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

Leyla Kursat & Judith Degen*

Abstract. [lk: this is taken from the abstract - shorten and add 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 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 (Pechmann 1989; Sedivy 2003; Gatt et al. 2011; Rubio-Fernández 2016; Degen et al. 2020). 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 (Degen et al. 2020; Davies & Katsos 2013; Koolen et al. 2013) 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 (Degen et al. 2020; Westerbeek et al. 2015; Mitchell et al. 2013).

Many explanations are offered in the literature for the source of overinformative referring expressions including pragmatic, semantic, lexical and visual ones. The pragmatic account, proposed by Sedivy (2003), takes into account the contrastive function of mentioning the color of objects that occur in predictable colors. According to this account, 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. [lk: add? 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.] In addition to listeners' expectations of informativity, production of overinformative referring expressions have been shown to depend on the lexical category of the noun (Rubio-Fernández 2016) and the semantics of the adjective involved (Rubio-Fernandez et al. 2019; Sedivy 2003). Based on their comparison of the looking patterns with color and scalar adjectives, Aparicio et al. (2018) report that asymmetries could be attributed to the different semantics of the adjectives, more specifically, the inherent relativity of scalar adjectives. [lk: add? The case of color and material adjectives shows that semantic explanations alone cannot account for all observed asymmetries with redundant adjective use.] Recent work by Viethen et al. (2017) 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. Accounts that focus on the role of perceptual factors take into consideration such contextual effects and argue that redundant adjective use is sensitive to relative visual salience of properties adjectives denote (Rubio-Fernandez et al. 2019).

[lk: fix?] In this paper, we investigate the role of perceptual difficulty in the production of overinformative referring expressions. We show that perceptual difficulty associated with establishing whether an object exhibits a property could be a contributing factor to the noise term assumed by the computational model of referring expression production developed by Degen et al. (2020). In doing so, we provide an pragmatic explanation for how systematically related perceptual factors are to informativity calculations.

The computational model of referring expression production developed by Degen et al. (2020) accounts jointly for a variety of overmodification phenomena and proposes a unified quantitative account for their emergence. Couched within the Rational Speech Act (RSA) framework (Goodman & Frank 2016), 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). Degen et al. (2020) 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.

Here, we take the first step in testing the hypothesis that the noise term reflects the perceptual difficulty of establishing whether the property denoted by the adjective holds of the contextually relevant objects. We test this in the domain of color and material adjectives and predict 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. Finally, in Exp. 3, we investigate the role of perceptual difficulty beyond the property type.

2. Experiment 1: Measuring perceptual difficulty. First, we collected perceptual difficulty norms for color and material properties of 81 images. Through a timed forced choice task we measured the perceptual difficulty of establishing whether the objects exhibit a color or material property.

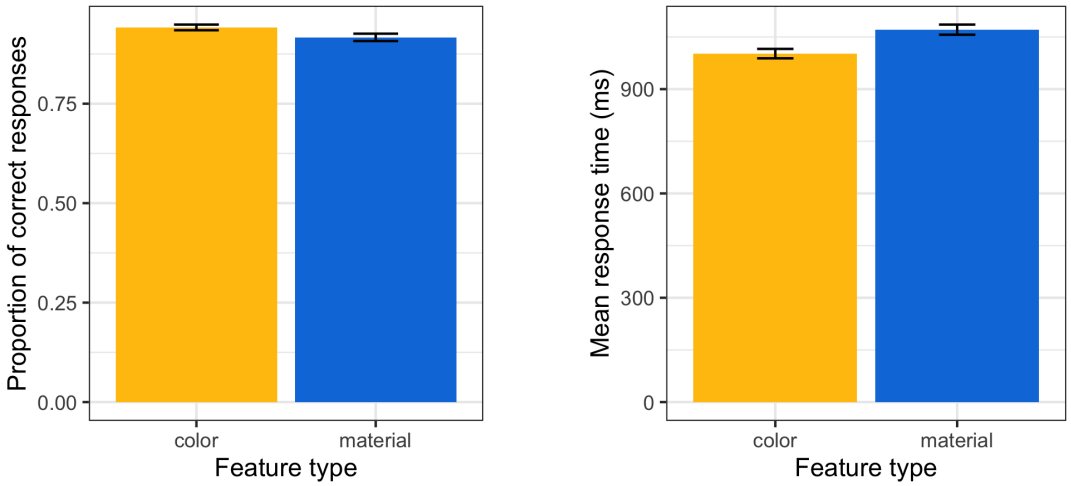
2.1. PARTICIPANTS. We recruited 120 participants through Amazon Mechanical Turk. We excluded participants who were self-reported non-native English speakers (n=4) and participants with accuracy lower than 75% (n=11).

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 selected 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.

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 re-



(a) Proportion of correct responses

(b) Response times of correct responses

Figure 1. Response patterns to color and material adjectives in the norming experiment

sponses 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: 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) (Hawkins 2015). 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.

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

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

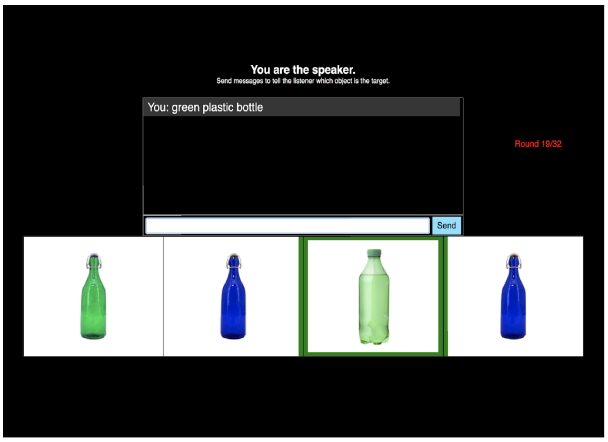


Figure 2. Example display from Exp. 2: speaker's perspective on a *low-difficulty (color redundant)* trial

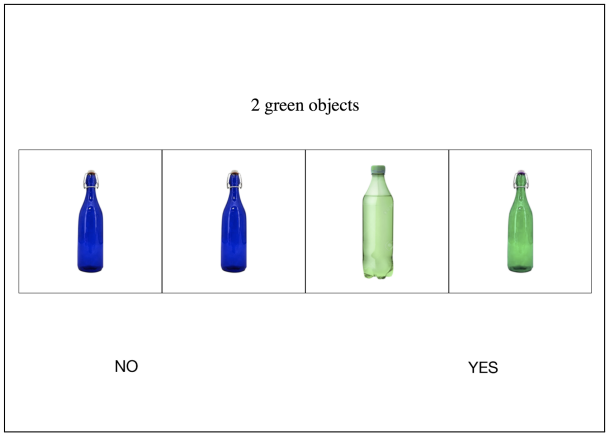


Figure 3. Example display from Exp. 3: color trial with correct number

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 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= 2.32, SE=0.64, 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.

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 (*dark blue plate*) and type [lk: ?] (*solo cup*) modifiers. The full data pattern with color, material and other modifiers revealed that when material was the redundant feature, participants mentioned only the color, and when color was the redundant feature, they either produced color-and-material utterances or overmodified with color and used a modifier denoting a property other than color or material.

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. However, these results do not reflect perceptual difficulty differences beyond the property type. Because the difference between response times for the two properties wasn't large enough, we weren't able to detect

perceptual difficulty differences within property type. In order to get more power and replicate this effect with a different dependent measure of perceptual difficulty, we ran a third experiment.

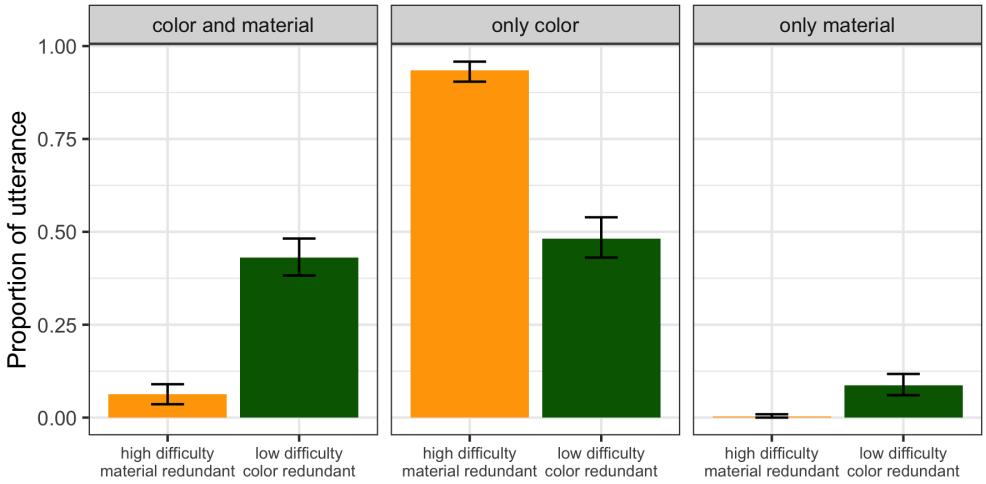


Figure 4. Proportion of redundant "color and material" utterances vs non-redundant utterances in high and low difficulty trials

4. Experiment 3: Perceptual difficulty in context. The goal of Exp. 3 was to get a more nuanced measure of perceptual difficulty that takes contextual factors into account.

4.1. PARTICIPANTS. We recruited 400 participants through Prolific. We excluded participants with accuracy 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 included a number and either a color or material adjective (see Fig. 3). On half of the trials, the statement was correct and on the other half, it was incorrect. The use of color and material adjectives was also balanced. We collected 100 judgements for each feature in the display, for all the different displays.

4.3. RESULTS. First, we analyzed the mean response times to the two property types to address the first question of interest, are properties denoted by material adjectives more perceptually difficult than ones denoted by color adjectives? To this end, we ran a mixed effects linear regression model predicting log-transformed response time from fixed effects of redundant property with random by-participant an by-item intercepts and slopes and slopes for redundant property. We found that responses to color adjectives were faster than the responses to material adjectives ((β= 0.24, SE=0.018, t=-59.62, p<.0001)), replicating the results of Exp. 1.

To address the more theoretically interesting question of interest, we analyzed the mean response times to redundant and sufficient properties of target items of Exp. 2. In order to identify the individual contribution of trial type and mean response time (to redundant adjective) on redundant adjective use, we first ran a simple linear model predicting log-transformed

response time from trial type, and then used the residuals of this model as the input to a mixed effects logistic regression predicting redundant adjective use. We found no effect of response time beyond trial type.

Figure [lk: add?] shows the correlation between trial type and response time before and after residualizing.

The distribution of mean log-transformed response times by trial type shown in Figure [lk: add?] shows that trial type explains the difference in redundancy, with no explanatory power from the response time measure.

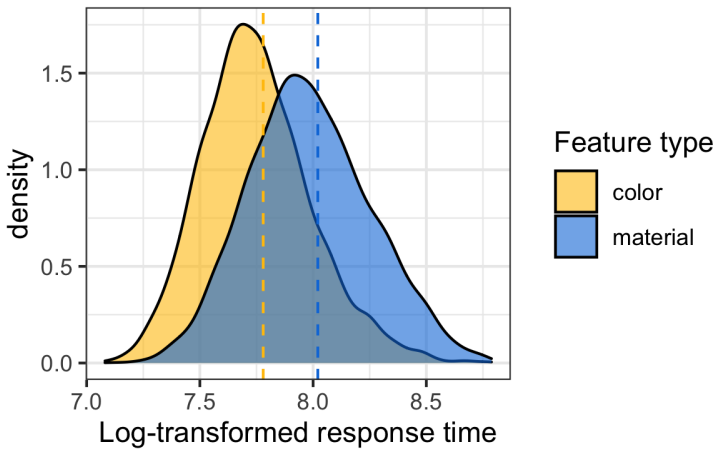


Figure 5. Density plot of log-transformed response times to color and material adjectives

5. General discussion. We tested the role of perceptual difficulty in explaining the asymmetry in overmodification patterns with color and material adjectives. The work thus far provides evidence for the weak version of the perceptual difficulty hypothesis, that the propensity to redundantly use color and material adjectives may be driven by the asymmetry in the perceptual difficulty involved in establishing whether or not a particular object is of a particular color or material. We have not found evidence for the strong version of the hypothesis, that beyond property type, perceptual difficulty modulates redundancy. The items we selected were structured such that there wasn't too much within-category variability, material being more difficult to asses than color categorically. A stronger test of the hypothesis would specifically select for bigger within-category perceptual difficulty variability.

An open issue discussed in this domain is whether overmodification helps the speaker, because the redundant attributes are the easiest to produce, or the listener, because they help to identify the target. Our task doesn't allow us to adjudicate between these two views. We provide evidence for the speaker internal pressure for redundantly mentioning the perceptually easy feature type, but we remain agnostic with regards to how this aids listeners in target identification.

References

Aparicio, Helena, Christopher Kennedy & Ming Xiang. 2018. Perceived informativity and referential effects of contrast in adjectivally modified nps. In *The semantics of gradability, vagueness, and scale structure*, 199–220. Springer.

Davies, Catherine & Napoleon Katsos. 2013. Are speakers and listeners 'only moderately gricean'? an empirical response to engelhardt et al.(2006). *Journal of Pragmatics* 49(1). 78–106.

Degen, Judith, Robert D Hawkins, Caroline Graf, Elisa Kreiss & Noah D Goodman. 2020. When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological Review* .

Gatt, Albert, Roger PG van Gompel, Emiel Krahmer & Kees van Deemter. 2011. Non-deterministic attribute selection in reference production .

Goodman, Noah D & Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences* 20(11). 818–829.

Hawkins, Robert XD. 2015. Conducting real-time multiplayer experiments on the web. *Behavior Research Methods* 47(4). 966–976.

Koolen, Ruud, Martijn Goudbeek & Emiel Krahmer. 2013. The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science* 37(2). 395–411.

Mitchell, Margaret, Ehud Reiter & Kees Van Deemter. 2013. Typicality and object reference. In *Proceedings of the annual meeting of the cognitive science society*, vol. 35 35, .

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

Rubio-Fernández, Paula. 2016. How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in psychology* 7. 153.

Rubio-Fernandez, Paula, Helena Aparicio Terrasa, Vishakha Shukla & Julian Jara-Ettinger. 2019. Contrastive inferences are sensitive to informativity expectations, adjective semantics and visual salience .

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

Viethen, Jette, Thomas van Vessem, Martijn Goudbeek & Emiel Krahmer. 2017. Color in reference production: the role of color similarity and color codability. *Cognitive science* 41. 1493–1514.

Westerbeek, Hans, Ruud Koolen & Alfons Maes. 2015. Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in psychology* 6. 935.