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 resolve reference and make inferences about novel referents’ categories. People use size adjectives contrastively to guide referent choice, 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). Finally, Overall, people are able to use descriptive contrast to resolve reference and make inferences about a novel object’s category, but limits to these abilities present further questions about the effect of context on listener interpretation."

Keywords: parent-child interaction; language development; communication

Word count: 1385

Using contrastive inferences to learn about new words and categories

Introduction

When trying to communicate, human listeners are faced with uncertainty. Novice listeners—children—face a continuous speech stream filled with unknown words referring to unformed concepts. Even seasoned listeners—adults—contend with noise, variable pronunciation, ambiguous meanings, and the occasional unknown word, too. Fortunately, listeners bring sensitive phonetic, syntactic, and semantic skills to the task, allowing them to reduce ambiguity during conversations and over developmental time. Most of these well-documented skills are concerned with the listener's understanding of the speaker's utterance alone. But communication occurs in context: in a rich world to which language refers. Listeners' ability to combine utterance information with context—their pragmatic ability—may be a powerful tool in resolving referential ambiguity and learning about the concepts language describes.

One potential pragmatic tool for reducing referential uncertainty is contrastive inference. Contrastive inferences are those inferences that derive from the principle that description should discriminate. This principle falls out of the more general Gricean maxim that speakers should say as much as they need to say and no more (Grice, 1975). 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; Sedivy, 2002, 2003). 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 contrastively, often describing more than is needed to establish reference (Engelhardt, Bailey, & Ferreira, 2006; Mangold & Pobel, 1988; Pechmann, 1989). How, then, do listeners interpret these descriptions?

Sedivy and colleagues carried out a visual world task demonstrating that adults 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. Participants then heard a referential expression: "Pick up the tall cup." Adults 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., the tall pitcher). 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 kind of 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 effect was demonstrated for size and material adjectives; the results for color adjectives were mixed (Sedivy, 2003; Sedivy et al., 1999). More recently, this contrastive processing effect was replicated with 5-year-old participants using size adjectives (Huang & Snedeker, 2008). These experiments demonstrate that listeners interpret at least some prenominal adjectives contrastively, and use this contrastive inference to guide their attention allocation. These results leave open, however, whether listeners use prenominal adjective contrast to resolve referential ambiguity and explicitly guide their referent choice.

Beyond contrasting a referent with other objects in the environment, description may draw a contrast between a referent and its category. In production studies, participants tend to describe atypical features more than they describe typical ones (Mitchell, Reiter, & Deemter, 2013; Rubio-Fernández, 2016; Westerbeek, Koolen, & Maes, 2015). For instance, they almost always include a color descriptor when referring to a blue banana, but not when referring to a yellow one. This, too, can be derived from a rational model of speaker behavior, but one with graded semantics in which the utterance “banana” fits a yellow banana better than a blue one (**???**). How do listeners interpret such adjective use? Suppose someone hears a referring expression to an unfamiliar object: “Look at that red sprocket.” In order to determine whether “red” was used in contrast to other objects in the environment or to the referent’s category, a rational listener must integrate contextual information. If there are many sprockets of different colors around, “red” was likely used to pick out an individual sprocket. If not, it may have been used to mark the abnormality of this sprocket—perhaps it is rare for sprockets to be red. In this way, it is possible for listeners to make inferences about the category of a novel referent using descriptive contrast.

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 in the array. 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. However, context should matter in these judgments: if the descriptor was necessary to identify the referent, an inference of contrast with the category is unwarranted.

In order to determine whether adults 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 avoid effects of typicality and familiarity that relate to level of description in production (Pechmann, 1989; Rubio-Fernández, 2016) on particular features (Mangold & Pobel, 1988). 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. This skill would aid not only adult speakers dealing with ambiguous or degraded communicative signal, but also children who need to establish new word–referent mappings. Across the developmental span, contrastive inference could help listeners exploit regularities in language and their environment to learn about both.

Experiment 1

In Experiment 1, we test whether adult participants use prenominal adjective contrast to choose a novel referent. To examine whether contrast occurs across adjective types, we test participants in two conditions: color contrast and size contrast. In a task similar to that of Sedivy and colleagues (1999), we present participants with arrays of novel fruit objects. On critical trials, participants see a target object, a lure object that shares the target's contrast feature but not its shape, and a contrastive pair that shares the target's shape but not its contrast feature. Participants hear 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. However, we do not expect listeners 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.

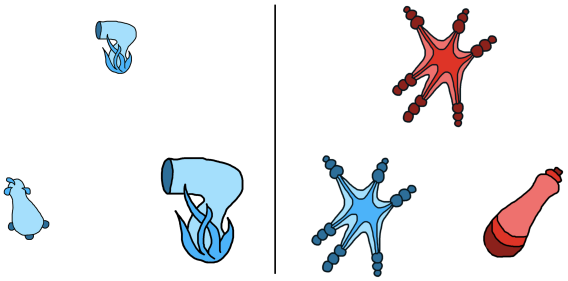

Figure 1. On the left: an example of a contrastive trial in which the critical feature is size. Here, the participant would hear the instruction "Find the small dax." On the right: an example of a contrastive trial in which the critical feature is color. Here, the participant would hear the instruction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

Method

Participants. 300 participants were recruited from Amazon Mechanical Turk. participants were assigned to a condition in which the critical feature was color (stimuli contrasted on color), and participants were assigned to a condition in which the critical feature was size.

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 has an instance in each of four colors (red, blue, green, or purple) and two sizes (big or small). There were two display types: unique target displays and contrastive displays. Unique target displays contain 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 match each other's (but not the target's) shape and critical feature. Contrastive displays contain a target, its contrastive pair (matches the target's shape but not critical feature), and a lure (matches the target's critical feature but not shape). The positions of the target and distractor items were randomized within a triad configuration.

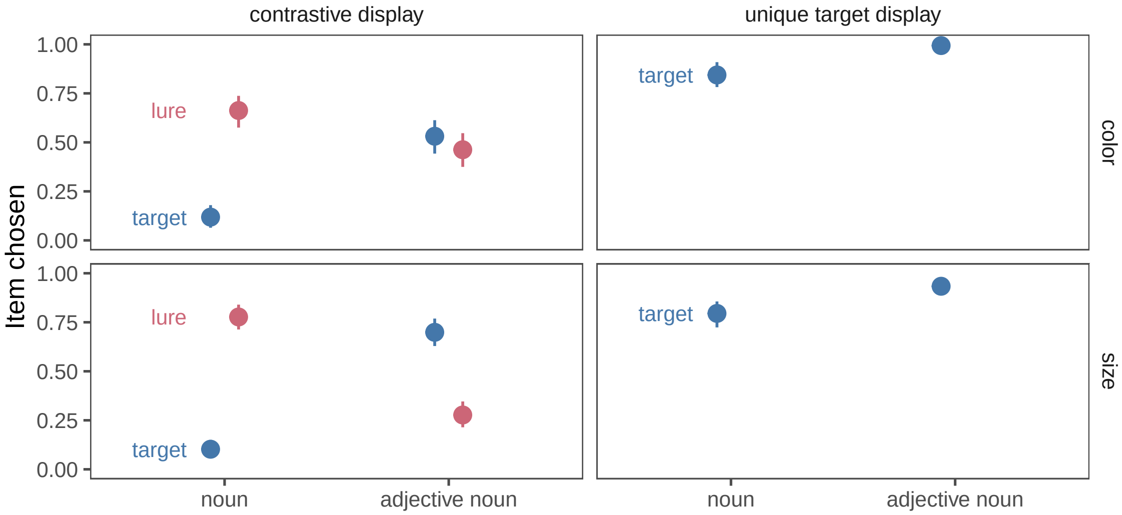

Figure 2. Proportion of times that participants chose the target and lure items as a function of condition and whether an adjective was provided. Points indicate group means; error bars indicate 95% confidence intervals computed by non-parametric bootstrapping.

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.

Results

We first confirmed that participants understood the task by analyzing performance on 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 ($\beta =$, $t =$, p) and size ($\beta =$, $t =$, p). 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. Adjective type (color vs. size) was not statistically related to target choice ($\beta =$, $p =$), and adjective description in the audio increased target choice ($\beta =$, $t =$, p). The two effects did not interact ($\beta =$, $t =$, $p =$). Participants had a general tendency to choose the target in unique target trials, which was amplified if the audio instruction contained the relevant contrast 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 contrast feature with the target, when the audio described the contrast feature. Participants chose the target more than the lure in the size condition ($\beta =$, $t =$, $p =$). However, participants in the color condition did not choose the target significantly more often than they chose the lure ($\beta =$, $t =$, $p =$). 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; this was true across color ($\beta =$, $t =$, $p =$) and size ($\beta =$, $t =$, $p =$) conditions. Adjective use therefore increased target choice ($\beta =$, $t =$, p) across contrastive trials. 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.

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 conditions of statements, and chooses randomly among all referents consistent with a statement (e.g. it is true that a red toma can be called "toma" and "red toma" but not "blue toma"). In planning their referential 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.

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

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

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

This computation naturally predicts a number of phenomena in pragmatics. For example, RSA explains scalar implicature–listeners treat "I ate some of the cookies" as a poor description of a case where the speaker at all of the cookies. The speaker's statement is literally true–the speaker eating some of the cookies is consistent with a world in which they ate all of then. However, this statement is ambiguous–it is true of both the world in which some cookies remain and the world in which there are no cookies left. Thus, if the speaker intends to convey that they ate all of the cookies, saying "I ate some of the cookies" will cause the literal listener to guess the wrong world half of the time. In contrast, the statement "I ate all of the cookies" is consistent only with world in which all of the cookies were eaten. Thus, if the speaker ate all of the cookies, this statement would accomplish their goal of communicating the state of the world more effectively. Scalar implicature arises from exact this inference: If the speaker actually ate all of the cookies, they should have said "I ate all of the cookies" because that would be a better utterance than "I ate some of the cookies." Since they produced "some," it is more likely that they wanted to communicate about the other possible world (Frank & Goodman, 2012).

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 (**???**; **???**; **???**; **???**). 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).

For this experiment, we build on a Rational Speech Act model developed by (**???**) to jointly resolve reference and learn 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." On the assumption that nouns generally refer to shapes, 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 constrat, 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 is 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

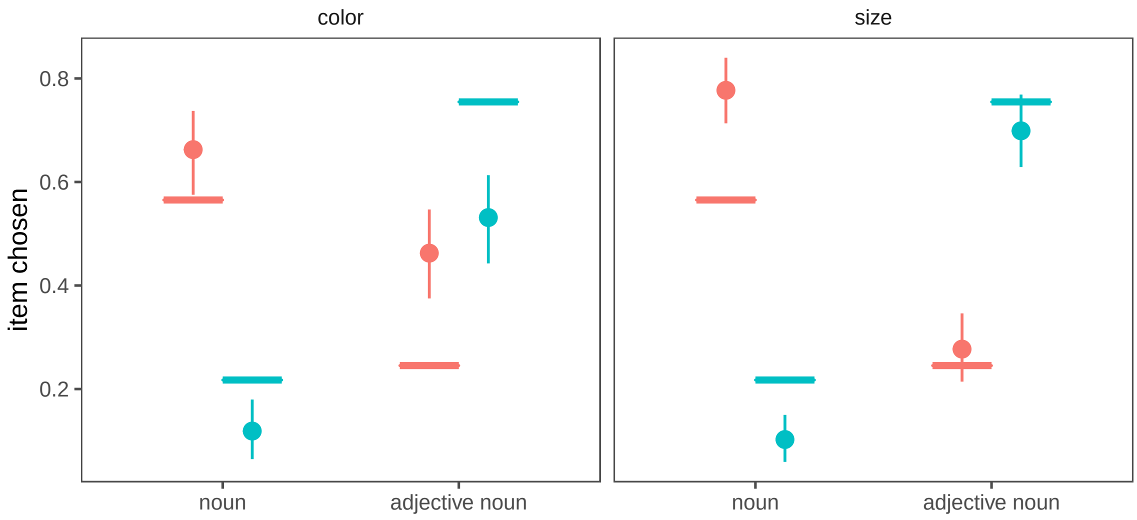

Figure 3. Proportion of times that people (and our model) chose the target and lure items as a function of adjective type and whether an adjective was provided. Points indicate empirical means; error bars indicate 95% confidence intervals computed by non-parametric bootstrapping. Solid lines show model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

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 is the target.

Because the model we described has no way of distinguishing between color and size adjectives, it makes the same predictions for both. Based on or pilot studies, we pre-registered and observed an asymmetry in which contrast inferences would be stronger for size than color. Why do we see an asymmetry in people? A recent model from Degen et al. (2020) does predict a color-size asymmetry. In this model, literal semantics are treated as continuous rather than discrete, so "blue" is neither 100% true or 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). We implemented this model using the same semantic values of color and size as given in the paper, but found that this specification predicts the opposite asymmetry of what we found in the paper. Because color has stronger semantics than size, listeners show a stronger contrast effect for color than size. We show this effect in appendix A.

What does explain the asymmetry we observed? Some possibilities:

Experiment 2

In our first experiment, we examined whether adult listeners 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. Work by Mitchell et al. (2013) and Westerbeek et al. (2015) demonstrates that speakers use more description when referring to objects with atypical features (e.g., a yellow tomato) than typical ones (e.g., a red tomato). This 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. In the following experiment, we test whether listeners use this type of contrast to learn about unfamiliar objects' categories.

If listeners do use this type of contrast, 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 likely was not used to mark atypicality. We therefore manipulate the context of the objects around the referent to see whether listeners adjust their inferences accordingly.

Method

Participants. Two hundred and forty participants were recruited from Amazon Mechanical Turk. One hundred and twenty participants were assigned to a condition in which the critical feature was color (red, blue, purple, or green), and 120 participants were assigned to a condition in which the critical feature was size (small or big).

Stimuli & Procedure. Stimulus displays showed two alien interlocutors, one on the left (Alien A) and one on the right (Alien B) side 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 (hereafter abbreviated as "contrast"), 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 (abbreviated as "different"), 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 (abbreviated as "same"), 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. 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 first 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 (contrast, different, or same), 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. The final model revealed significant effects of utterance type ($\beta_{adjective} =$, $t =$, p), critical feature ($\beta_{size} =$, $t =$, p) and a marginally lower prevalence for same search type relative to contrast search type ($\beta_{same} =$, $t =$, $p =$). Prevalance judgments for different trials was not reliably different from contrast trials ($\beta_{different} =$, $t =$, $p =$). Participants robustly inferred that described features were less prevalent in the target's category than unmentioned features. This atypicality inference was marginally stronger for trials on which the distractor had the same feature as the target, making the descriptor particularly unhelpful, than on trials in which the descriptor was necessary to distinguish between two objects of the same type. Overall, however, participants failed to substantially adjust their inferences according to the context of the referring expression.

Thus, participants treated all adjectives as marked, and inferred lower typicality, regardless of whether they could felicitiously be interpreted as contrasting between potential target referents. But were participants nonetheless sensitive to this information in their response times? We investigated this question by analyzing participants' time to advance after seeing the aliens' referential exchange. Though this task was not speeded, we hypothesized that participants would advance more quickly after seeing referential exchanges that were easier to process. After dropping all response times less than 1 second and longer than 10 seconds, and log transforming them because of the right skew in response time data, we predicted participants' time to advance on each trial of the experiment from utterance type, context type, critical adjective type, and the interaction between utterance type and context type (`log(rt) ~ adjective * search + type + (1 |subj)`). This model showed a reliable effect of utterance type ($\beta_{adjective} =$, $t =$, p)–participants were faster when an a descriptor was provided despite having to process an additional word. There was no main effect of critical adjective type ($\beta_{size} =$, $t =$, $p =$), nor context type ($\beta_{different} =$, $t =$, $p =$; $\beta_{same} =$, $t =$, $p =$), but the interactions between utterance type and context type trended towards significance for both non-contrast searches ($\beta_{adjective*different} =$, $t =$, $p =$; $\beta_{adjecgive*same} =$, $t =$, $p =$). Directionally, these results indicate that participants took longer to process utterances which were under-described (contrast trials with no adjective) than those with appropriately no description, and processed trials with an appropriate level of description (contrast trials with an adjective) more quickly than those with superfluous description.

Discussion

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 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 differences between color and size or trade-offs between these two types of inference, here we replicate Experiment 2 in a larger sample of participants. … [some explanation of why the new control condition is interesting as well …]

Method

Participants. Four hundred participants were recruited from Amazon Mechanical Turk. Two hundred were assigned to a condition in which the critical feature was color (red, blue, purple, or green), and 200 participants were assigned to a condition in which the critical feature was size (small or big).

Stimuli & Procedure. Stimulus displays showed two alien interlocutors, one on the left (Alien A) and one on the right (Alien B) side 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 two levels: within-category contrast and between-category contrast. In the within-category contrast condition (hereafter abbreviated as "contrast"), 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 (abbreviated as "different"), 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. 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 the color shown below?" with an image of the target object they just saw available on the screen. They answered on a slider scale from 0 to 100.

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 first 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 (contrast, different, or same), 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. The final model revealed significant effects of utterance type ($\beta_{adjective} =$, $t =$, p), critical feature ($\beta_{size} =$, $t =$, p) and a marginally lower prevalence for same search type relative to contrast search type ($\beta_{same} =$, $t =$, $p =$). Prevalance judgments for different trials was not reliably different from contrast trials ($\beta_{different} =$, $t =$, $p =$). Participants robustly inferred that

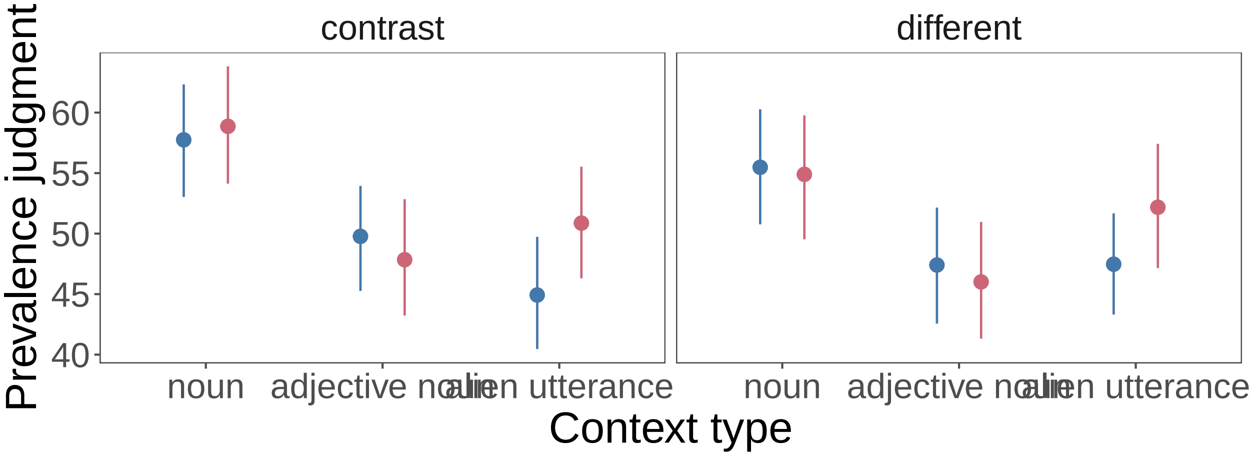

Figure 4. The proportion of the novel category participants judged to have the feature of the target object, by condition. The left panel shows judgments on trials in which no adjective was used in the referring expression (e.g., “Pass me the blicket”), and the right panel shows judgments on trials in which an adjective was used (e.g., “Pass me the [purple/small] blicket”). This is crossed by the type of object context (contrast, different, same) on the x-axis.

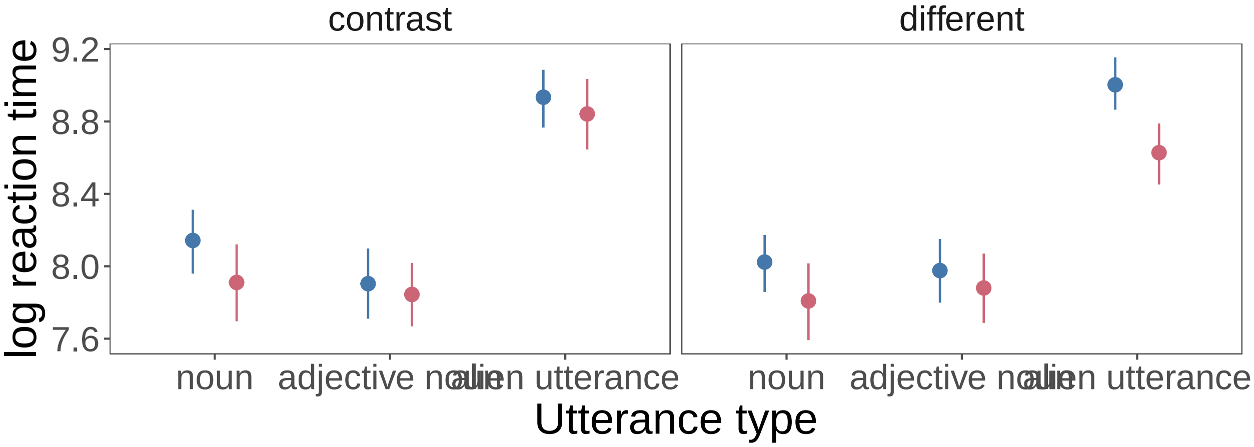

Figure 5. The log reaction time participants took to advance after seeing the referential exchange, by condition.

described features were less prevalent in the target's category than unmentioned features. This atypicality inference was marginally stronger for trials on which the distractor had the same feature as the target, making the descriptor particularly unhelpful, than on trials in which the descriptor was necessary to distinguish between two objects of the same type. Overall, however, participants failed to substantially adjust their inferences according to the context of the referring expression.

Thus, participants treated all adjectives as marked, and inferred lower typicality, regardless of whether they could felicitiously be interpreted as contrasting between potential target referents. But were participants nonetheless sensitive to this information in their response times? We investigated this question by analyzing participants' time to advance after seeing the aliens' referential exchange. Though this task was not speeded, we hypothesized that participants would advance more quickly after seeing referential exchanges that were easier to process. After dropping all response times less than 1 second and longer than 10 seconds, and log transforming them because of the right skew in response time data, we predicted participants' time to advance on each trial of the experiment from utterance type, context type, critical adjective type, and the interaction between utterance type and context type (`log(rt)` $\sim$ `adjective * search + type + (1 |subj)`). This model showed a reliable effect of utterance type ($\beta_{adjective} =$, $t =$, p)–participants were faster when an a descriptor was provided despite having to process an additional word. There was no main effect of critical adjective type ($\beta_{size} =$, $t =$, $p =$), nor context type ($\beta_{different} =$, $t =$, $p =$; $\beta_{same} =$, $t =$, $p =$), but the interactions between utterance type and context type trended towards significance for both non-contrast searches ($\beta_{adjective*different} =$, $t =$, $p =$; $\beta_{adjecgive*same} =$, $t =$, $p =$). Directionally, these results indicate that participants took longer to process utterances which were under-described (contrast trials with no adjective) than those with appropriately no description, and processed trials with an appropriate level of description (contrast trials with an adjective) more quickly than those with superfluous description.

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 cost of referring to atypical object with a bare noun. This cost could arise from two possible sources. One possibility is that speakers are motivated to help listeners to select the correct referent not just eventually but as quickly as possible. In this case, they should prefer not to produce bare nouns (e.g. "bird") to refer to atypical examples of that category (e.g. "penguin"). This is roughly the kind of inference encoded in (**???**)'s continuous semantics Rational Speech Act model. Alternatively, the cost may arise from difficulty of lexical access: In order to refer to an object, speakers need to retrieve the linguistic label appropriate. The difficulty of retrieving a lexical label may be proportional to how difficult it is [CITE?]. We are agnostic about whether this cost comes from speaker or listener design, and it makes no difference for the model's predictions. BETA DISTRIBUTION SIZE PRINCIPLE (**???**).

Listeners draw inferences about speakers as in regular RSA, but now need to figure out whether an adjective was caused by this retrieval term, or reference, or both.

Discussion

In this series of experiments, we asked whether listeners could use pragmatic contrast to resolve referential ambiguity and make inferences about a referent's category. In our first experiment, participants were able to use size adjectives contrastively to establish a novel word–referent mapping. Their contrastive inference goes beyond the implicit attention allocation shown in prior eye-tracking paradigms (Huang & Snedeker, 2008; Sedivy et al., 1999), determining explicit referent choice. This finding bolsters contrastive inference as a viable tool for referential disambiguation. In our second experiment, participants interpreted size and color adjectives contrastively to make inferences about a novel referent's category.

Participants failed, however, to use color adjectives contrastively in choosing referents.

What makes size different from color? One possibility is that the scalar nature of size supports a contrastive interpretation. We tested whether using relative color adjectives (e.g., bluer, greyer) or adjectives describing value (bright, dark) on saturated and desaturated stimuli would encourage the contrastive inference. We also tested whether adding a prosodic cue to contrast (e.g., "Find the *blue* dax") would encourage contrastive inference. Participants persisted in interpreting color non-contrastively, never consistently choosing the intended target over the lure. Though we do not claim that contrastive color inferences cannot be used to explicitly choose referents, it seems that a contrastive interpretation is difficult to elicit using color, while it emerges under similar conditions using size.

Another 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; see Chapter 10 of Gibson & Pearlmutter, 2011). This account explains well why color is not interpreted contrastively here, but fails to explain why presumably rare adjectives (bluer, bright) do not receive contrastive treatment in our task. 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.

Description is not limited to conveying contrast between present objects: it can also convey contrast with an object's category. In Experiment 3, we tested whether listeners inferred that a described feature of a novel object was atypical of its category, and how this inference was affected by the distractor objects present. We find that listeners infer atypicality from use of descriptors. However, they do not reserve this inference for cases of over-description alone: listeners inferred atypicality of a described feature even when the descriptor was necessary to establish reference. Listeners, then, seem not to rationally weigh the potential contrasts intended by the listener and trade off between them. Rather, participants' behavior in this task is better described by a coarse heuristic: use of description implies atypicality in relation to the category. Despite not being very sensitive to the referential context in their overt judgments, participants in our third experiment did show facilitation from contrast in processing. Directionally, participants advanced more quickly on trials in which a descriptor was used and was necessary to establish reference than on trials when a supplied descriptor was unnecessary. Overall, our results suggest that the atypicality inference is robust to the point of being difficult to suppress: it is not discounted, even when a descriptor is needed to distinguish between present objects. Participants do trend toward showing effects of the object context in their reaction times, but this processing effect does not consistently extend to overt judgments about the target's category.

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. 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). We plan to research further the development of these contrastive skills, as well as their potential as tools for extracting information from language and context.

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

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.

Engelhardt, P. E., Bailey, K. G. D., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, *54*(4), 554–573.

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

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.

Gibson, E. A., & Pearlmutter, N. J. (2011). *The Processing and Acquisition of Reference.* MIT Press.

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

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.

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.

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

Sedivy, J. C. (2002). Invoking Discourse-Based Contrast Sets and Resolving Syntactic Ambiguities. *Journal of Memory and Language*, *46*(2), 341–370.

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