Perceptual Difficulty Differences Predict Asymmetry in Overmodification with Color and Material Adjectives

Leyla Kursat & Judith Degen*

Abstract. [lk: this is taken from the abstract - write about exp3]. When referring to objects, speakers are often more specific than they need to be for establishing unique reference. Adjectival overspecification patterns are not random, but structured: color adjectives are produced redundantly more often than size or material adjectives; color adjectives are more likely to be produced redundantly with increasing scene variation; and adjectives are more likely to be produced redundantly, the more atypical the property denoted by the adjective is for the object under discussion. The only current computational model of referring expression production that accounts jointly for all of these patterns is couched within the Rational Speech Act framework and assumes that adjectives differ in how noisy, and consequently, how useful they are for the purpose of establishing reference. One hypothesis about the nature of this noise is that it reflects the perceptual difficulty of establishing whether the property denoted by the adjective holds of the contextually relevant objects. Here, we take a first step towards testing the prediction that systematic differences in the overmodification patterns observed for color and material adjectives can be explained by a difference in perceptual difficulty of establishing whether objects are of a particular color or material. In Exp.1, we norm the perceptual difficulty associated with establishing whether an object exhibits a color or material and select objects with highest and lowest perceptual difficulty for testing in Exp. 2. In Exp. 2, we test in a reference game whether adjectives that denote more perceptually difficult properties indeed are less frequently produced redundantly.

Keywords. reference; perception; overinformativeness; experimental pragmatics

1. Introduction. When referring to objects, speakers aim to be sufficiently informative in choosing which features of the object to include in their utterances. However, they are often more specific than they need to be for establishing unique reference. Recent research has identified systematic differences in these adjectival overspecification patterns, but the question remains open why such patterns emerge.

One way in which we observe structure in the production of overinformative referring expression is through the asymmetry in the redundant use of color, size and material adjectives. When size or material is the sufficient property to single out the intended reference, participants routinely include color adjectives in their utterances. However, in contexts where color is sufficient for unique reference, speaker's don't tend to mention size or material redundantly (**Pechmann1989, Sedivy2003, GattEtAl2011, RubioFernandez2017, DegenEtAl2020**). Moreover, speaker's knowledge of the typicality of the properties of objects and the features of the context interact with these asymmetries. Color adjectives are more likely to be produced redundantly with increasing scene variation (**DegenEtAl2020, DaviesKatsos2013, KoolenEtAl2013**) and adjectives are more likely the be produced redundantly, the more atypical the

[*] Authors: Leyla Kursat, Stanford University (lkursat@stanford.edu) & Judith Degen, Stanford University (jdegen@stanford.edu).

property denoted by the adjective is for the object under discussion (**DegenEtAl2020, WesterbeekEtAl2015, Mitchell2013**).

The computational model of referring expression production developed by **DegenEtAl2020** accounts jointly for all of these patterns and proposes a unified quantitative account for their emergence. Couched within the Rational Speech Act framework (**GoodmanFrank2016**), this model treats speakers and listeners as agents recursively reasoning about each other's mental states to communicate. The simple RSA model assumes that objects have deterministic lexical meanings and that speakers choose utterances that maximize informativeness with respect to those meanings. This model doesn't generate overinformative referring expressions mainly because it calculates the informativeness (and cost) of mentioning the redundant property to be equal to the informativeness of mentioning the alternative (only mentioning the sufficient property). **DegenEtAl2020** extend this model by focusing on the calculation of informativeness and relaxing the Boolean semantics to non-deterministic continuous semantics that return real values between 0 and 1. By allowing utterances to be informative about objects to varying degrees, this continuos semantics assumes that adjectives differ in how noisy, and consequently, how useful they are for the purpose of establishing reference.

This assumption raises an important question regarding the nature of the adjectival noise. Many explanations are offered in the literature for the source of the overinformative patterns including lexical, visual, communicative and semantic ones and [lk: need to say something about how they could also account for the noise?].

[lk: connect these accounts in one cohesive paragraph!!!]

- **Sedivy2003** proposes a pragmatic account that is based on listeners' expectations of informativity. According to her, highly predictable properties of objects are not encoded as part of their "default descriptions" and therefore, their use triggers a contrastive inference and provides referential disambiguation. Although this account explains the typicality effects that have been shown to modulate overmodification patterns, it fails to account for the role of contextual information in referential communication.

- Recent work by **ViethenEtAl2017** highlights the context-dependency of overmodification patterns by showing that a decrease in the color contrast between the objects in the context reduces the use of color adjectives in referring expressions.

- The semantic account for the asymmetry between redundant use of different adjectives attributes these patterns to the inherent relativity of the adjective types the adjectives belong to (**AparicioEtAl2016, Aparicio2018**). Because the meaning of scalar adjectives are dependent to a comparison class, they are thought to be more context-dependent than color adjectives and therefore less informative in general. The asymmetry between overmodification with color and material adjectives shows that this is not true.

- Visual/perceptual factors..

[lk: overview of this study]

[lk: redundant vs useful, efficient vs informative, speaker oriented vs listener oriented - can also discuss these in general discussion]

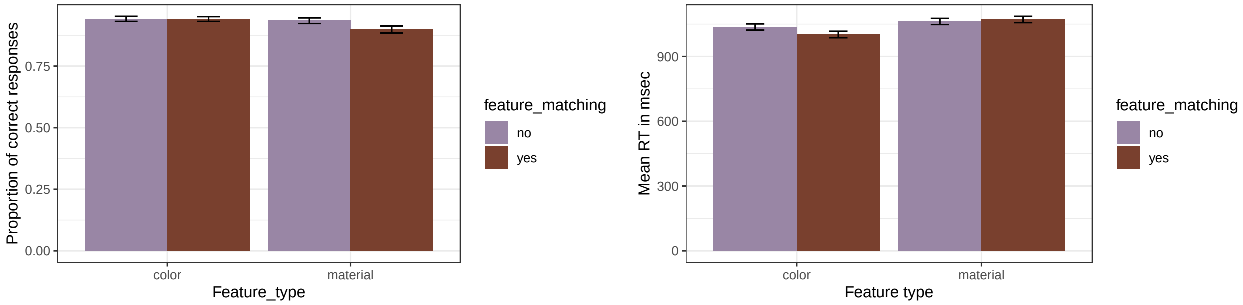
2. Experiment 1: Measuring perceptual difficulty. First, we need to norm the perceptual difficulty associated with establishing whether an object exhibits a color or material. Through a timed forced choice task we collected norms for color and material properties of 81 images.

2.1. PARTICIPANTS. We recruited 120 participants through Amazon Mechanical Turk. We excluded participants who were self-reported non-native English speakers (n=4) and participants with accuracy lower than 75% (n=11).

2.2. PROCEDURE. Participants saw images of objects with color or material adjectives and were asked to indicate whether the object had the property denoted by the adjective or not. Their task was to indicate ”yes” or ”no” by pressing the F or J key as quickly as possible. If participants did not respond within 4 seconds, the trial timed out. When participants responded correctly, a green border appeared around their selection, and when they responded incorrectly a red border appeared.

We collected perceptual difficulty norms for 12 objects that each occurred in two or three different materials and in three different colors. All resulting 81 images were separately normed for object nameability, feature nameability, object typicality and feature typicality. Every participant saw each image once and we collected 30 judgements for each image with matching and not-matching color and material adjectives. Color and material adjectives that didn’t match the images were randomly generated online for each participant from a pool of adjectives denoting properties of other images in the experiment.

2.3. RESULTS. Figure 1 shows the proportion and response times of correct responses to color and material adjectives. Overall, material adjectives resulted in higher error rates (β= 0.40, SE=0.09, p<.0001) and greater response times (β=5.46, SE=4.73, t=11.55, p<.0001) than color adjectives. We grouped the more perceptually difficult image-material pairs into a *high difficulty* group and less perceptually difficult image-color adjectives into a *low difficulty group* for testing in Exp. 2.



(a) Proportion of correct responses

(b) Response times of correct responses

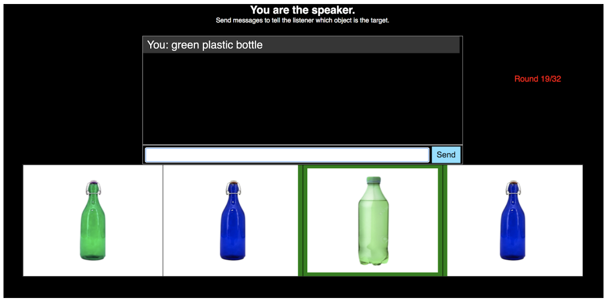
Figure 1. Responses to color and material adjectives (matching and not-matching images) [lk: replace with better formatted plots]

In order to categorize the image-word pairs in this way, we first grouped them in multiple different ways, always grouping the top 15 image-word combinations in the high-difficulty group and the bottom 15 image-word combinations in the low-difficulty group. First, we grouped pairs by response times, regardless of response correctness and match between the adjective and image. Then we grouped the pairs by response times of correct responses, and then by response times of correct responses to matching features and by response times of correct responses to not-matching features. We also grouped pairs by error rates in three different ways. Finally, we looked at the overlap between these groups and grouped the 8 image-material adjective pairs with the highest error rate and response times into a *high difficulty* group, and the 8 image-color adjective pairs with the lowest error rate and response times into a *low difficulty group*. [lk: way too wordy!!]

3. Experiment 2: Production of referring expressions. The goal of Exp. 2 was to elicit production probabilities of redundantly mentioning color and material adjectives for the high- and low-difficulty items normed in Exp. 1. In a free production interactive reference game we tested whether adjectives that denote more perceptually difficult properties are less frequently produced redundantly.[1]

3.1. PARTICIPANTS. We recruited 100 participants through Amazon Mechanical Turk and randomly paired them into speaker-listener dyads to play a real time communication game (50 pairs) (**Hawkins2015**). We excluded games where participants reported a native language different from English.

3.2. PROCEDURE. On each trial, participants saw a display with 4 images and chat box. Both the speaker and listener saw the same images in different positions. One of the images was designated as the target image, and marked by a green border in the speaker's display. The speaker's task was to describe this target image to the listener using the chat box to send messages. The listener's task was to guess the target image by clicking. After the listener made a selection, both participants received feedback about whether the target image was selected and advanced to the next trial.



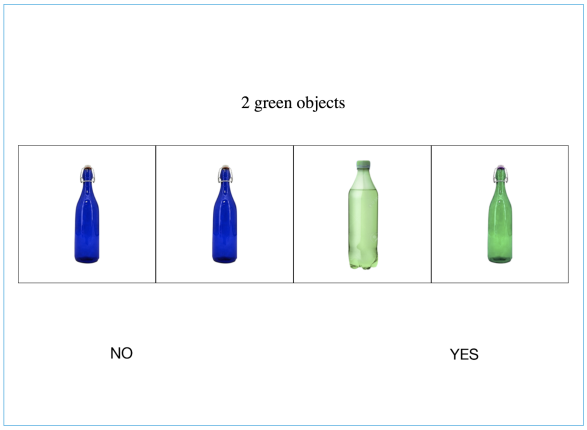
You are the speaker.

Send messages to tell the listener which object is the target.

You: green plastic bottle

Round 19/32

Figure 2. Example display from Exp. 2: speaker's perspective on a *low-difficulty (color redundant)* trial



2 green objects

Figure 3. Example display from Exp. 3: correct color trial

Participants completed 32 trials. Of these, half were critical trials and half were filler trials. On critical trials (Figure 2), the 4 images were of the same object and either *color* or *material* was redundant for distinguishing the target. One of the images, the competitor, always shared the redundant feature with the target and the two distractors shared the sufficient feature with the competitor. On 8 high-difficulty trials, mentioning the material was redundant; on 8 low-difficulty trials, color was redundant. On filler trials, the 4 images were of different objects and both color and material mention were redundant for unique reference. Filler items were of

[1] Procedure, materials, analysis and exclusions were preregistered at https://osf.io/57c6u.

4 different types: the competitor either shared the color, material, both or none of the features with the target.

3.3. RESULTS. We first classified the produced utterances as 'color-and-material' (redundant), 'only-color' or 'only-material'. Proportion of redundant "color and material" utterances and non-redundant utterances are shown in Figure 4. We conducted a mixed effects logistic regression predicting redundant adjective use from fixed effects of redundant property, with random by-subject and by-item intercepts and slopes for redundant property. There was a main effect of redundant property, such that speakers were more likely to redundantly mention color than material (β= 2.11, SE=0.62, p<.0001), replicating the previously observed asymmetry between overmodification with color and material adjectives on a new set of items. Our analysis of the responses to filler trials showed that the preference to mention color transferred to trials in which neither color nor material mention was required for unique reference [lk: add figure if there is space].

In a second step, we manually checked for the use of modifiers other than the color and material adjectives. We found that on 39% of all utterances, participants used a different kind of modifier to reference the target. These modifiers included shape (rectangular table), size (long table), shade (darker green plate) and [lk: ?] (solo cup) modifiers. The full data pattern with color, material and other modifiers revealed that when material is the redundant feature, participants tend to only mention color, and when color is the redundant feature, they either produce color-and-material utterances or overmodify with color and use another modifier. [lk: fix this sentence]

Given the norms from Exp. 1, these results suggest that the more difficult it is to judge whether an object has a feature, the less likely speakers are to redundantly mention that feature, providing initial support for the perceptual difficulty hypothesis. [lk: talk about limitations of this study - motivation for testing perceptual difficulty in context] [lk: do not want to assume that color modifiers have inherently higher semantic values than size modifiers - more nuanced semantics]

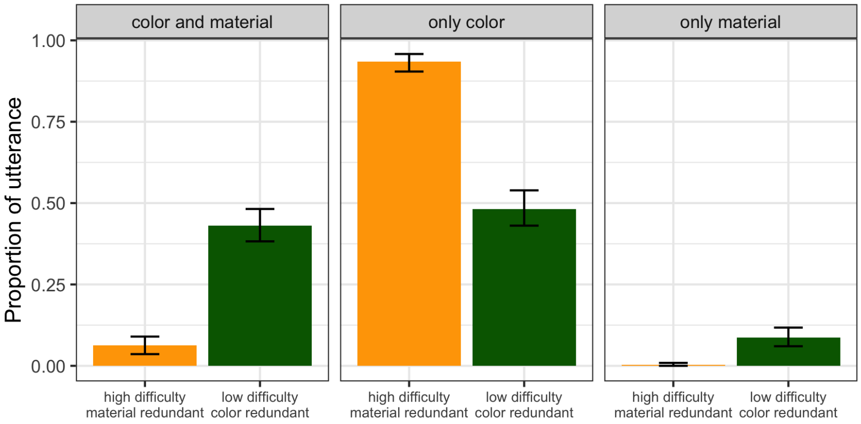


Figure 4. Proportion of redundant "color and material" utterances vs non-redundant utterances in high and low difficulty trials

4. Experiment 3: Perceptual difficulty in context. To get a more fine grained measure of perceptual difficulty that takes the contextual factors into account, we ran a third experiment.

4.1. PARTICIPANTS. We recruited 400 participants through Prolific. We excluded participants with accuracy was lower than 75% (n=24) and responses that were too slow (2.5 standard deviations away from the mean response time) (217 responses).

4.2. PROCEDURE. Exp. 3 was identical to Exp. 1 but instead of seeing the images in isolation, participants saw the displays from the production experiment. These displays appeared with short descriptions that were of the form ”X [adjective] objects” and either included a color or material adjective.

4.3. RESULTS. Overall, participants responded to material adjectives with higher error rates ([lk: add]) and in greater response times ([lk: add]) than color adjectives, replicating the results of Exp. 1. [lk: OR We ran a mixed effects linear regression model predicting log-transformed response time from fixed effects of redundant property with random by-participant intercepts. We observed a main effect of redundant property such that responses to material adjectives were slower than the responses to color adjectives ([lk: add]).]

To address the more theoretically interesting question of interest, we formalized the notion of perceptual difficulty in context. We computed a perceptual difficulty difference score for the target item in each context by subtracting the mean response time for the sufficient property from the mean response time for the redundant property. Higher difficulty scores indicate that the redundant property of the target object is less perceptually difficulty than its sufficient property. [lk: which model to discuss?]

5. General discussion.

6. Citations and references. Use author-date notation for in-text citations, for example: ... as noted recently by Jameson (2012) and Mateus (2014), drawing on insights from various other researchers (e.g., Nelson 1986:223–28, Martin 2003, Wellington & Johnson 2016), there have been numerous technological advances in the procedures used to publish research articles online. Include URLs or DOIs with active hyperlinks in citations. The *Semantics and Pragmatics* stylesheet sp.bst meets the LSA formatting requirements for the list of references, and so may be used. See info.semprag.org.