Mentioning atypical properties of objects is communicatively efficient

Elisa Kreiss, Judith Degen, Robert X.D. Hawkins, Noah D. Goodman
ekreiss@uos.de, {jdegen,rxdh,ngoodman}@stanford.edu
Department of Psychology, 450 Serra Mall
Stanford, CA 94305 USA

Abstract

[jd: write once modeling results clear]

Keywords: keywords

Reference to objects is one of the most basic functions of language. How do speakers decide which of an object's properties to include in a referring expression? This problem of content selection ([jd: cite the dutch]) has plagued cognitive science for decades. For example, in Fig. 1c, the utterances 'blue banana', 'banana', 'blue fruit', etc. all uniquely establish reference to the same target: the blue banana. How do speakers decide between these? One factor that has been identified as affecting speakers' choice of referring expression is the expression's *contextual informativeness*. Assuming that properties either do or do not apply to objects, 'banana' would be the appropriate choice in Fig. 1c (where no other banana competes with the target banana), but 'blue banana' when there is also a competing brown banana (as in Fig. 1b). However, previous research has established that this is not the case: speakers generally prefer to mention properties of objects to the extent that they are atypical, even when doing so is unnecessary for uniquely establishing reference (Sedivy, 2003; Mitchell, 2013; Westerbeek, Koolen, & Maes, 2015; Rubio-Fernandez, 2016). That is, speakers are more likely to redundantly call a blue banana 'blue banana' but a yellow banana simply 'banana'. Why is this so?

An account of why more typical properties are less likely to be mentioned is still lacking. Some ([jd: cite]) have proposed that it is due to a speaker-internal pressure to mention salient properties; others ([jd: cite]) have proposed that speakers mention properties to facilitate the listener's visual search. Here, we ask the computational-level question: when should a rational speaker with the goal of correctly communicating the intended referent be expected to mention an object's color?

Following the bulk of the previous literature, we assume that a speaker's choice of referring expression is governed by multiple factors, including the expression's contextual informativeness and its cost. Following Graf, Degen, Hawkins, and Goodman (2016); Degen, Graf, Hawkins, and Goodman (in prep.) [jd: who else?], we model speaker behavior formally within the Rational Speech Act (RSA) framework (Frank & Goodman, 2012; Goodman & Frank, 2016). However, we show that with a deterministic semantics for nouns and color adjectives, RSA cannot capture typicality effects in language. Therefore, also following Graf et al. (2016); Degen et al. (in prep.), we allow the semantics of expressions to assume a continuous value. That is, we allow 'banana'-hood or 'blue'-ness to apply to a given object to some degree rather than deterministically. This change directly affects expressions' contextual informativeness, resulting in precisely the expected typicality effects, as we demonstrate below.

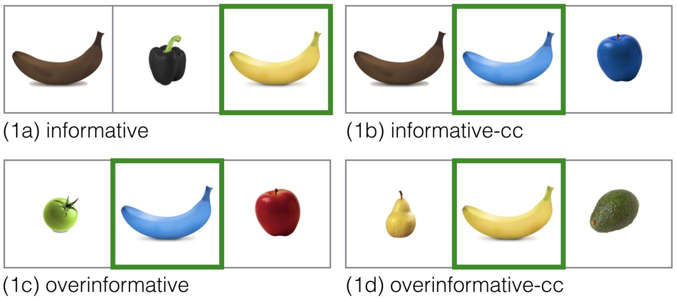
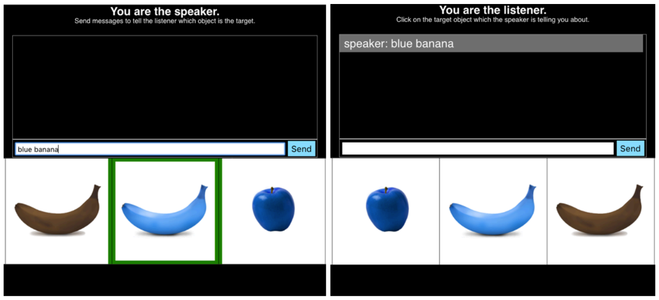


Figure 1: The four context conditions, exemplified by the *banana* domain. The target is outlined in green; the color and type of the distractors differ with each condition (see text). [jd: elisa: please make this figure with exactly the same labels as in results Figure 4 (informative, informative-cc, etc, and explain in figure caption – in making labels shorter, also make them more legible or we'll be rejected outright for illegible figures.]

In order to quantitatively evaluate the model, we collected freely produced referring expressions to objects in a web-based two-player reference game experiment (see Fig. 2). We presented participants with color-diagnostic objects that varied in how typical their colors were. Objects were presented in different contexts to vary the contextual informativeness of mentioning color (see Fig. 1).

We expected to replicate color typicality effects on referring expressions in at least those contexts where color use would be 'overinformative' (Fig. 1c) and 1d)), i.e., not strictly necessary for uniquely establishing reference. We also included control contexts in which mentioning color was informative (Fig. 1a) and 1b)). In these conditions, it is necessary to mention an object's color to unambiguously establish reference. Thus, we can test for typicality effects even in situations where mentioning color is seemingly necessary.

We begin by reporting the production experiment, including norming studies we conducted to empirically elicit typicality values for all utterance-object pairs. We then report a comparison of different RSA models with graded semantics against a deterministic semantics baseline. Model comparison makes use of the empirically elicited typicality values to derive behavioral predictions, which we compare against the data obtained in the reported production experiment.



You are the speaker.
Send messages to tell the listener which object is the target.

blue banana

You are the listener.
Click on the target object which the speaker is telling you about.

speaker: blue banana

Figure 2: Experimental setup.

Experiment: color reference game

Participants and materials

We recruited 120 self-reported native speakers of English over Mechanical Turk. The experiment was a multi-player reference game in which one participant was randomly assigned to the role of the speaker, and the other one to the role of the listener. The speaker's task was to communicate a target object out of three-object contexts to the listener. The target was always marked with a green border (see Fig. 2). The listener clicked on the object they thought was the target. The speaker and the listener could communicate freely through a chat box.

The stimuli were selected from seven food items (apple, avocado, banana, carrot, pear, pepper, tomato) which each occurred in three different colors that intuitively differed in typicality for that item. For example, the banana occurred in the colors yellow, brown, and blue. Each item occurred as a target and as a distractor. A pepper additionally occurred in a fourth color which only functioned as a distractor due to the need for a green color competitor.

Each presented context consisted of three objects, one target and two distractors. The contexts always corresponded to one out of four possible conditions. The different context conditions are referred to as "informative without a color competitor" (Fig. 1a), "informative with a color competitor" (Fig. 1b), "overinformative without a color competitor" (Fig. 1c), and "overinformative with a color competitor" (Fig. 1d). A context is referred to as overinformative when mentioning the type of the item, e.g., banana, would be sufficient for an unambiguous identification of the target. An additional mention of color would mean that the speaker uses the color adjective overinformatively, i.e., they are adding "unnecessary" information. However, in this condition the target never has a color competitor, i.e., if the target is brown, there is no distractor of the same color in the context. This means that an only-color utterance would lead to an unambiguous identification, too. This is not possible in the overinformative condition with a color competitor (Fig. 1d). In the informative conditions, the speaker needed to mention the color in addition to the type to provide an unambiguous utterance. Again, one context type did (Fig. 1a) and one did not include a color competitor among its distractors (Fig. 1b).

The item selection was randomized but conditioned on the corresponding context condition, i.e., the items had to fulfill the properties dictated by the condition. In the end, each participant saw 42 different contexts. All of the differently colored items were the target exactly twice but the context in which they occurred was drawn randomly from the four possible conditions mentioned above. All in all, there were 84 different configurations, i.e., seven target food items, each of them in three colors, where each could occur in four contexts. Trial order was randomized.

Procedure

Participants were randomly paired up and each was randomly assigned either to the role of speaker or listener. They communicated through a real-time multi-player interface as described in Hawkins (2015). The virtual environment of the experiment can be seen in Fig. 2. 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 speaker and listener received feedback about whether the right object had been selected.

Annotation

After collecting the data, the different utterances were labeled as belonging to one of the following categories: type-only ("banana"), color-and-type ("yellow banana"), color-only ("yellow"), category-only ("fruit"), color-and-category ("yellow fruit"), description ("has green stem"), color-modifier ("funky carrot"), and negation ("yellow but not banana"). Before sorting, two participants were excluded because they did not finish the experiment. Trials on which the speaker did not produce any utterances and trials on which the listener did not identify the target correctly were excluded as well. The remaining utterances (94% of the original set) were cleaned manually for misspellings and abbreviations, e.g., "banan" for banana. Finally, there were 10 speakers who consistently used roundabout descriptions instead of direct referring expressions (e.g., "monkeys love..." for banana). These participants were excluded because they were clearly not trying to simply communicate the target.

The 1942 resulting utterances were then categorized according to the categories laid out above. Only five utterances (0.003%) were assigned to the category "other".

Typicality norming

To evaluate the model we report below, we collected empirical typicality values for each utterance-object pair. Ratings were collected across three separate studies. The first study collected typicalities for adjective&noun-object pairs, e.g.,

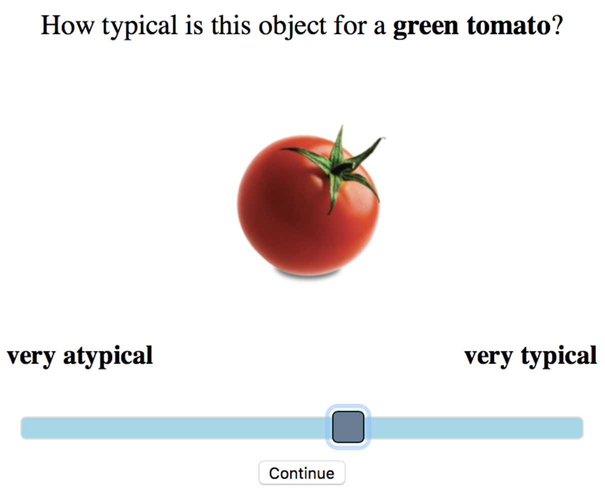

How typical is this object for a **green tomato**?

Figure 3: A typical trial in the 'Adj Noun' typicality norming study.

"yellow banana" as applied to a yellow banana, a blue banana, an orange pear, etc. The second study collected typicalities for noun-object pairs, e.g., "banana" as applied to a yellow banana, a blue banana, an orange pear, etc. The third study collected typicalities for adjective-color pairs, e.g., "yellow" as applied to a color patch of the average yellow from the yellow banana image or to a color patch of the average orange from the orange pear image. On each trial, participants saw one of the images used in the production experiment in isolation and were asked: "How typical is this object for a UTTERANCE", where UTTERANCE was replaced by an utterance of interest. In the color typicality study, they were asked "How typical is this color for the color COLOR?", where COLOR was replaced by one of the relevant color terms. They then adjusted a sliding scale with endpoints labeled "very atypical" and "very typical" to indicated their response. A screenshot of a typical trial is shown in Fig. 3. An overview of the differences between the three typicality norming studies differed is shown in Table 1. Each participant saw each 'correct' combination of utterances and objects (22 total) as well as a number of randomly sampled trials from the total set of items in each study, shown in the table.

Slider values were coded as falling between 0 ('very atypical') and 1 ('very typical'). For each utterance-object combination we computed mean typicality ratings. For example, mean typicalities for the banana items are shown in Table 2.

The way we elicited typicalities differs somewhat from the approach taken in previous work (e.g. Westerbeek et al., 2015), where participants are typically asked "How typical is this color for this object?" We deviated from this because for the purpose of testing the model (see below), what is required is the degree of applicability of an utterance to an object, rather than the degree to which an object's color is representative of that object. We expect, however, that the simple noun-object typicalities will yield very similar results as the Westerbeek question because the employed objects are color-diagnostic – asking whether an blue or a yellow banana is a typical *banana* is similar to asking whether or not the bananas' most salient property – their color – is typical.

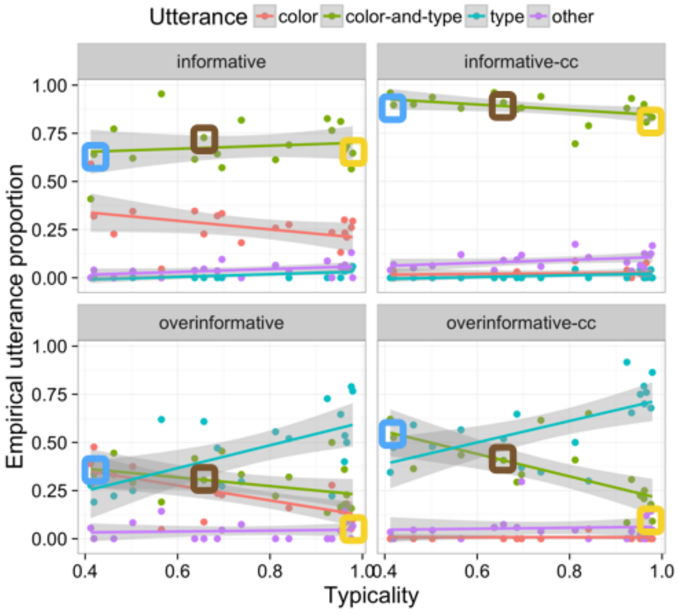


Figure 4: For each target, proportion of color-only (*yellow*), type-only (*banana*), color-and-type (*yellow banana*), and other (*funky carrot*) utterances as a function of mean object typicality for the type-only utterance, across conditions. COLOR *banana* cases are circled in their respective color.

Results and discussion

Proportions of type-only (*banana*), color-and-type (*yellow banana*), color-only (*yellow*), and other (*funky carrot*) are shown in Fig. 4. Visually inspecting just the explicitly marked *yellow banana*, *brown banana*, and *blue banana* cases suggests a clear typicality effect in the overinformative conditions as well as a smaller typicality effect in the informative conditions.

The following questions are of interest. First, do we replicate the previously documented typicality effect on redundant color mention beyond the one-item visual inspection? Second, do we observe typicality effects even when color is informative (i.e., technically necessary for establishing unique reference)? Third, are speakers sensitive to the presence of color competitors in their use of color or are typicality effects immune to the nature of the distractor items?

To address these questions we conducted a mixed effects logistic regression predicting color use from fixed effects of typicality, informativeness, and color competitor presence. We used the typicality norms obtained in the noun-object typicality elicitation study reported above as the continuous typicality predictor. Informativeness was coded as a binary variable (color *informative* vs. color overinformative) as was

Table 1: Overview of typicality norming studies.

Utterances	Example	Images	Participants	Trials	Items	Excluded
Adj Noun	*yellow banana*	object	66	110	484	4
Noun	*banana*	object	75	90	198	1
Adj	*yellow*	color patch	75	90	176	None

Table 2: Mean typicalities for banana items. Combinations where a deterministic semantics would return TRUE are marked in boldface.

	Banana items			Other
Utterance	yellow	brown	blue	
banana	**.98**	**.66**	**.42**	.05
yellow banana	**.98**	.33	.17	.05
brown banana	.28	**.90**	.18	.04
blue banana	.20	.18	**.91**	.06

color competitor presence (absent vs. present). All predictors were centered before entering the analysis. The model included by-speaker and by-item random intercepts, which was the maximal random effects structure that allowed the model to converge.

There was a main effect of typicality, such that the more typical an object was for the noun, the lower the log odds of color mention (β = -3.19, *SE* = 0.36, $p < .0001$), replicating previously documented typicality effects. Model comparison revealed that including interaction terms was not justified by the data, suggesting that speakers produce more typical colors less often even when the color is in principle necessary for establishing reference (i.e., in the informative conditions). There was also a main effect of informativeness, such that color mention was more likely when it was informative than when it was overinformative (β = 3.32, *SE* = 0.16, $p < .0001$). Finally, there was a main effect of color competitor presence, such that color mention was less likely when a color competitor was present (β = -0.66, *SE* = 0.13, $p < .0001$). This suggests that speakers are indeed sensitive to the contextual utility of color – color typicality alone does not capture the full set of facts about color mention.

In this section we have reported the replication of previously documented typicality effects on color mention as well as a novel demonstration of typicality effects even when color is informative. We have also shown that color mention is sensitive not only to typicality, but also to the nature of the distractors – when there is another distractor of the same color as the target, color is dispreferred compared to when there is not. Thus far we have only been concerned with analyzing the use of color, i.e., we collapsed COLOR and COLOR-AND-TYPE utterances into one category. However, our goal is to formulate a cognitive model that captures the production of referring expressions more generally. To this we turn next.

Modeling referential expressions

We consider a family of computational models characterizing the communicative challenge a speaker agent faces in the reference game scenarios above. These models are all situated within the broader Rational Speech Act (RSA) framework, which has successfully explained a range of sophisticated language phenomena through a recursive process of social reasoning between speaker and listener agents (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Goodman & Frank, 2016). More formally, we define a *literal listener* L_0 that selects between contextual referents $c \in \mathcal{C}$ proportionally to the meanings given by a lexicon $\mathcal{L}$:

$$L_0(c|u,\mathcal{C}) \propto \mathcal{L}(u,c)P(c)$$

We assume uniform prior beliefs $P(c)$ over referents. We then introduce a pragmatic speaker S_1, which selects an utterance $u \in \mathcal{U}$ to communicate an intended referent c_i by trading off *informativity* with *cost*:

$$S_1(u|c_i) \propto \exp\left(\alpha \log(L_0(c_i|u,\mathcal{C})) - \mathrm{cost}(u)\right)$$

where cost is usually defined as a function of an utterance's length or corpus frequency [jd: make sure we report the right cost function in final version] and α is a parameter controlling the speaker's "optimality": as $\alpha \to \infty$, they will choose utterances that maximize informativity. We explore several variations of this model in our model comparison, first considering different formulations of the lexicon $\mathcal{L}$ and then several novel ways of enriching the speaker to marginalize over possible *noise* in the listener's perception of the context:

Model alternatives

Baseline: The simplest version of the lexicon $\mathcal{L}$ uses truth-conditional meanings, such that a given utterance is either true or false of a given referent: $\mathcal{L}(u,c) = \delta_{u(c)}$. This is the traditional formulation in formal semantics, and the one most frequently used in previous RSA models.

Typicality: [jd: Motivate with banana example?] Next, we consider the elaborated model in Graf et al. (2016), which focused on nominal reference contexts where a speaker must choose between different taxonomic levels of reference (e.g. 'dalmatian,' 'dog,' or 'animal'). The primary innovation introduced by Graf et al. (2016) was a shift to a graded, real-valued meaning function based on the *typicality* of the referent relative to the utterance category: $\mathcal{L}(u,c) =$

typicality(u, c). This gives rise to phenomena where, for example, a speaker is more likely to use an overinformative utterance for a particularly atypical category member when a more typical referent is in context because they reason that L_0 would interpret the simpler utterance to mean the more typical referent. In the color typicality contexts we examine here, this corresponds to speakers being more likely to use an overinformative color-and-type utterance for a particularly atypical category member (blue banana) when there are other objects present that receive greater-than-zero typicality for the unmodified noun (*banana*) compared to when the target item is instead a highly typical category member (yellow banana).

Uniform Perceptual Noise: An additional source of uncertainty in language production comes from considerations of perceptual noise (in the speaker or in the listener, Jaeger, 2010). We propose that perceptual noise is another critical factor in explaining the overproduction of redundant modifiers. In this variation of the model, the speaker supposes that with some noise parameter p_{noise}, the listener might have misperceived one or more of the objects in context, thus leading to a corrupted context $\mathcal{C}'$:

$$S_{noisy}(u|c_i) \propto \exp\left(\alpha \log\left(\sum_{\mathcal{C}'} P(\mathcal{C}')L_0(c_i|u,\mathcal{C}')\right) - \text{cost}(u)\right)$$

In the simplest version of this noisy-context model, the prior over possible misperceptions $P(\mathcal{C}')$ is uniform, such that the true context has probability $P(\mathcal{C}) = p_{noise}$, and the rest of the probability mass is spread evenly across all possible replacements of one or more objects $c \in \mathcal{C}$. Intuitively, this has the effect of making the speaker more cautious about using less specific utterances: even if there is only one 'banana' in context, the speaker reasons that the listener may misperceive one of the distractors as another banana with some small probability, hence it may be useful to include a color modifier just in case.

Similarity-Based Perceptual Noise: This model is equivalent to the Uniform Perceptual Noise except for the prior $P(\mathcal{C}')$ over possible corruptions. Rather than choosing uniform across possible corruptions, we define a similarity metric such that misperceptions closer in similarity space to the true context (e.g. sharing one or more complete objects, only differing in color, and so on) are proportionally more likely: $P(\mathcal{C}') \propto \text{sim}(\mathcal{C}, \mathcal{C}')$. [rdh: Need to explain this similarity metric more once we've settled on one]

Model evaluation

[jd: waiting for results... – what we'll want is a faceted scatterplot of all the model posterior predictives we end up comparing plus ca. 2 paragraphs of discussion]

Discussion and conclusion

The work reported here makes both empirical and theoretical contributions. Empirically, we both replicated previously reported color typicality effects on the production of 'overinformative' referring expressions (Sedivy, 2003; Westerbeek et al., 2015; Rubio-Fernandez, 2016) and demonstrated unexpected color typicality effects on the production of 'informative' referring expressions as well: even when color is necessary for reference because a distractor object of the same category is present, color is sometimes dropped if the distractor is particularly atypical – i.e., speakers will sometimes refer to a yellow banana in the context of a blue banana simply as a *banana*. Theoretically, we have provided a unified account of color typicality effects in language production within the Rational Speech Act framework (Goodman & Frank, 2016) with graded rather than deterministic semantics. The best-fitting model [jd: add a sentence on its characteristics and why it was the most successful, using a qualitative example].

While the modeling reported here constitutes an intriguing extension of a graded-semantics RSA model from the nominal (Graf et al., 2016) to the modified NP domain, there remain a number of interesting open questions for future research. First and foremost, while we have shown that computing an utterance's contextual informativeness using empirically elicited typicalities for one-word and two-word utterances yields precisely the kinds of noisy literal listener distributions that pragmatic speakers seem to consider, it is an open question how the two-word typicalities are come by compositionally. That is, given the typicalities of the one-word utterances (e.g., *banana* and *blue*), what is the right way of combining these to arrive at the appropriate two-word utterance typicalities for a particular object? This is a question that has received attention time and time again whenever a non-deterministic semantics is considered (Kamp & Partee, 1995). Given the tools at hand, we are now in a position to formally test multiple hypotheses about the compositional semantics of terms with continuous meanings. This is an exciting avenue for future research.

[jd: discuss perceptual noise models vs typicality semantics models? ie different kinds of uncertainty]

[jd: add concluding sentence]

Acknowledgments

This work was supported by ONR grant N00014-13-1-0788 and a James S. McDonnell Foundation Scholar Award to NDG. RXDH was supported by the Stanford Graduate Fellowship and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-114747.

References

Degen, J., Graf, C., Hawkins, R. D. X., & Goodman, N. D. (in prep.). Over overinformativeness: Rationally redundant referring expressions.

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

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818-829.

Goodman, N. D., & Stuhlmüller, A. (2013, jan). Knowledge and implicature: modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–84. doi: 10.1111/tops.12007

Graf, C., Degen, J., Hawkins, R. X. D., & Goodman, N. D. (2016). Animal , dog , or dalmatian ? Level of abstraction in nominal referring expressions. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2261–2266). Austin, TX: Cognitive Science Society.

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

Jaeger, T. F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Kamp, H., & Partee, B. (1995, nov). Prototype theory and compositionality. *Cognition*, *57*(2), 129–91.

Mitchell, M. (2013). Typicality and object reference. *Proceedings of the 35th . . .* , 3062–3067.

Rubio-Fernandez, P. (2016). How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology*, *7*(153). doi: 10.3389/fpsyg.2016.00153

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

Westerbeek, H., Koolen, R., & Maes, A. (2015). Stored object knowledge and the production of referring expressions: the case of color typicality. *Frontiers in Psychology*, *6*(July), 1–12. doi: 10.3389/fpsyg.2015.00935