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

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.

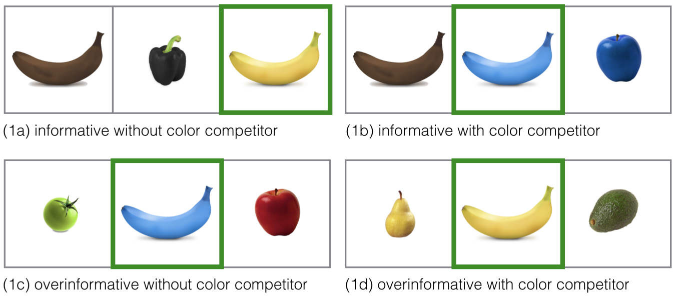

(1a) informative without color competitor
(1b) informative with color competitor
(1c) overinformative without color competitor
(1d) overinformative with color competitor

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).

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). [jd: continue here]

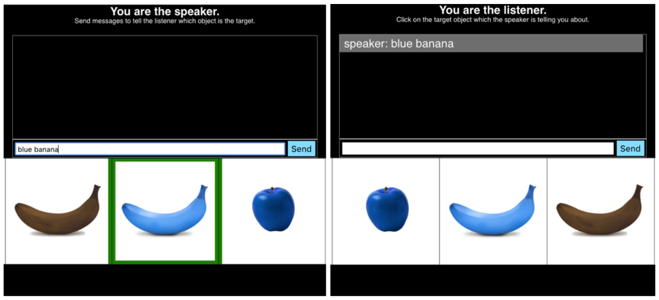
We expected to replicate color typicality effects on referring expressions in at least those contexts where color use would be 'overinformative' (Fig. 1c) and 1d)), i.e., not strictly necessary for uniquely establishing reference. We included contexts in which mentioning color was informative (Fig. 1a) and 1b)) as a control condition as well as to test (to our knowledge, for the first time) whether typicality effects are observed even in situations where mentioning color is seemingly necessary for uniquely establishing reference.

We first report the production experiment. We then report 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.

Experiment: color reference game

Methods

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 partici-



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.

pant 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 which each occurred in three different colors, e.g., one of the seven food items was the banana that occurred in the colors yellow, brown, and blue. All of those stimuli occurred as targets and distractors. A pepper additionally occurred in a fourth color which only functioned as a distractor due to the need for an adequate 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.

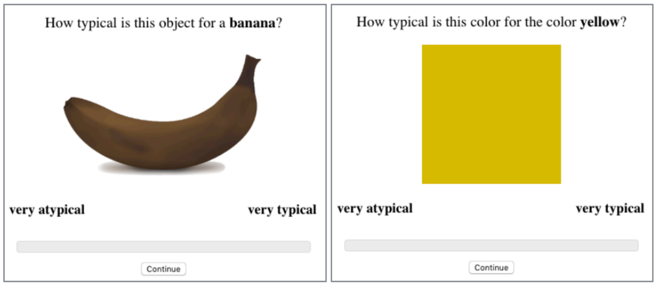
Procedure Participants were randomly paired up and each was randomly assigned either to the role of speaker or listener. They communicated through a real-time multi-player interface as described in Hawkins, Stuhlmüller, Degen, and Goodman (2015). The virtual environment of the experiment can be seen 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 had to be excluded because they did not finish the experiment, and therefore could not be paid. Furthermore, trials in which the speaker did not produce any utterances, and trials in which the listener could not identify the target correctly were excluded, too. The remaining utterances (94% of the original set) were cleaned manually for misspellings and abbreviations, e.g., "banan" for banana. Afterwards, the speakers that continuously refused to name objects but used unnatural descriptions instead (e.g., "monkeys love..." for banana) were excluded, too, as they misunderstood the task and the resulting expressions were therefore not relevant for the study.

The 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 For further analysis of the objects, we conducted two norming studies - an object and a color norming study. Both had 75 participants that were recruited over Mechanical Turk.

In the object norming study, the participants were shown a colored food item, e.g., a blue banana, and asked "How typical is this object for an X?" (X being a type, e.g., banana, or a category, e.g., fruit). The participants could rate the fitness on a continuous draggable scale from "very atypical" to "very typical". Every object in the set was paired with every



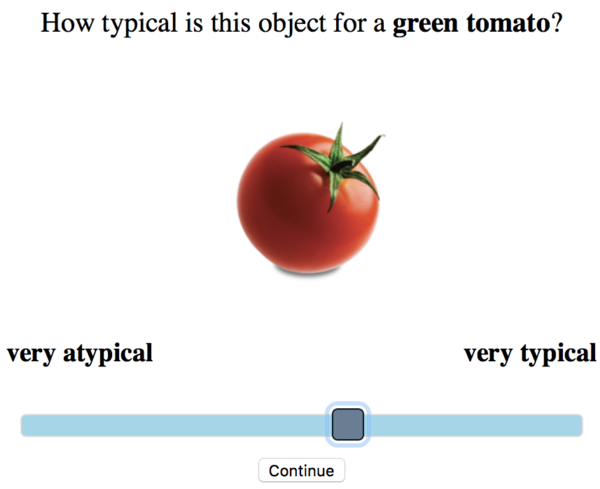
How typical is this object for a **banana**?

very atypical very typical

How typical is this color for the color **yellow**?

very atypical very typical

Figure 3: Typicality norming studies for object and color patch norming.



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

very atypical **very typical**

Figure 4: Typicality norming study.

type and category expression from the set, resulting in 189 different trials.

The aim of the color norming study was to identify how typical a certain object's color is for different color terms, i.e., how blue is the given color. The color patches were constructed by determining the average color of the object, stems and leaves excluded. The participants were presented with this color patch, and had to rate its typicality given a certain color term. Again, they had a continuous draggable scale with the end marks ”very atypical” and ”very typical”. Eight color terms were paired with 21 color patches and which resulted in 168 trials.

Due to the amount of trials, each participant only saw a subset of 90. The decision on which trials were shown was random with a slight preference on fitting combinations, e.g., seeing a blue banana and asking for ”banana”, in contrast to seeing a blue carrot and asking for ”apple”. The amount of data per combination ranged from 13 to 60. The assumed typicality values are the averaged slider values for each combination ranging from 0.004 (very atypical) to 0.989 (very typical).

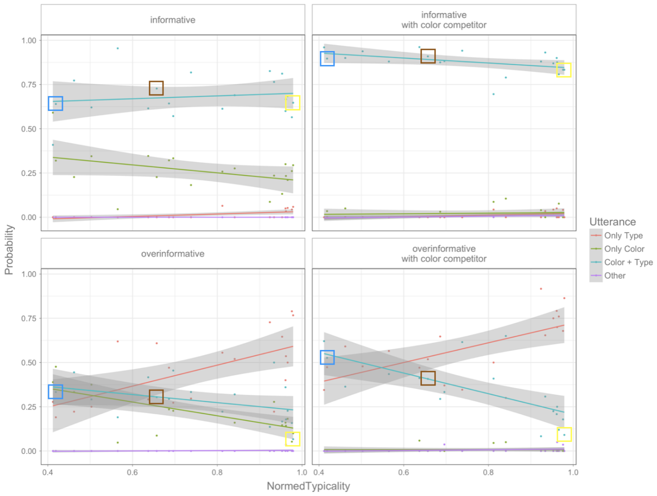


Figure 5: Proportion of Color (”blue”), Type (”banana”), and ColorAndType (”blue banana”) utterances as a function of mean object typicality ???for the Type utterance???, across conditions. ”COLOR banana” cases are circled in their respective color.

To get empirical typicality values of the objects, we conducted a norming study. 20 participants have been recruited via Mechanical turk. They were shown a colored food item, e.g., a blue banana, and asked ”How typical is this object for an X?” (X being a color-type combination, e.g., blue banana, or red apple). The participants could rate the fitness on a continuous draggable scale from ”very atypical” to ”very typical”. Every object in the set was paired with every color-type combination from the set, resulting in 484 different trials. Each participant answered 110 trials in which type consistent trials (showing all kinds of bananas when utterance contains banana) were always present, and non-fitting trials which were selected randomly. The amount of data per combination ranged from ... to ... . The assumed typicality values are the averaged slider values for each combination ranging from 0 (very atypical) to 0.968 (very typical).

Results

[jd: judith needs to fill this in]

Modeling referential expressions

Next, 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 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:

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

where cost is usually defined as a function of an utterance's length or corpus frequency 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:

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: 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: $\mathcal{L}(u,c) = \text{typicality(u, c)}$. 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.

Uniform Perceptual Noise: 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]

Discussion and conclusion

Acknowledgments

This work was supported by ONR grant N00014-13-1-0788 and a James S. McDonnell Foundation Scholar Award to NDG and an SNF Early Postdoc. Mobility Award to JD. 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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23335578 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., Stuhlmüller, A., Degen, J., & Goodman, N. D. (2015). Why do you ask ? Good questions provoke informative answers . In *Proceedings of the 37th annual conference of the cognitive science society.*

Mitchell, M. (2013). Typicality and object reference. *Proceedings of the 35th ...*, 3062–3067. Retrieved from http://csjarchive.cogsci.rpi.edu/Proceedings/2013/papers/0547/paper0547.pdf

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. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/12647560`

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. Retrieved from `http://journal.frontiersin.org/Article/10.3389/fpsyg.2015.00935/abstract` doi: 10.3389/fpsyg.2015.00935