Pick the Yellow ... Strawberry? Production Expectations Explain Variation in Contrastive Inference

Elisa Kreiss, Judith Degen

Stanford University

Pick the Yellow ... Strawberry? Production Expectations Explain Variation in Contrastive Inference

Introduction

Communication generally involves a *speaker* who sends out a communicative signal (e.g., speaks or signs a phrase, draws a picture, or points to an object), and a *listener* who perceives that signal and tries to understand its meaning. For example, if a speaker says "Pick up the yellow banana!", a listener can easily pick out the top left object in Figure 1A as the speaker's intended referent. A key feature of (human) communication is that listeners incorporate information as soon as it becomes available. For instance, when listeners have so far only heard "Pick up the yellow...", they already disregard the two non-yellow objects in the display because those are incompatible with the information received at that point (Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995).

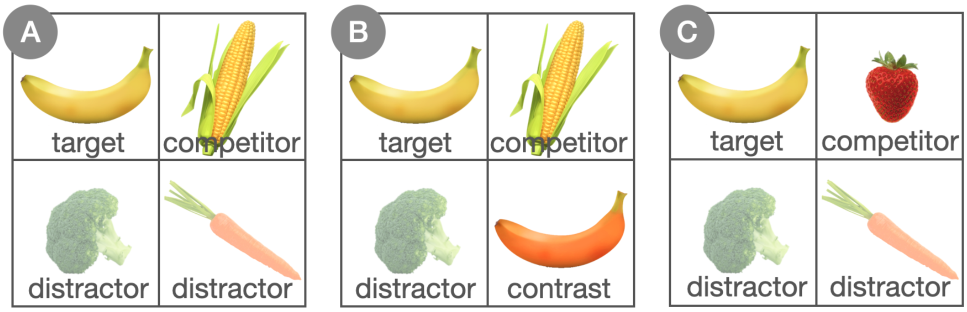


Figure 1. Three contexts of objects.

Moreover, listeners leverage contextual information to draw pragmatic inferences about the object the speaker is most likely referring to [ek: decide on which work to cite here]. While listeners don't show a preference for neither the yellow banana (the *target*) nor the yellow corncob (the *color competitor*) in Figure 1A, this changes as soon as a differently colored banana is introduced (as shown in Figure 1B). Again upon hearing "Pick up the yellow...", listeners now already show a preference for the banana over the corncob (Aparicio, Xiang, & Kennedy, 2016; Rubio-Fernandez, Terrasa, Shukla, & Jara-Ettinger, 2019; Sedivy, 2003; Sedivy et al., 1999). Since the original findings in the size adjective domain (Sedivy et al., 1999), this pattern of preference has been referred to as the *contrast effect* or *contrastive inference.*

It is important to distinguish three concepts here: *preference*, *inference*, and *contrastive inference*. *Preference* describes listeners' behavioral data that shows which objects are currently considered as most likely referents. This can be evidenced for example by listeners' increased fixations of an object (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Tanenhaus et al., 1995). An *inference* here describes any change in preference that is not due to changes in truth-conditions. For example, the target can already be uniquely identified after "Click on the yellow..." if the yellow corncob is replaced by a red strawberry (see Figure 1C). The resulting boost in target preference is therefore not due to an inference. In comparison, if we exchange one of the distractors for an orange banana (see Figure 1B), target preference increases even though target and competitor are still the only possible referents. This change in target preference is therefore an example of *inference.* We call it a *contrastive inference* because the target preference increases due to the introduction of a contrast (here the orange banana).

To explain how listeners draw this rapid contrastive inference, Sedivy (2003) proposes that it arises from listeners' pragmatic reasoning about the informativity of an utterance. Following the Maxim of Quantity (Grice, 1975), a cooperative speaker can be expected to not include more information than necessary into their utterance. In Figure 1B, listeners can therefore reason that the speaker could have referred to the corncob more simply as *the corncob.* According to this account, the use of a modifier to refer to the corncob would be considered *descriptive* and should not result in an inference. Instead, including the modifier is only necessary to distinguish the yellow from the orange banana. Here, the modifier is used in its contrastive function. Since the contrastive use of the modifier is considered more informative than the descriptive one, listeners interpret the modifier contrastively whenever possible. This reasoning predicts a preference for the target when a contrast is present (Figure 1B) and no preference when the contrast is absent (Figure 1A).

In order to make this prediction, this account needs to postulate three central assumptions. Firstly, the speaker needs to be considered cooperative for the inference to occur (Grodner & Sedivy, 2011). Secondly, adjectives can take on two separate roles (*descriptive* and *contrastive*). Thirdly, the use of an adjective in its contrastive function is more informative than when it's used descriptively. Finally, descriptive adjective use does not trigger any target inference. While previous literature appears to confirm that when a speaker is perceived as uncooperative, contrastive inference does not arise (Grodner & Sedivy, 2011; Ryskin, Kurumada, & Brown-Schmidt, 2019), none of the other assumptions has previously been tested.

Furthermore, some previous empirical results are at odds with this account of contrastive inference. In a series of studies investigating contrastive inference in the color adjective domain, the contrast effect appeared less stable than predicted under this account. Sedivy (2003) reports that the contrastive inference arises in contexts where the target object has a predictable color (such as the yellow banana in Figure 1) but not when it is replaced by an object with an unpredictable color like a cup, which comes in many colors. She shows that these objects differ in how likely a speaker is to produce the color modifier for the object in isolation: in the absence of a contrast, a yellow banana is usually called *the banana* while a yellow cup is often called *the yellow cup*, which Sedivy calls these objects' *default descriptions.* To account for this finding, she extends the contrastive inference account by another assumption: Only in cases where the modifier is not part of the default description is its observation surprising and the adjective's contrastive function is inferred. If its observation is not surprising, listeners simply accept the adjective in its descriptive function.

At the same time, Sedivy (2003) also finds that when using yellow banana-like items as targets and yellow cup-like items as competitors, listeners prefer the cup (the competitor) over the banana (the target) even if no contrast is present[1]. This is a puzzle for any account of contrastive inference so far.

In this paper, we propose a highly speaker-centric account of contrastive inference

[1] This surprising finding is mentioned in footnote 5 in Sedivy (2003).

which is based in the Rational Speech Act framework (Frank & Goodman, 2012; Goodman & Frank, 2016). In contrast to previous accounts, this model makes predictions only on the assumption of speaker cooperativity and listeners reasoning about speakers' most likely utterances. We will show that this leads to new predictions of contrastive inference that cannot be explained by other accounts without violating at least one of their underlying assumptions. Furthermore as to our knowledge, it is the first account of contrastive inference that makes *quantitative* predictions on how likely a listener is to prefer the target over the competitor when a contrast is absent and present. From that we can derive the boost in target preference that is due to the presence of a contrast. It allows us to abstract away from specific factors that can affect interpretation (such as adjective classes, descriptive vs. contrastive interpretation, or salience) by only considering production and prior probabilities.

We show that this model closely predicts listeners' inferences using color adjectives without assuming any additional linguistic and cognitive factors. Simply from production expectations alone, we can derive the patterns in interpretation, indicating that production and comprehension are closely linked. Motivated by the model predictions, we provide empirical evidence that not only the target, but also the competitor matters for eliciting contrastive inference. By changing whether a speaker is expected to produce an adjective for the target and competitor, we find high variation of target preference/inference within the color domain, which calls into question generalized statements about color-adjective inference patterns. Furthermore, our results suggest that listeners' interpretations are also affected by their prior beliefs about what the speaker is most likely referring to, making contrastive inference more malleable to previous exchange than is predicted by other accounts and discussed in the literature.

A Bayesian account of contrastive inference

The Rational Speech Act framework (Frank & Goodman, 2012; Goodman & Frank, 2016) is a probabilistic (and thus non-deterministic) Bayesian account of natural language which ascribes a central role to the speaker in pragmatic interpretation. The core idea of the model is that a listener and a speaker recursively reason about each other: A pragmatic listener L_1 wants to infer the meaning of an utterance u, as formulated by the pragmatic speaker S_1. Possible referents r are assigned a probability proportional to the probability that S_1 will produce u to convey r multiplied by the listener's prior belief in r $P(r)$, as defined by Bayes' Rule.[2]

$$P_{L_1}(r|u) \propto P_{S_1}(u|r) * P(r) \tag{1}$$

To simplify the following example, we will assume that listeners have a uniform prior $P(r)$ over all objects in the display[3]. Then the RSA model predicts a direct relationship between the production probabilities P_{S_1} and the listener's distribution over possible referents P_{L_1}.

While RSA has typically been applied to the analysis of full utterances, it can straightforwardly be extended to generate predictions at the sub-sentential level[4]. To generate RSA predictions for an incomplete referring expression such as *Click on the yellow...*, we take P_{S_1} to correspond to the contextual probability of color mention for each referent in the display. This corresponds to marginalizing over the probabilities of all continuations of the utterance (i.e., *Click on the yellow banana/corncob/lettuce/...!*). Let's investigate this account's qualitative predictions:

Consider the example contexts in Figure 2A and Figure 2B. Upon hearing the modifier *yellow*, the pragmatic listener P_{L_1} considers how likely a speaker is to include this modifier in their referring expression for each object in the display. Since only the target (yellow banana) and the competitor (corncob) are yellow, we assume that the

[2] The pragmatic speaker model and further recursive steps are spelled out in detail elsewhere (Goodman & Frank, 2016). Since we will elicit speaker probabilities empirically, we need not be concerned with the details of the speaker model.

[3] The simplifying assumption is justified by the results of Exp. 2.

[4] One exception is the Incremental Iterated Response Model of Pragmatics, which is also shown to qualitatively predict contrastive inference in general Cohn-Gordon, Goodman, and Potts (2019).

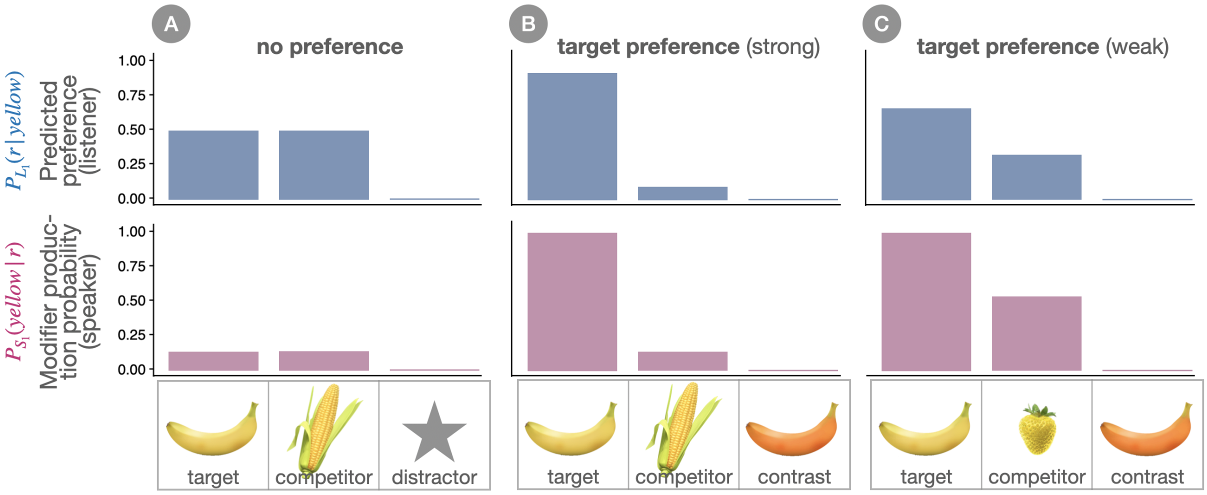


Figure 2. Qualitative RSA predictions.

production probabilities of *yellow* for the other objects in the display are 0. This only leaves the target and the competitor as potential referents.

Hypothetical modifier production probabilities for target and competitor are shown in the middle row of Figure 2. Assume that in the absence of a contrast object (Figure 2A), speakers are equally unlikely to include the color modifier when referring to the target banana (probability 0.1) and its color competitor, the corncob (0.1). Pragmatic listener predictions are obtained by renormalizing these probabilities, resulting in a target preference of 0.5, i.e., the pragmatic listener does not prefer one potential referent over the other.

Does RSA predict the target preference and therefore contrastive inference in context Figure 2B? Assuming that the presence of the contrasting orange banana does not affect the speaker's modifier production probability for the competitor corncob but does increase modifier production probability for the target banana to 0.9, renormalizing the production probabilities results in a target preference of 0.9. This way RSA can reproduce the classic contrastive inference pattern without assuming a distinction between descriptive and contrastive functions of modifiers.

Unlike previous accounts of contrastive inference, modifier production probabilities are expected to directly drive any target inference. The term *contrastive inference* itself therefore doesn't denote a distinct psycholinguistic process of inference but simply a change in inference strength that is due to the presence of a contrast. Crucially, this model predicts that *any* modifier use can result in an inference and interfere with the size of a contrastive inference.

Since the target preference depends on the modifier production probabilities of the target and the competitor, the competitor takes on a central role in contrastive inference predictions. This suggests that increasing the modifier production probabilities for the competitor should lead to a decrease in target preference. It has been established that speakers are more likely to include color modifiers in referring expressions for objects in isolation when they appear in an atypical rather than in a typical color (Degen, Hawkins, Graf, Kreiss, & Goodman, 2020; Rubio-Fernández, 2016; Westerbeek, Koolen, & Maes, 2015). Thus the atypical yellow strawberry in Figure 2C is more likely to elicit a color modifier than the typical corncob in Figure 2B. Assuming a modifier production probability of 0.6, this contrast-present context yields a much smaller increase in target preference compared to the contrast-absent context. In other words, the size of the target preference is predicted to be dependent on the choice of competitor in the contrast-present vs. contrast-absent conditions, keeping target typicality constant. This predicts that not only features of the target (Rubio-Fernandez et al., 2019; Sedivy, 2003), but also features of the competitor will affect the size of contrastive inference when evaluated against a contrast-absent-baseline of no preference.

We have shown that a speaker-centric model can predict the classic contrastive inference pattern without making assumptions about a contrastive function which is inherent to the adjective. Instead, contrastive inference is derived the same way as an inference based on typicality is – by reasoning about speaker's modifier production probabilities. Since this model doesn't assume a separate process associated with drawing contrastive vs. other types of inferences, it makes new predictions on when contrastive inference should occur which are incompatible with predictions made by contrastive-function accounts. In this work, we explicitly manipulate the modifier production probability of the target and competitor by varying the typicality of the color they appear in. In order to generate RSA predictions, we elicited the modifier production probabilities (i.e., an estimate of $P_{S_1}(u|r)$) in a free production interactive reference game (Section). This allowed us to generate pragmatic listener probabilities for each display. We compare these predictions with predictions contrastive-function-based models of contrastive inference make. Those predictions can then be evaluated against the inferences listeners draw. The results clearly point to an expectation-based account of inference more generally, and contrastive inference in particular as well.

Obtaining quantitative model predictions: production experiment

The goal of this experiment was to obtain color modifier production probabilities for the items in the displays ultimately used in the contrastive inference experiment (Exp. 2). In particular, we elicited production probabilities for those items that functioned as targets and competitors in Exp. 2.[5] Probabilities were elicited in a free production interactive reference game. We expected modifier production probability to be higher for atypical objects and in the presence of a contrast. For instance, we expected speakers to call a yellow banana simply *the banana*, but an orange banana *the orange banana*. We treated the elicited modifier production probabilities as the pragmatic speaker probabilities in the RSA model evaluation.

Method

Participants. We recruited 282 participants ([ek: XXX] female, mean age: [ek: XXX]) over Amazon's Mechanical Turk, who were randomly paired to form director-matcher dyads (i.e., 141 pairs in total). Each participant was paid $2.30 (approximately $11-$14/hr with a median completion time of [ek: XXX]). We restricted participation to workers with US-based IP addresses and a previous work approval rate of at least 97%.

Materials. Each context included four objects, as displayed in Figure 4. The items used were carefully normed for color-diagnosticity Tanaka and Presnell (1999),

[5] We assumed that the production probability of the relevant color modifier was close to 0 for the remaining distractor objects in the display and did not elicit these explicitly.

typicality, and nameability. The pool of items consisted of 10 types (e.g., broccoli), each of which could occur in a typical (green broccoli) and atypical color (red broccoli). The colors were counterbalanced such that each color occurred on two objects as a typical instance and on two objects as an atypical instance. All items are displayed in Figure 3 together with their empirically elicited typicality rating. More details on the norming studies can be found in Appendix A.

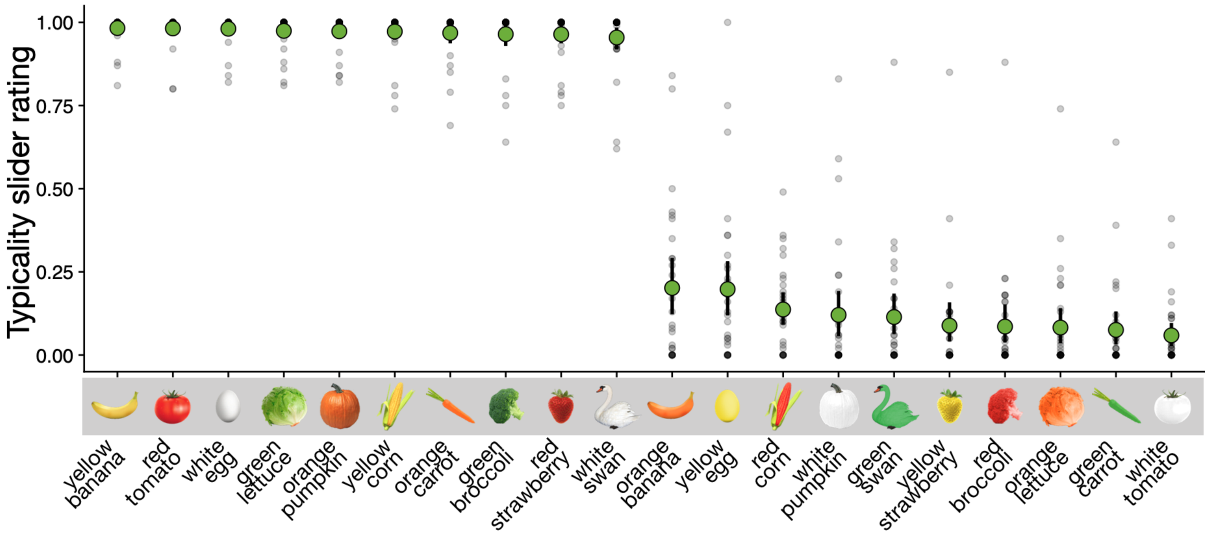
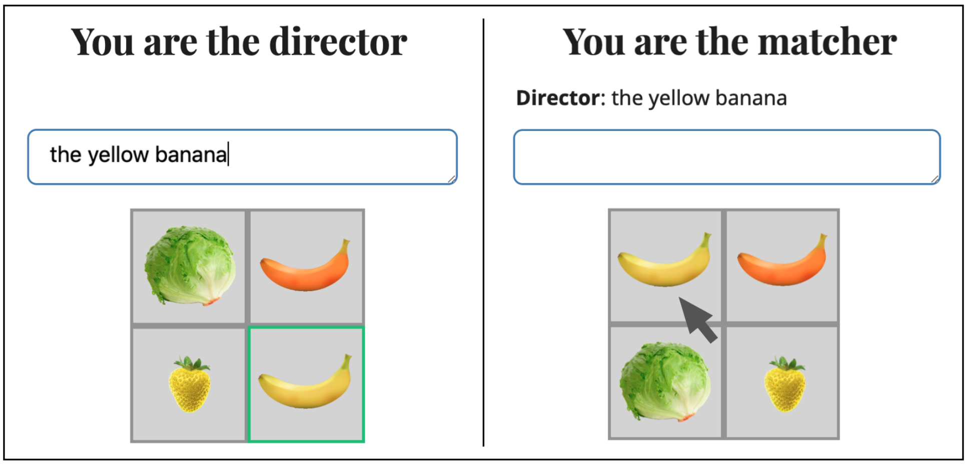


Figure 3. Final set of stimuli, ordered by typicality. Each object occurs in a typical and atypical color.

Procedure. The speaker-listener pairs could communicate freely through a real-time multi-player interface similar to Hawkins (2015). The speaker was instructed to communicate a target object out of a four-object context to the listener. The target could be identified by a green border surrounding it. The speaker and the listener saw the same set of objects but in a randomized order to avoid trivial position-based references such as “the left one”. After the listener clicked on the presumed target, both the speaker and listener received feedback about whether the correct object had been selected.

On critical trials, participants saw the critical displays from Exp. 2. The object to be communicated could be either the object that functioned as the target or the object that functioned as the competitor in that display in Exp. 2, as exemplified in Figure 4. We continue to refer to ‘target’ and ‘competitor’ in the reporting of this experiment,



You are the director

the yellow banana

You are the matcher

Director: the yellow banana

Figure 4. Example display for the interactive reference game (Exp. 1). Both, the speaker (here *Director*) and listener (*Matcher*) see the same four objects but in a scrambled order. Additionally, the speaker sees a green border around one of the objects, marking the intended target which the listener needs to select.

terms which refer to the function of the object to be communicated in Exp. 2. Contexts varied in the typicality of the target and the competitor and the presence of a contrast, resulting in eight conditions. Participants saw each context exactly once. Throughout the experiment, half of the critical trials required the speaker to communicate the 'target' and in the other half the 'competitor'.

In contexts where the contrast was absent, the distinction between target and competitor was meaningless and thus one of the color competitor objects was arbitrarily coded as the target and the other as the competitor. Fillers were eight randomly created contexts where the 'contrast' or the 'distractor' from [ek: context-ref] was the object to be communicated. Overall, each dyad saw 60 contexts (32 critical trials) in randomized order.

Data pre-processing and exclusion. [ek: update numbers] Exclusions were performed on the 141 speakers, since they provided the utterances. Participants were excluded when they participated multiple times in the experiment (1 participant; 139 pairs remaining) and when they did not use a noun from the display in at least half of the cases (27 participants; 112 pairs remaining). These participants clearly misunderstood the task, using expressions such as *yellow monkey* instead of *yellow banana*, or *should be yellow, must have teeth to eat* for *corn.* All speakers indicated that their native language was English.

Results

We excluded two dyads because of multiple participation and 27 dyads for primarily using playful descriptions, e.g., *should be yellow, must have teeth to eat* for the *red corn* object, which left 112 dyads for the analysis.

Figure 5 shows the proportion of color modifier mentions for the target and competitor in each condition. We conducted a Bayesian mixed effects logistic regression predicting color mention for each item from centered fixed effects of contrast presence, target typicality, and competitor typicality, as well as random by-participant intercepts (the most complex random effects structure that allowed the model to converge).

There was strong evidence of contrast presence ($E = 5.25,\ CI = [4.82, 5.69]$), such that when a contrast to the object was present (e.g., another banana, see target proportions in the upper row in Figure 5), participants were more likely to mention the color modifier than in the absence of a contrast (see target proportions in the lower row in Figure 5 and competitor proportions overall). This was especially true when the object was atypical[6]. There was also strong evidence for the object's typicality ($E = 2.82,\ CI = [2.52, 3.12]$), such that participants were more likely to include a color modifier when referring to an atypical object than a typical one.

[6] A full interaction model did not converge because color was *always* mentioned in the contrast-present condition with atypical targets, which did not allow the model to generate estimates for interactions involving these conditions. We did not find evidence for any other interactions.

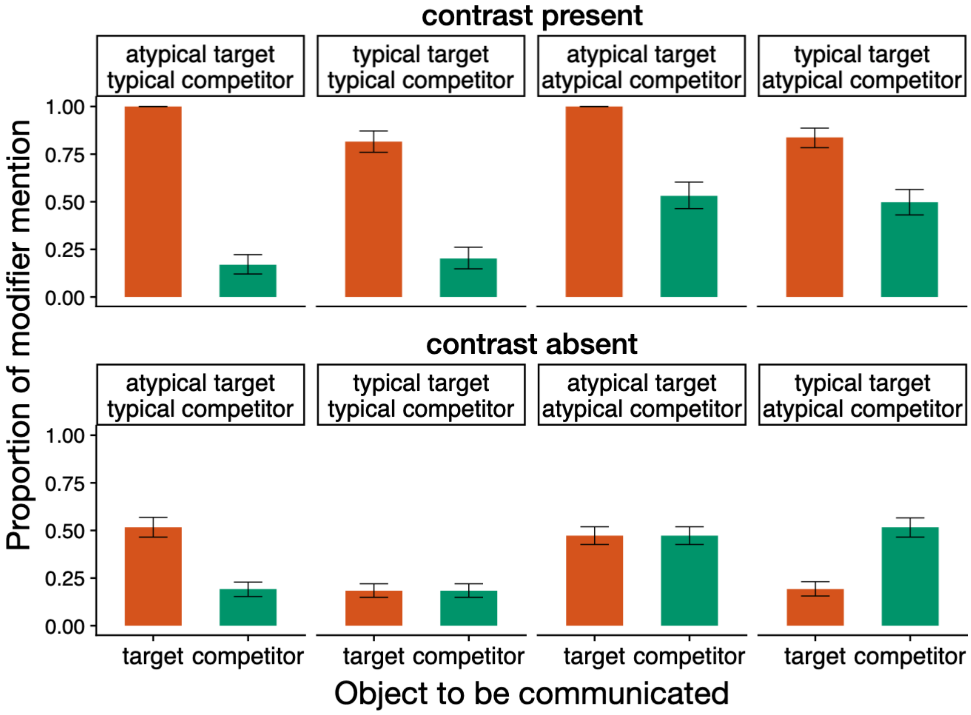


Figure 5. Proportion of modifier mentions in each condition for objects that functioned as target and competitor in Exp. 2. Error bars indicate 95% bootstrapped confidence intervals.

The results of this production experiment show that the probability of a speaker's modifier use is modulated by an object's color typicality, replicating previous results Westerbeek et al. (2015). [ek: Do other contrastive inference accounts really assume that though?:] The results also confirm the assumption made by many contrastive inference studies that speakers are more likely to produce the color modifier in the presence of a contrast Aparicio, Kennedy, and Xiang (2018); Grodner and Sedivy (2011); Sedivy et al. (1999), though this probability is modulated by the typicality of the object.

Model predictions

The production data in Figure 5 were used to obtain RSA model predictions of target preference and we compare these with four variants of the contrastive-function account. Since the assumptions around the contrastive-function account are generally not explicitly formulated, we decided on four variants which represent the foundation of the account (i.e., *vanilla* and *default*), or are relevant possible extensions (i.e., *descriptive inference* extensions). All variants assume that contrastive functions of adjectives are separate from their descriptive form and that the contrastive use is the more informative choice. Consequently, all of these accounts predict all-or-none elicitations of (contrastive) inference: either the target is preferred or there is no preference. While recent work on contrastive inference has started to investigate variation in contrastive inference data (Aparicio et al., 2018; Rubio-Fernandez et al., 2019), this has so far only addressed variation between adjective classes and has been explained by other aspects than the size of inference itself (e.g., through visual salience of the property).

Figure 6 displays the predictions contrastive-function accounts and an RSA model of contrastive inference make when varying target typicality, competitor typicality, and contrast presence. It shows the predictions on how likely the listener considers the target object to be the intended referent after observing the modifier (e.g., “the yellow...”). If neither target nor competitor are preferred, the predicted target consideration is 0.5. If the target is preferred over the competitor, this value increases (and vice versa).

Firstly, we discuss the four variants of the contrastive-function account. The **vanilla** version simply predicts that an inference only occurs when a contrast is present. This is independent of any properties of the target or the competitor. Consequently, there is no preference when the contrast is absent and a target preference when the contrast is present.

Due to the observation that contrastive inference only appears to arise with objects that don’t have color in their *default description*, such as bananas, Sedivy (2003) proposed the **default** account. It introduces the restriction to the vanilla account that the contrastive interpretation is only activated if the use of color is surprising given the default description. Just like the vanilla account, it uniformly predicts no preference when the contrast is absent. But it only predicts a contrastive interpretation when the target is typical.

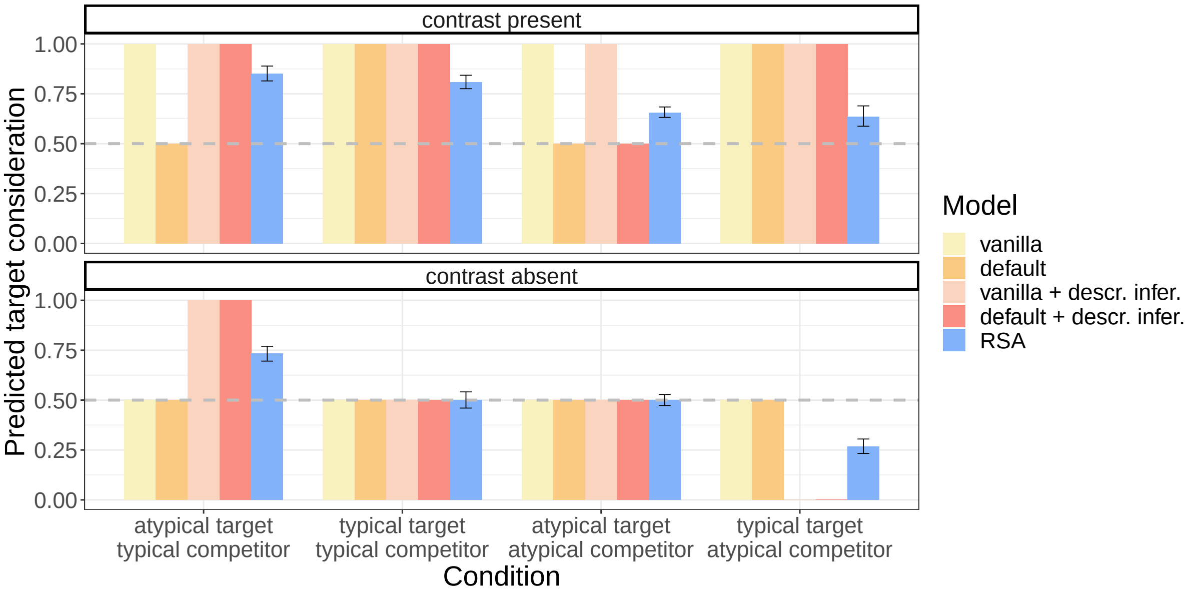


Figure 6. Predictions of contrastive function accounts vs. RSA account. Plotted are the probabilities how likely the target object is considered as the intended referent upon observing "the yellow…". 1 -(minus) target preference/consideration corresponds to competitor consideration which is not displayed here. [ek: neither blackandwhite nor redgreenblind friendly]

We also consider a variation of the vanilla and default account (displayed in light and darker red respectively in Figure 6). In these models, we allow for **descriptive inference** such that if the speaker uses color in their reference, listeners show a preference for the atypical over the typically colored object. To our knowledge, this version of a contrastive function account has not been discussed in the literature but we believe it's a natural extension. Parallel to the assumption that an adjective's contrastive use is more informative than its descriptive use, the use of the descriptive could be considered more informative than the bare form. Therefore, those models predict a preference for atypical objects when observing the modifier. However this inference only affects the contrast-absent conditions, since it will be overruled by the contrastive interpretation when a contrast is present.

Let's turn to the RSA model predictions. The first most salient difference is the fact that RSA predictions are non-deterministic. While contrastive-function accounts assume a deterministic process on whether an inference arises or not, RSA predicts that listeners can be more or less certain about the speaker's most likely intended referent which is why the strength of target inference varies dependent on the condition. In the contrast absent conditions, RSA predicts that listeners prefer the atypical item over the typical one, since speakers are more likely to use the modifier for atypically colored objects. The direction of the preference is most closely related to the *descriptive inference* accounts, since they are also based on this production asymmetry. However those parallels fall apart in the contrast present conditions. Most crucially, the predicted target preference when the competitor is atypical (i.e., in the two right-most conditions) is predicted to be lower than when the competitor is typical. None of the contrastive-function accounts predict a variation in inference (strength) or preference dependent on the competitor typicality.

This is the point where the differentiation between *preference* and *contrastive inference* becomes crucial. Consider the right-most condition where the target is typical (e.g., a yellow banana) and the competitor is atypical (e.g., a yellow strawberry). When the contrast (i.e., another banana) is present, the RSA model predicts only a small preference for the target over the competitor, since the speaker could have used the modifier for differentiation or due to the competitor's atypicality. However, there is a strong predicted contrastive inference since the target is predicted to be *dis*preferred in the contrast-absent condition.

Reversely, we can consider the left-most condition where the target is atypical and the competitor is typical. Here the model predicts a target preference already in the contrast-absent condition and almost no target boost in the contrast-present condition. This suggests that listeners don't draw a contrastive inference in this condition (that exceeds the atypicality inference), even though there is a preference for the target when a contrast is present.

These two cases show the relevance of separating the behavioral pattern of preference in the contrast present condition from the difference in preference matched with the contrast absent condition.

Overall, contrastive-function accounts of contrastive inference make qualitatively and quantitatively different predictions. Most crucially, the RSA model of target preference (1) makes non-deterministic predictions, (2) predicts that listeners draw target inferences even when adjectives are used non-contrastively, and (3) predicts that the nature of the competitor affects target preference.

We conducted a comprehension experiment where we elicited listeners' beliefs about the most likely referent given an utterance such as "the yellow..." to investigate whether contrastive inference is best explained by assuming contrastive functions of adjectives or a production-centric view of inference in general.

Comprehension experiment

To investigate which object listeners consider to be the most likely referent after observing the color adjective, we conducted an incremental decision task Qing, Lassiter, and Degen (2018). This is an offline task that allows for eliciting participants' belief distributions at multiple points in the unfolding referring expression.

Method

Methodologically, contrastive inference behavior has generally been investigated in eyetracking experiments (Aparicio et al., 2016; Grodner & Sedivy, 2011; Heller, Grodner, & Tanenhaus, 2008; Rubio-Fernandez et al., 2019; Ryskin et al., 2019; Sedivy et al., 1999). A contrastive effect is considered to be present if the target preference arises earlier in the contrast condition than in the no-contrast condition. In this paradigm, target preference is approximated by the number of looks to the target (over looks to the competitor). This linking hypothesis has recently been called into question (Qing et al., 2018). While explicit belief about the intended referent appears to be correlated with an increased proportion of looks, this correlation decreases when the adjective is unexpected (Qing et al., 2018) [ek: more examples!].

Recently, those offline measures of preference have gained wider popularity (Alsop, Stranahan, & Davidson, 2018; Kronmüller, Morisseau, & Noveck, 2014; Qing et al., 2018). In an *incremental decision task*, the utterance is revealed word-by-word to the listener, who makes an explicit guess about the most likely referent at each step. Qing et al. therefore argue that their method offers a more direct window into listeners' beliefs. In this paradigm, contrastive inference is the average difference in target preference between the contrast and no-contrast condition in the adjective window. In contrast to the eyetracking methodology, the strength of the inference can easily be translated to the increase in target preference.

Due to the more intuitive interpretability and easier data collection procedure, we investigate target preference patterns using the incremental decision task. This paper provides further evidence that the incremental decision task can be used to elicit contrastive inference. [ek: Can eyetracking be used as a method for prior manipulation experiments (i.e., where the preadjective baseline is non-uniform)?]

Participants. We recruited 239 participants over Amazon's Mechanical Turk, 121 of which saw atypical color competitors and 118 saw typical color competitors in the critical trials[7]. Each of them were paid $1.80 for their participation (10-16/hr). We restricted participation to workers with IP addresses in the US and an approval rate of previous work above 97%. 27 participants were excluded because they indicated that they did the experiment incorrectly, English was not their native language, or they gave more than 20% erroneous responses[8]. 211 participants remain, 108 of which were in the atypical competitor and 103 were in the typical competitor condition.

Materials. The critical displays were identical to the critical displays in Exp. 1.

Procedure. This experiment was a one-player comprehension-only adaptation of the production study described above and was implemented as an incremental decision task Qing et al. (2018): Participants read sentences of the form "Click on the yellow banana", which contained a referring expression, and their task was to select the

[7] The experiment was preregistered on `https://osf.io/27dn8`. Originally, we recruited 80 participants and then ran a follow-up with 140 more to get enough data for the evaluation of the RSA model. The results from the first 80 participants do not differ from the full data set, which is why we present them collapsed.

[8] An erroneous response is defined as a selection of a non-target object after observing the fully disambiguating noun.

target in the display. Crucially, the sentence was only gradually revealed. Participants made a selection at each of three time points: (1) before receiving any information about the referent (i.e, after observing "Click on the", *prior window*), (2) after observing the adjective ("Click on the yellow"), *adjective window*, and (3) after observing the full referring expression with the disambiguating noun ("Click on the yellow banana"), *noun window*.

To center the position of the mouse after each selection, a button appeared in the center of the grid which had to be clicked to reveal the next word or to advance to the next trial.

[ek: The data from before adjective onset we use to estimate the prior distribution. For cleaner effects, we don't use the data from this experiment to investigate post-adj selections due to a strong non-switching bias. The results presented here are from the post-adj-only experiments]

Target typicality and contrast presence were within-participant manipulations, competitor typicality was a between-participants manipulation [9] . All critical trials used color modified referring expressions. Filler trials were included that primarily used unmodified utterances and referred to one of the other three items in the display to avoid learning effects. Participants completed 55 trials (20 critical) in random order. To minimize the risk that the speaker was perceived as pragmatically uncooperative Grodner and Sedivy (2011); Pogue, Kurumada, and Tanenhaus (2016); Ryskin et al. (2019), trials with modified utterances that referred to a typical object with no contrast only appeared after the 15th trial. To familiarize participants with the task, they first

[9] The complexity of the 2x2x2 design and considerations of power required that either the number of trials per participant be high or one manipulation be between-participants. We decided for a smaller number of trials to minimize the probability of strategic responses or response fatigue developing over the course of the experiment. Contrast presence and target typicality could not be manipulated between-participants since these regularities are easily detectable by a participant within an experiment. Between-participants manipulations are considered more conservative Charness, Gneezy, and Kuhn (2012) and random by-participant intercepts and slopes were included in the analyses to account for random by-participant variability.

completed four practice trials in the director role.

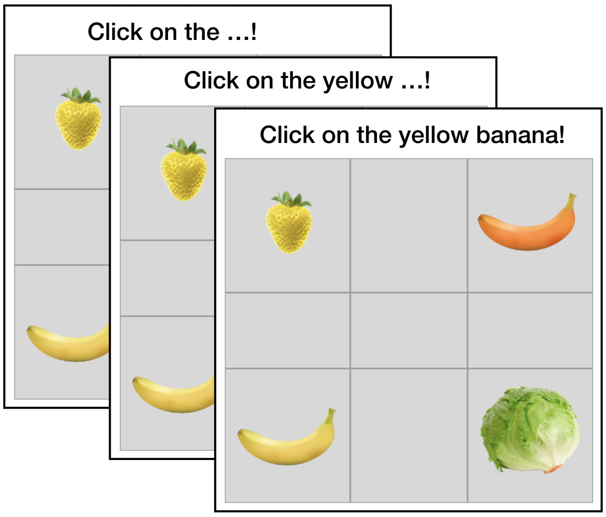


Figure 7. Design of the incremental decision task. The referring expression was placed above the grid and revealed gradually. After each new word participants made a selection indicating their best guess about the intended target.

Data pre-processing and exclusion. We excluded participants who did the Hit multiple times (1), who indicated that they did the Hit incorrectly or were confused (13), who indicated that they had a native language other than English (6), and who gave more then 20% erroneous responses (7). An incorrect response is defined as a click to a non-target object after observing the fully disambiguating noun, i.e., participants are excluded who selected the wrong final object more than 11 times. Overall, we excluded 27 people, which is 11% of the subjects. 211 participants remain, 108 of which were in the atypical competitor and 103 were in the typical competitor condition.

Results

[ek: modelcompr-results] shows the proportion of target and competitor selections in the adjective window (lighter colors) alongside the RSA model predictions derived from the Exp. 1 production probabilities (darker colors), grouped by condition.[10] We conducted a Bayesian mixed effects logistic regression on adjective window choices,

[10] Neither of the other two objects in the display was chosen after observing the adjective.

predicting the log odds of target over competitor selections from centered fixed effects of contrast presence, target typicality, competitor typicality, and their interactions, prior window selection, as well as the maximal random effects structure that allowed the model to converge[11].

There was strong evidence for an effect of contrast presence $(E = 0.34,\ CI = [0.13, 0.53])$, such that when there was a contrast object (top panels), there was a general preference for target over competitor selections, replicating the standard contrastive inference effect. This preference was largest when the target was atypical and the competitor was typical and disappeared when the target was typical and the competitor was atypical, following the qualitative predictions discussed in the modeling section above and exemplified in [ek: example-context]. There was also strong evidence for an effect of competitor typicality $(E = -0.54,\ CI = [-0.90, -0.17])$, such that when the competitor was atypical, target selections decreased, which is again in line with our predictions.

Although object selections in the prior window were approximately at chance, there was strong evidence that it affected participants' specific selections of their adjective window choices $(E = 1.46,\ CI = [1.29, 1.63])$. These results suggest that when participants' prior selection is congruent with the newly revealed adjectival information, they stick with their previous choice.

Overall, these results suggest that the color typicality of not just the target, but of competitor objects in the display, too, affects the inferences listeners draw about the intended referent. An atypical competitor alone can promote the competitor over the target when the contrast is absent and can even make the target preference disappear when a contrast is present.

If one quantifies contrastive inference as an increased target preference in the adjective window in the contrast-present condition compared to its item-matched contrast-absent condition, the contrastive inferences is small or even non-existent when

[11] Random effects: $(1 + \text{contrast} * \text{target\_typicality}|\text{participant}) + (1 + \text{contrast} * \text{competitor\_typicality}|\text{target}) + (1 + \text{contrast} * \text{target\_typicality}|\text{competitor})$

the target is atypical and the competitor typical (left column of [ek: modelcompr-results]). This may explain why contrastive inferences did not occur with target items of unpredictable colors Sedivy (2003). However, even though those items have been reported to have a higher modifier production probability in isolation Sedivy (2003), future work still needs to establish how those objects of unpredictable colors relate to (a)typically colored objects.

Model predictions (with prior)

Comprehension experiment – prior manipulation

and what other models have to say about that

Discussion

[ek: speaker specific adaptation (Pogue et al., 2016)] [ek: mention that done in English, prenominal, should extend to other languages (connect to that literature)] [ek: given our methodological restrictions (binary choice): we cannot differentiate between binary threshold variance (draw inference or not), or an inherent probabilistic process]

Conclusion

Acknowledgments

QP1 committee, ALPS lab, CAMP, CUNY, CogSci, MagPie team (reference game setup); Previous versions of this work have been presented at CogSci, CAMP and CUNY

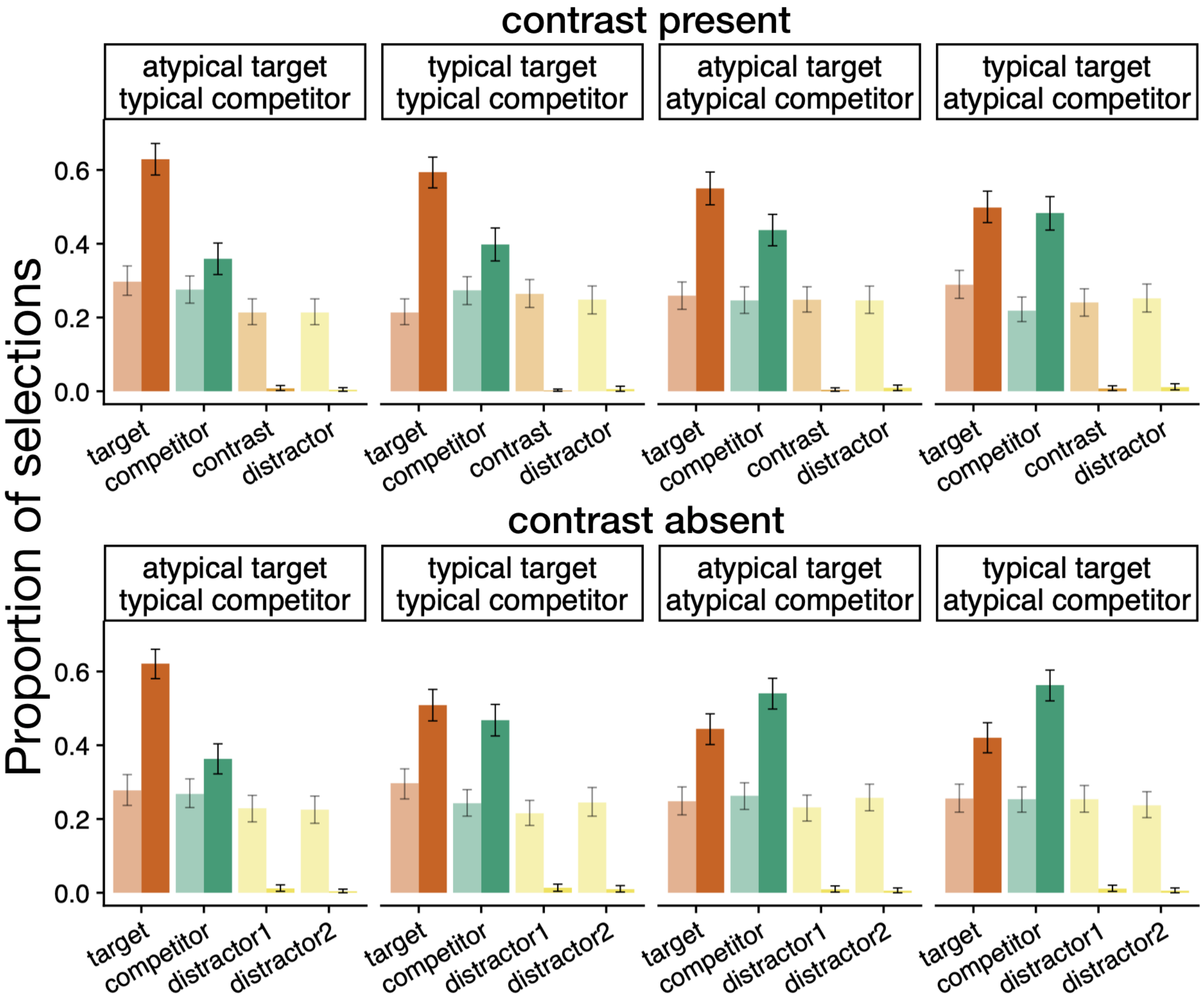


Figure 8. Results for the comprehension study, showing the proportion of selections for each item in the display and each condition. The bars in lighter colors indicate the selections before, the darker bars are the selections after the adjective was observed. Error bars are 95% bootstrapped confidence intervals.

Appendix

Norming

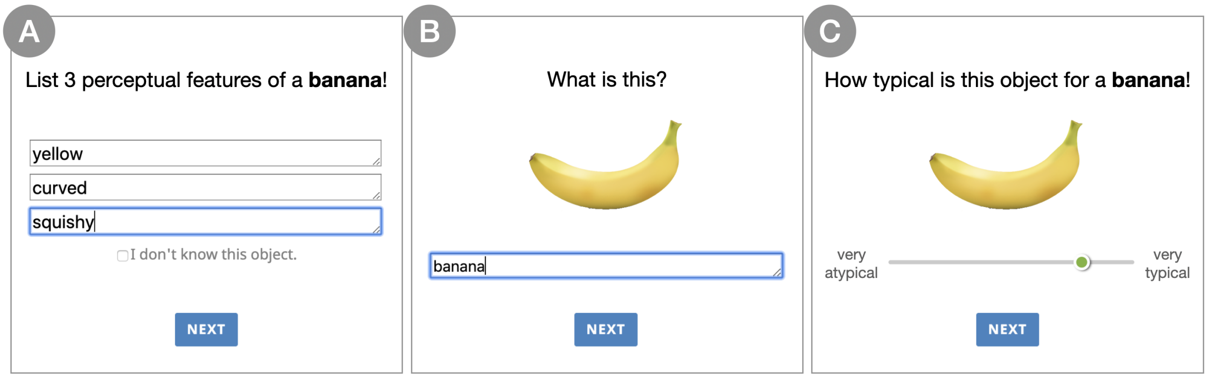
The items used in the experiment were carefully selected according to the results of four norming studies. First of all, the objects needed to be color-diagnostic (Section [ek: coldiagnorming]), since those objects show the highest difference in color modifier use dependent on their color typicality (Sedivy, 2003; Tanaka & Presnell, 1999; Westerbeek et al., 2015), our central manipulation. In addition, the items needed to be shape-diagnostic, such that they would still be recognizable when presented in a non-prototypical color. Plums, oranges, limes and lemons for example are objects with low shape-diagnosticity and are therefore hard to identify when presented in an atypical color (Section [ek: freeprodnorming]). Furthermore the items needed to be easily recognizable and nameable (Section [ek: nameabilitynorming] and [ek: freeprodnorming]). Previous literature suggests that unexpected utterances and labels might be a confound in eyetracking experiments (Qing et al., 2018). To make the items further suitable for eyetracking studies, target and color competitor were never cohort competitors of each other (e.g., Cole and Jakimik (1980), Marslen-Wilson (1984)), i.e., they were always distinguishable at noun onset.

Furthermore, to our knowledge, previous experiments which manipulated the color typicality of objects did not counterbalance the colors used for typical and atypical instances. Certain colors (e.g., blue) primarily occurred as an atypical instance while other colors (e.g., green) occurred as a typical one. However, color hues vary in their a priori preference and elicit different emotional responses (see Elliot and Maier (2014); Palmer, Schloss, and Sammartino (2013) for reviews on color preference and psychology). The color hue imbalance for typical and atypical objects might therefore be a non-negligible confound to the typicality effect. Each color in this data set occurs twice as a typical instance and twice as an atypical instance to counteract this potential confound.

Finally, the typical and atypical instance of each object was normed to ensure that the color manipulation of the images shows the desired difference in typicality ratings

(Section [ek: typicalitynorming]).

Since half of the conditions require the target and color competitor both to be (a)typical, we needed two (a)typical instances for each color. Motivated through this experimental design, the final set of stimuli consists of ten color diagnostic objects (each of them in a typical and atypical instantiation), evenly distributed over five colors. To determine the most suitable items, our initial set of potential stimuli comprised six colors (green, orange, pink, red, white, yellow), each with at least four presumably typical color-diagnostic instances (25 items in total).



List 3 perceptual features of a **banana**!

yellow

curved

squishy

□ I don't know this object.

What is this?

banana

How typical is this object for a **banana**!

Figure A1. Norming studies. [ek: possibly remove button]

In the end, the final set of stimuli comprises 10 objects, each occurring in a typical and atypical color. Items can occur in the colors yellow, red, green, orange and white. Each color occurs twice as typical and twice as atypical. The objects are *banana, broccoli, carrot, corn, egg, lettuce, pumpkin, strawberry, swan,* and *tomato*. The full set of stimuli is displayed in Figure 3.

References

Alsop, A., Stranahan, E., & Davidson, K. (2018). Testing contrastive inferences from suprasegmental features using offline measures. *Proceedings of the Linguistic Society of America*, *3*(1), 71–1.

Aparicio, H., Kennedy, C., & Xiang, M. (2018). Perceived informativity and referential effects of contrast in adjectivally modified nps. In *The semantics of gradability, vagueness, and scale structure* (pp. 199–220). Springer.

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

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1–8.

Cohn-Gordon, R., Goodman, N., & Potts, C. (2019). An incremental iterated response model of pragmatics. *Proceedings of the Society for Computation in Linguistics*, *2*(1), 81–90.

Cole, R. A., & Jakimik, J. (1980). A model of speech perception. *Perception and production of fluent speech*, 133–163.

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*.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of psycholinguistic research*, *24*(6), 409–436.

Elliot, A. J., & Maier, M. A. (2014). Color psychology: Effects of perceiving color on psychological functioning in humans. *Annual review of psychology*, *65*, 95–120.

Frank, M. C., & Goodman, N. D. (2012, May). Predicting pragmatic reasoning in language games. *SCIENCE*, *336*, 1.

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.

Grodner, D., & Sedivy, J. C. (2011). 10 the effect of speaker-specific information on pragmatic inferences. In *The processing and acquisition of reference* (Vol. 2327, pp. 239–272). MIT Press.

Hawkins, R. X. D. (2015). Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, *47*(4), 966-976.

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*(3), 831–836.

Kronmüller, E., Morisseau, T., & Noveck, I. A. (2014). Show me the pragmatic contribution: a developmental investigation of contrastive inference. *Journal of child language*, *41*(5), 985–1014.

Marslen-Wilson, W. D. (1984). Function and process in spoken word recognition: A tutorial review. In *Attention and performance: Control of language processes* (pp. 125–150). Erlbaum.

Palmer, S. E., Schloss, K. B., & Sammartino, J. (2013). Visual aesthetics and human preference. *Annual review of psychology*, *64*, 77–107.

Pogue, A., Kurumada, C., & Tanenhaus, M. K. (2016). Talker-specific generalization of pragmatic inferences based on under-and over-informative prenominal adjective use. *Frontiers in psychology*, *6*, 2035.

Qing, C., Lassiter, D., & Degen, J. (2018). What do eye movements in the visual world reflect? A case study from adjectives. In *Proceedings of the 40th annual conference of the cognitive science society.*

Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in psychology*, *7*.

Rubio-Fernandez, P., Terrasa, H. A., Shukla, V., & Jara-Ettinger, J. (2019). Contrastive inferences are sensitive to informativity expectations, adjective semantics and visual salience.

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

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of psycholinguistic research*, *32*(1), 3–23.

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

Tanaka, J. W., & Presnell, L. M. (1999). Color diagnosticity in object recognition. *Perception & Psychophysics*, *61*(6), 1140–1153.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Westerbeek, H., Koolen, R., & Maes, A. (2015, Jul). Stored object knowledge and the production of referring expressions: the case of color typicality. *Front. Psychol.*, *6*. Retrieved from `http://dx.doi.org/10.3389/fpsyg.2015.00935` doi: 10.3389/fpsyg.2015.00935