature distributions, creating the ambiguity necessary for our test of referential disambiguation. But the ability to disambiguate novel referents, or to establish reference with incomplete information, is also the broader problem of learning about the world. We know that distributional information in the world affects people’s pragmatic use and interpretation of description Westerbeek, Koolen, & Maes (2015). Here, we ask: can people use pragmatic inferences from description to learn about unfamiliar things in the world?

Experiment 1

In Experiment 1, we test whether participants use prenominal adjective contrast to choose a novel referent. In a similar task with familiar objects, Sedivy and colleagues showed that people interpret at least some prenominal adjective use as contrastive (Sedivy, K. Tanenhaus, Chambers, & Carlson, 1999). In their task, four objects appeared on a screen: 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. These experiments demonstrate that listeners interpret at least some prenominal adjectives contrastively, and use this contrastive inference to guide their attention allocation. This kind of contrastive inference can be derived from a rational speaker framework in which listeners reason that speakers using an utterance with a description, rather than one without, chose to do so to make a useful contribution to listener understanding (Frank & Goodman, 2012). 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 [pechmann_incremental_1989]. These experiments demonstrate that listeners use contrast among familiar referents to guide their attention allocation, though not their explicit referent choice, which occurs after the noun disambiguates the object.

In a referential disambiguation task, we presented participants with arrays of novel fruit objects (Figure 1). On critical trials, participants saw a target object, a lure object that shared the target's contrast feature but not its shape, and a contrastive pair that shared the target's shape but not its contrast feature. Participants heard an utterance denoting the feature: "Find the [blue/big] dax." For the target object, use of the adjective is necessary to disambiguate from the same-shape distractor; for the lure, the adjective would be 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 test participants in two conditions: color contrast and size contrast. Though we expect participants to shift toward choosing the item with a contrastive pair in both conditions, we do not expect them to treat color and size equally. Because color is often used redundantly in English while size is not (Nadig & Sedivy, 2002; Pechmann, 1989), we expect size to hold more contrastive weight, encouraging a more consistent contrastive inference.

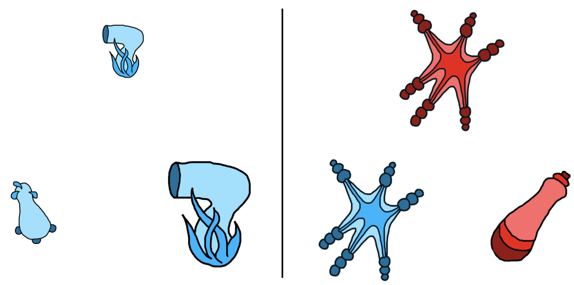


Figure 1. On the left: an example of a contrastive trial in which the critical feature is size. Here, the participant would hear the instruction "Find the small dax." On the right: an example of a contrastive trial in which the critical feature is color. Here, the participant would hear the instruction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

Method

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

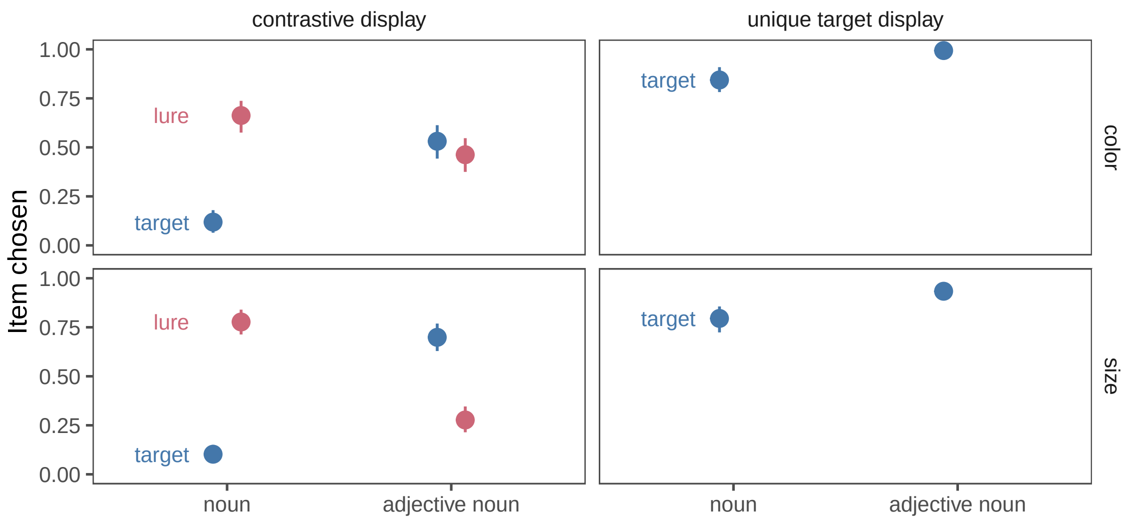


Figure 2. Proportion of times that participants chose the target and lure items as a function of condition and whether an adjective was provided. Points indicate group means; error bars indicate 95% confidence intervals computed by non-parametric bootstrapping.

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 has a unique shape and is 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 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 critical feature), and a lure (matched the target's critical feature but not shape). The 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 ("Find the [blue/big] dax"), and half of trials had audio instructions with no adjective description

("Find the dax"). 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 respond 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) or if they missed any of four color perception check trials (resulting n = 163).

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 ($\beta = 6.64$, $t = 4.10$, $p < .001$) and size ($\beta = 2.25$, $t = 6.91$, $p < .001$), indicating that participants consistently chose the unique object on the screen when given an instruction like "Find the (blue) dax." 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 ($\beta = 3.85$, $t = 3.52$, $p < .001$), and adjective type (color vs. size) was not statistically related to target choice (β = -0.48, t = -1.10, p = .269). The two effects did not significantly interact (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 contained the relevant adjective.

Our key test was whether participants would choose the target object on contrastive trials in which description was given, reflecting use of a contrastive inference to choose a novel referent. To do this, we compare participants' rate of choosing the target to their rate of choosing the lure, which shares the relevant critical feature with the target, when the audio described the critical feature. Overall, participants chose the target with a contrasting pair more often than the unique lure ($\beta = 0.53$, $t = 3.83$, $p = < .001$). Considering the adjective type conditions (color vs. size) separately, participants chose the target more than the lure in the size condition ($\beta = 0.86$, $t = 4.41$, $p = < .001$). However, participants in the color condition did not choose the target significantly more often than they chose the lure ($\beta = 0.15$, $t = 0.75$, $p = .455$). 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.

Discussion

When faced with unfamiliar objects referred to by unfamiliar names, people must resolve ambiguity to understand their conversational partner and learn more about the lexicon. In Experiment 1, we tested whether people could use contrastive inferences to resolve ambiguous reference to novel objects. We found that participants have a general tendency to choose objects that are unique in shape when reference is ambiguous. However, when people 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 find that people are able to use contrastive inferences about size to successfully resolve which unfamiliar object an unfamiliar word refers to.

Model 1

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 statements (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 a 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). 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 extension of RSA is that the pragmatic learner is a pragmatic listener who has 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 learners 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 2 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 dax." 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 (SHAPE BIAS CITE). Given this, the two possible mappings are $\{m_1 : hairdryer - \text{“}dax'', pear - \text{“}?''\}$, and $\{m_2 : hairdryer - \text{“}?'', pear - \text{“}dax''\}$ The literal semantics of each object allow them to be referred to by their shape label (e.g. "dax"), 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 dax," the model must now choose a referent. If the true mapping for "dax" is the hair dryer (m_1), this utterance is ambiguous to the literal listener, as there are two referents consistent with the literal meaning dax. Consequently, whichever of the two referents the speaker intends to point out to the learner, the speaker's utility will be relatively low. In contrast, if the true mapping for "dax" is the pear (m_1), 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 "dax."

If instead the speaker produced "Find the small dax," the model will make a different inference. If the true mapping for "dax" 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 "dax" means pear (m_1). However, if "dax" means pear, the speaker's utterance was inefficient because the single word utterance "dax" would have identified the target to the literal listener and incurred less cost. Thus, the model can infer that "dax" 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. "dax") should lead people to infer that the intended referent is the unique object (lure), whereas hearing a modified noun (e.g. "small dax") 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.

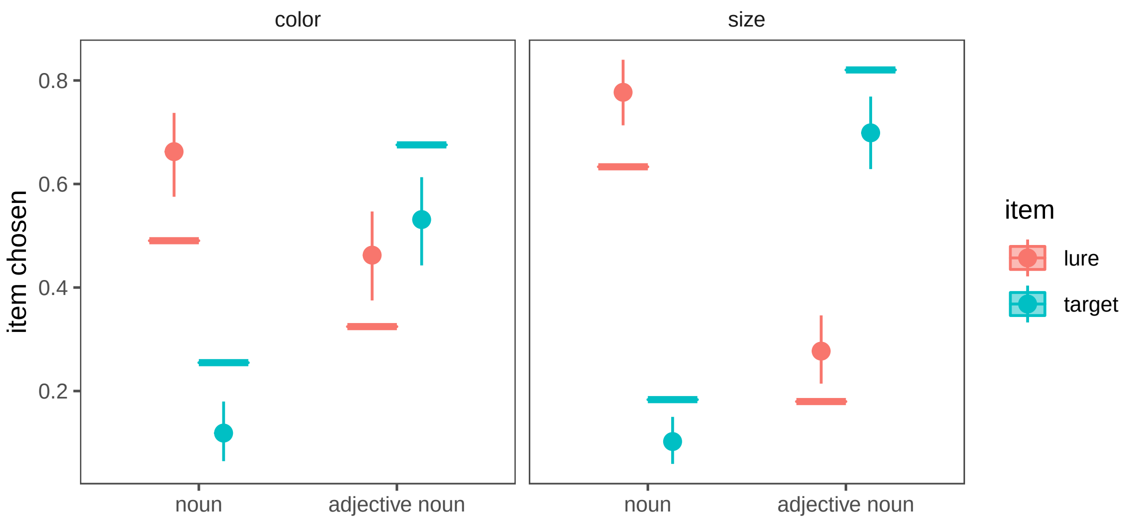


Figure 3. Proportion of times that people (and our model) chose the target and lure items as a function of adjective type and whether an adjective was provided. Points indicate empirical means; error bars indicate 95% confidence intervals computed by non-parametric bootstrapping. Solid lines show model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

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 in each condition, we used Empirical Bayesian inference to estimate the likely range of parameter values. These estimates varied substantially across conditions, with the rationality parameter in the color condition estimated to be 2.00 with a 95% credible interval of [1.37, 2.63], and the rationality parameter in the size condition estimated to be 3.98 [3.22, 4.74].

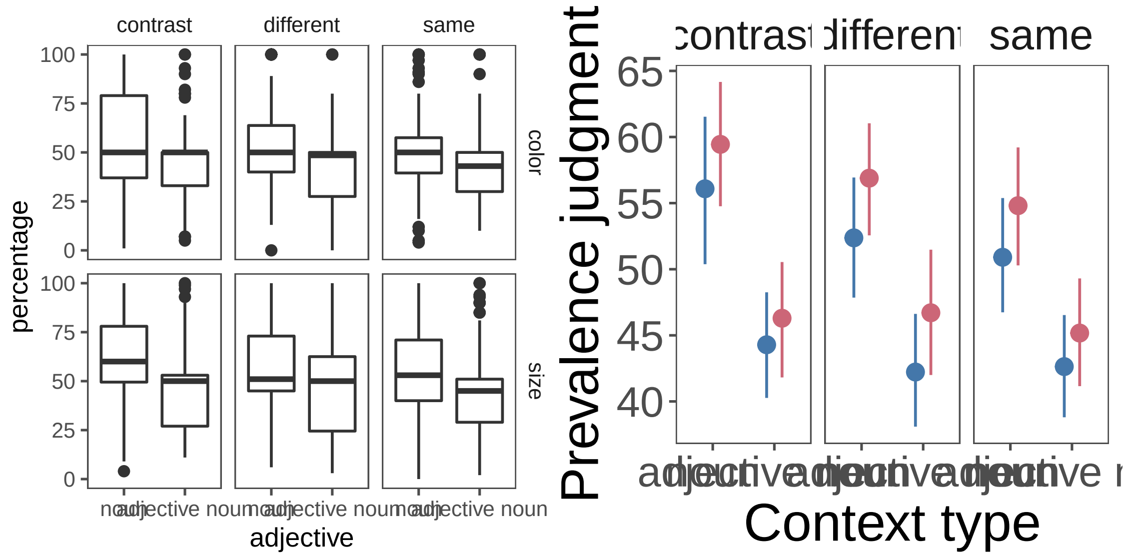
Figure 3 shows the model predictions along with the empirical data from Experiment 1. The model broadly captures the contrastive inference–when speakers produce an adjective noun combination like "red dax," the model selects the target object more often than the lure object. The extent to which the model 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 more likely, the model overpredicts the extent of the contrastive inference that people make. Intuitively, it appears that in over 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. "dax"). 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 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 by treating color as having stronger literal semantics (i.e. "blue dax" is a better description of a small blue dax than "small dax" is). However, this model predicts the opposite asymmetry of what we found. Because color has stronger semantics than size, listeners show a stronger contrast effect for color than size. We show this effect in appendix A. Thus, though a continuous semantics can explain our asymmetry, this explanation is unlikely given the continuous semantics that predicts other empirical color–size asymmetries does not predict our findings.

Experiment 2



In our first experiment, we examined whether people would interpret description as implying contrast with other present objects. However, as discussed earlier, description can imply contrast with sets other than the set of currently available referents. One of these alternative sets is the referent's category. Speakers use more description when referring to objects with atypical features (e.g., a yellow tomato) than typical ones (e.g., a red tomato) Rubio-Fernández (2016). 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. 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 (2016) demonstrates that speakers often use 'redundant' color adjectives to describe colors when they are central to the category's meaning (e.g., colorful t-shirts) or when they are atypical (e.g., a purple banana). Therefore, color may be no less contrastive with respect to the category's feature distribution. In the following experiment, 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 the prior experiment, 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. 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.

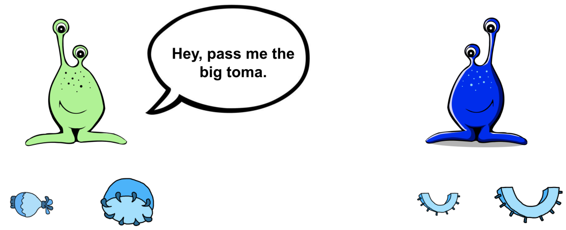


Figure 4. Experiment 2 stimuli. In the above example, the critical feature is size and the object context is a within-category contrast: the alien on the right has two same-shaped objects that differ in size.

Method

Participants. Two hundred and forty 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 (Figure 4) 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. Alien A, in a speech bubble, asked Alien B for one of its fruits (e.g., "Hey, pass me the red gade.") Alien B replied, "Here you go!" and the referent disappeared from Alien B's side and reappeared on Alien A's side.

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. 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 (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. 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. Thus, in the within-category contrast condition, the descriptor is necessary to distinguish the referent; in the between-category contrast condition it is unnecessary but potentially helpful; and in the same feature condition it is unnecessary and unhelpful. Note that in all context conditions, the set of objects onscreen was the same in terms of the experiment design; in each condition, they were rearranged such that the relevant referents (the objects under Alien B) were different. 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. We manipulated the critical feature type (color or size) between subjects.

Participants performed 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?" and 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 XX participants, leaving XX for further analysis.

Results

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: effects utterance type (adjective or no adjective), context type (within category, between category, or same feature), 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, and fixed effects were removed if they did not improve model fit (XXX CHECK THIS). The final model revealed a significant effect of utterance type ($\beta_{adjective}$ = -11.80, t = -3.90, $p < .001$), such that prevalence judgments were lower when an adjective was used than when it was not. Participants also made lower prevalence judgments in the same-feature context type relative to within-category context type (β_{same} = -5.41, t = -2.25, p = .025), but there was no significant effect of between-category relative to within-category contexts ($\beta_{between}$ = -3.92, t = -1.63, p = .104). There was not a significant interaction between context and presence of an adjective in the utterance ($\beta_{same*adjective}$ =

3.71, $t = 1.09$, $p = .277$; $\beta_{between*adjective} = 1.58$, $t = 0.46$, $p = .644$). That is, participants slightly adjusted their inferences according to the object context, though not in a way that depended on whether an adjective was used in the utterance. However, they robustly inferred that described features were less prevalent in the target's category than unmentioned features.

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, 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 abouttomatoes would infer that tomatoes are relatively unlikely to be red.

[add paragraph about diff in color/size asymmetry between exps 1 and 2, people tracking production norms on the level of the type of contrast set]

Model 2

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). 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, Hawkins, Graf, Kreiss, and Goodman (2020)'s continuous semantics Rational Speech Act model.

[FIX THIS] Extensions of this framework have successfully accounted for a variety of other pragmatic inferences, including inference that speech is hyperbolic (e.g. waiting "a million years" means waiting a long time), 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). Further, a recent extension of the framework using continuous rather than discrete semantics has given an account of the kinds of differences between color and size modification that we observed in our experimental data (Degen, Hawkins, Graf, Kreiss, & Goodman, 2020).

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 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 of them from the modified category is much higher[1]

[add paragraph about how the typicality is one term in speaker utility]

If speakers use this utility function, listeners who do not know the feature distribution

[1] This is a generalization of Xu and Tenenbaum (2007)'s size principle to categories where exemplars are not equally likely.

for a category can use speakers' production to infer it. Intuitively, speakers should prefer not to modify nouns with adjectives because they incur a cost for producing that adjective. If they did, it must be because they thought the learner would have a difficult time finding the referent from a bare noun alone because of typicality and/or competing referents. 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 listener'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 listeners prior is thus updated to be $\text{Beta}\,(2, 2)$.

As in Experiment 1, we used Empirical Bayesian methods to estimate the 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 0.96 [0.69, 0.85] than color adjectives 0.96 [0.45, 0.85], but the two inferred confidence intervals were overlapping, suggesting that people treated the adjective types as more similar to each other.

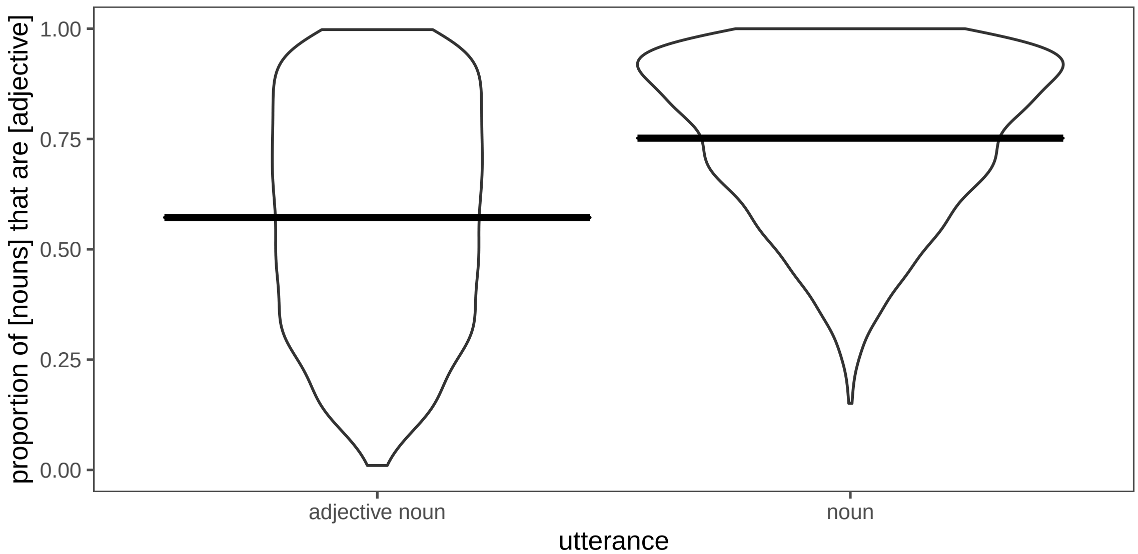


Figure 5. Model estimates of typicality judgments for one object seen alone and labeled either [noun] or [adjective noun].

Figure **??** 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 if it is referred to with an adjective (e.g., “red dax”) than if it was not mentioned (e.g., “dax”). The model also infers the prevalence of the critical feature to be numerically more likely in the contrast condition, like people do. That is, in the contrast condition when an adjective is used to distinguish between referents, the model thinks that the target color is slightyly 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 more prevalent.

Discussion

In contrast to the reference-first view that these two kinds of inferences trade off strongly–that is, adjectives are used primarily for reference, and such use blocks the inference that they are marking typicality–the model captures the graded way in which people interpolate between them. When an adjective is helpful for reference, whether it is used or

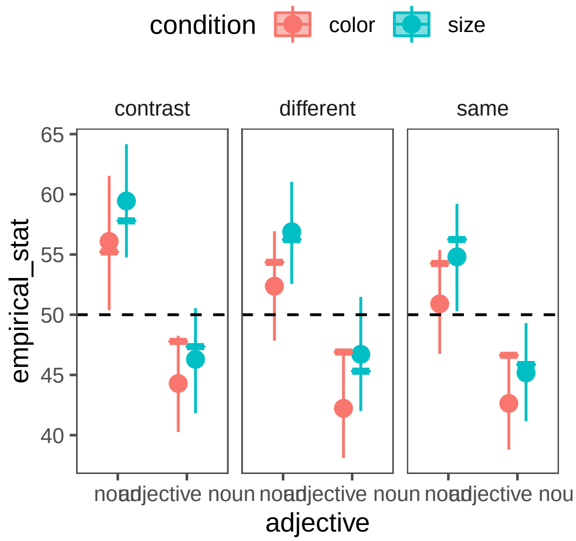


Figure 6. Model predictions for Experiment 2.

not makes both the model and people give it slightly less weight in inferring the typical features of the target object, but the weight is still significant. Our model's explanation for this is that while people choose their language in order to refer successfully, their choices also reflect their knowledge of features of those objects. In the model as constructed, we cannot distinguish between listener and speaker design explanation for the impact of feature knowledge. One possibility is that the pressure from this feature knowledge is communicative as well speakers could be intentionally transmitting information to the listener about the typical features of their intended referent. Alternatively, the influence of this feature knowledge could be unintentional, driven by pressures from the speaker's semantic representation. We consider these implications more fully in the General Discussion. In either case, listeners can leverage the impact of speakers' feature knowledge on their productions in order to infer the typical features of the objects they are talking about, even if this is their first exposure to these novel objects.

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.

To strengthen our findings in a way that would allow us to better detect potential trade-offs between these two types of inference, here we replicate Experiment 2 in a larger sample of participants. In addition, we test 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 cannot 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 test the model here by testing the same model, with the same inferred parameters as in Experiment 2, on data from Experiment 3.

Method

Participants. Four hundred 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 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, the descriptor 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 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). 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 XX participants, leaving XX for further analysis.

Results

We began by fitting a pre-registered maximum mixed-effects linear model: effects 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 ($\beta = 13.05$, $t = 4.88$, $p = < .001$) and a marginal effect of the adjective utterance type compared to the alien utterance type ($\beta = 5.13$, $t = 1.95$, $p = .052$). The effects of context type (within-category or between-category) and adjective type (color or size) were not significant ($\beta_{between} = 2.70$, $t_{between} = 1.23$, $p_{between} = .220$; $\beta_{size} = 5.68$, $t_{size} = 1.70$, $p_{size} = .090$). There was a significant interaction between the adjective utterance type and the size condition ($\beta =$ -8.78, $t =$ -2.31, $p = .022$). 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. They also inferred that an object referred to in an intelligible utterance with description was marginally less typical than an object referred to with an alien utterance, and this effect was slightly stronger in the size condition. They did not substantially adjust their inferences based on the object 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. This model revealed a significant effect of utterance type: participants' prevalence judgments were lower when an adjective was used than when it was not ($\beta =$ -7.92, $t =$ -3.38, $p = .001$). No other effects were significant. This replicates the main effect of interest in Experiment 2: that 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.

Model 3

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 show in Figure 7, the model predictions were well aligned with peoples' 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 on 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, peoples' judgments were indeed around 50%, although in the color condition they were directionally lower. This small effect may arise from a fundamental difference between polar adjectives like size (where objects can be big or small) and adjectives like color where there may be many nameable alternatives (e.g. red, blue, green, etc.). Thus, the results of Experiment 3 confirm the modeling assumptions we made in estimating peoples' 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 think about not only the set of present objects as providing the context of referents, but also the broader set of categories that they belong to.

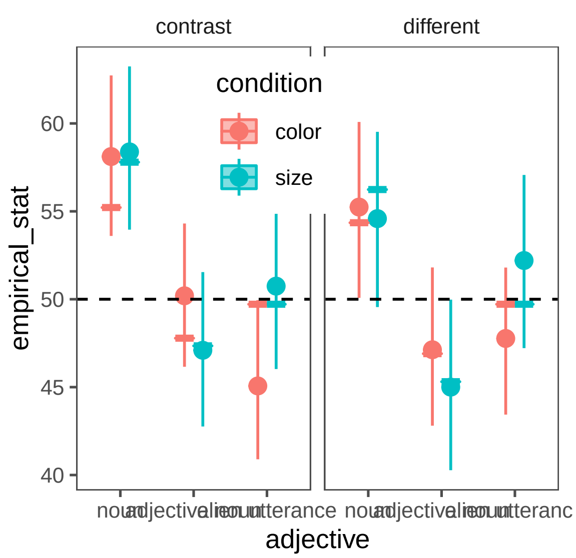


Figure 7. Model predictions for Experiment 3

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* effect 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

Overall, we found that people are able to use descriptive contrast to infer the referent of a novel word and to make inferences about a novel referent's category. In our first experiment, participants were able to resolve referential ambiguity using a contrastive interpretation of size adjectives, though not reliably with color adjectives. In our second and third experiments, participants inferred that a described referent was atypical of its category on that feature: hearing "big toma" or "blue toma" led them to think that most tomas were not that size or color, respectively. In real life it is often unclear whether description is meant to contrast with present objects or imply atypicality. In Experiments 2 and 3, participants did not significantly adjust their prevalence judgments based on the interaction of adjective use and object context—that is, they did not adjust their inferences about typicality based on how redundant description was in context. Contexts in which description was necessary to identify the referent did not preempt inferences of atypicality.

In Experiment 1, participants notably failed to use color adjectives contrastively in choosing referents. What makes size different from color? One possibility is that color adjectives are often used redundantly, and therefore receive less contrastive weight than adjectives consistently used to differentiate between referents. Sedivy (2003) puts forth such an account, finding that color adjectives tend not to be interpreted contrastively in eye-tracking measures except in contexts that make their use unlikely. In comparison, adjectives describing material (e.g., plastic) and size are interpreted contrastively, which corresponds to less redundant use of material and size adjectives in production (Sedivy, 2003). Further work is necessary to determine whether contrastive inferences hew to production norms, and whether implicit indications of contrast usually extend to explicit referent choice.

In Experiment 2, we asked whether utterances like "Pass me the blue dax" lead people to infer that daxes are generally less likely to be blue. We found that people robustly infer that mentioned features are atypical of the object's category, across both color and size adjectives and in varying object contexts.

In Experiment 3, we replicated Experiment 2 and asked what kinds of inferences people make about novel object typicality when they cannot understand the referring expression. We found that people tend to infer that the feature is as prevalent as their direct experience would suggest, around the same as our model's estimate after observation of the objects and before hearing an utterance. This is significantly less than their prevalence judgment when they hear the object referred to with a noun and no adjective (e.g., "Pass me the dax"). That is, people infer that an object is fairly typical when it is referred to in a sentence they understand, but think it is less typical—only as typical as their prior indicates—when it is referred to in a completely incomprehensible utterance. This suggests that even simple mentions, such as "Pass me the toma," prompt inferences about the typicality of the object in its category (namely, that this toma is typical). While the effects we show here are appropriately described as atypicality inferences from description, this result suggests that people's inferences about typicality are not simply inferring 'markedness' from the use of an adjective; any mention of an object can engender inferences about its category.

The relative robustness of contrastive inferences about typicality across contexts and adjective types compared to contrastive inferences among present referents raises questions about the relative importance of these two kinds of contrast in language understanding. Most prior work has focused on contrast with present referents as the main phenomenon of interest, with object typicality as a modulating factor; our results emphasize the role of contrast with an object's category, particularly when ambiguity is at play. A reference-first view of utterance interpretation might predict that use of description would be largely explained away if the description was necessary for reference (e.g., the 'red' in 'red dax' is explained by a blue dax being present to distinguish from). Contrary to this possibility, we find that both our participants and a probabilistic model that integrates both referential utility and typicality make inferences of atypicality even when the adjective was necessary to establish reference. The model slightly weakens its inference of atypicality in this case, and participants' inferences do not significantly differ based on object context. Future work will explore whether people make subtle trade-offs between contrast with present referents and with the referent's category.

[add RSA stuff]

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. Further, use of adjectives has been shown to allow children to make contrastive inferences among familiar present objects (Huang & Snedeker, 2008) and, when paired with other contrastive cues such as prosody, about novel object typicality (Horowitz & Frank, 2016); future work will explore whether adjective contrast alone is a viable learning tool in early childhood. Tasks using a mixture of novel adjectives and words suggest that children as young as 3 can make contrastive inferences about adjectives (Diesendruck, Hall, & Graham, 2006; Gelman & Markman, 1985; Huang & Snedeker, 2008). Contrastive inferences allow people to learn the meanings of new words and the typical features of new categories, pointing to a broader potential role of pragmatic inference in learning about the world.

Conclusion

Taken together, these experiments show that people use contrastive inference to map novel words to novel referents and to make inferences about the typicality of novel referents' features. Hearing "small toma" allows people to narrow possible referents not only to small objects, but objects with larger counterparts nearby. Hearing "big toma" in a referential context leads them to think that most tomas are not that size. However, these two abilities do not appear to interact. A referential felicitous use of description does not block an inference of atypicality. These results do not yet provide an explanation of *why* these skills do not interact: the inference may be too complex, the stimuli too novel, or listeners may use contrast more heuristically than rational models of pragmatic inference assume (Frank & Goodman, 2012). Understanding the origins of these independent but non-interpendent inferential abilities, as well as asymmetries between comprehension and production and adjectives like color and size, will be an important next challenge in our development of theories of human pragmatic inference.

Acknowledgements

This research was funded by a James S. McDonnell Foundation Scholar Award to the last author.

References

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.

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*.

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.

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.

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.

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.

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

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*.

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

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

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.