# Blue banana – the linguistic phenomenon and not the discontinuous corridor of urbanisation in Western Europe

**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

When asked, referring to objects is something most people don't think about, even though it doesn't seem to be an easy task. There are several possibilities how to refer to objects in general. Looking at Fig. 1c, the utterances 'blue banana', 'banana', 'blue fruit', etc. would all lead us to the same target: the blue banana. But what do people actually choose to say in this context? And, more generally speaking, what governs how much information speakers include in referring expressions? One important aspect is for speakers to include just enough (but no more) information for their interlocutor to uniquely select an intended referent from a set of potential referents (cite Grice). Therefore, 'banana' would be the appropriate choice in Fig. 1c, but 'blue banana' when there is also a competing brown banana (as in Fig. 1b). However, this is not always the one we use.

The banana itself is a color-diagnostic object (Tanaka and Presnell, 1999), meaning different colors can have differently strong associations with, or be differently symptomatic of the object. A banana is normally associated with being yellow, a bit less with being brown, and not at all with being blue. When rating the association between the object and its color, we talk about *typicality*. For a cup, all colors have approximately the same typicality, and it is therefore a non-color-diagnostic object.

Various studies (Westerbeek, Mitchell, ) have shown that the color of a color-diagnostic object governs the chosen referring expression significantly. Westerbeek (cite) looked at contexts in which the referred-to object can unambiguously be distinguished from the other objects by only mentioning the type. Additionally, there has always been one object present that had the same color as the target (similar to Fig. 1d). Therefore, only considering unambiguous reference expressions, the use of color terms in any way would be overinformative, i.e., a true but unnecessarily added information. Westerbeek (cite) found that the lower the typicality for a color given the referred-to object is, the more likely one is to mention the color overinformatively.

Looking at the example from the beginning again, Westerbeek's result would suggest that people would also tend to say 'blue banana' in Fig. 1c to refer to the target object even if 'banana' would already be a sufficient way to do so.

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

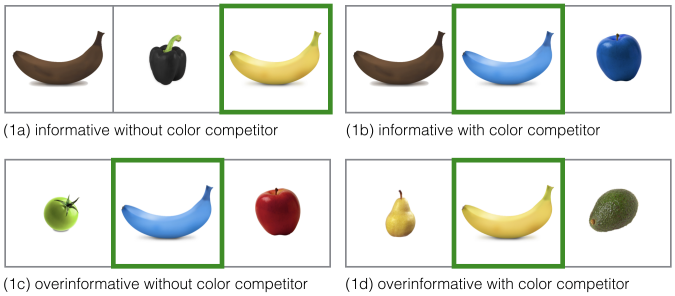


Figure 1: The four context conditions, exemplified by the *banana* domain. The target is outlined in green; the color and type of the distractors differ with each condition (see text).

To answer this question, we presented color-diagnostic objects with varying typicalities in different contexts to the participants. The experiment was done in form of a two-player reference game (see Fig. 2) hosted by Mechanical Turk. One of the two players was given the role of the listener, the other one the role of the speaker. They could communicate freely with each other through a chat box which ensured natural language data. In each context, both saw the same objects but only the speaker had the special marking of the referent which he had to communicate to the listener. The listener then had to deduce the correct target from the speakers utterance and mark it by clicking on the object. The context were manipulated by a differing typicality for the target object, and by distractor variation (either sharing the target's type, color or none of those). This way we could evaluate the effect of typicality and different kinds of contexts on referring expressions. We expected peoples utterances to 1) show a typicality effect at least in those contexts where color use would be overinformative, and 2) (what do we expect from informative???).

In this paper, we also describe how this effect can be modeled by using the Rational-Speech-Acts (RSA) framework (Frank & Goodman, 2012; Goodman & Stuhlmller, 2013). This model already showed good performance in various language interpretation tasks (e.g. Goodman & Stuhlmller, 2013; Kao, Wu, Bergen, & Goodman, 2014), but has only recently been applied to production data(, also with promising

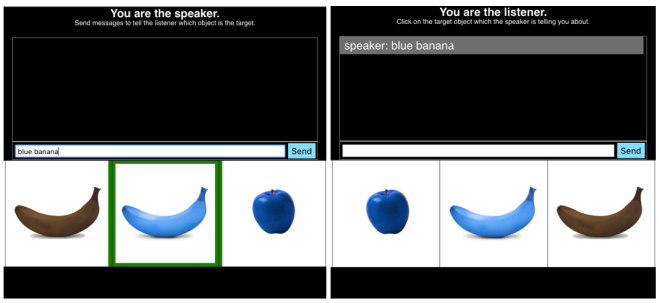


Figure 2: Experimental setup.

results???) (but see Franke, 2014; Orita, Vornov, Feldman, & Daume III, 2015, Graf, 2016). A speaker in RSA is treated as an approximately optimal decision maker who chooses which utterance to use to communicate to a listener. The speaker has a utility which includes terms for the cost of producing an utterance (in terms of length or frequency) and the informativeness of the utterance for a listener. The listener is treated as a literal Bayesian interpreter who updates her beliefs given the truth of the utterance. These truth values are usually treated as deterministic (an object either is a dog or it is not); here we relax this formulation in order to incorporate typicality effects. That is, we elicit typicality ratings in a separate experiment, and model the listener as updating her beliefs by weighting the possible referents according to how typical each is for the description used. We evaluate the quantitative model predictions against our production data. The model also allows us to evaluate the need for each extra componenttypicality, length, frequencyand determine whether the empirical bias toward reference at the basic level (Rosch et al., 1976) can be accounted for without build- ing it in as a separate factor. (from Caroline)
In this particular case, the RSA model can capture the effects of the context on the reference production, and the typicality effect.

## Experiment: color reference game

### Methods

**Participants and materials** We recruited 60 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 had to communicate which out of three objects was indicated as the target, and the listener clicked the one they assumed to be it. 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. The 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 being the target (the item that had to be referred to), and two distractors. The contexts always corresponded to one out of four possible conditions. The different context types 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 anymore in the overinformative condition with a color competitor (Fig. 1d). In the informative conditions, one always has to say the color in addition to the type to make an unambiguous utterance. Again, one context type does (Fig. 1a), and the other one does not have a color competitor under its distractors (Fig. 1b).
The item selection is random but conditioned on the corresponding context condition, i.e., the items need to fulfill the properties dictated by the condition. In the end, each subject sees 42 different contexts. All of the differently colored items are the target for exactly two times but the context in which they occur is drawn randomly from the four possible conditions mentioned above. All in all, we looked at 84 different configurations, i.e., seven target food items, each of them in three colors where each could occur in four contexts. The trial order was randomized.

**Procedure** The participants were randomly formed to pairs and each of them was randomly assigned either to the role of the speaker, or to the one of the 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". It was the speaker's task to tell the listener which of the three displayed objects was the target. The target could be identified by the green border around it. The listener then could either ask further questions, or immediately click on the object they thought was the correct one. Afterwards, both got a feedback showing whether the right object had been selected by the listener, or not.

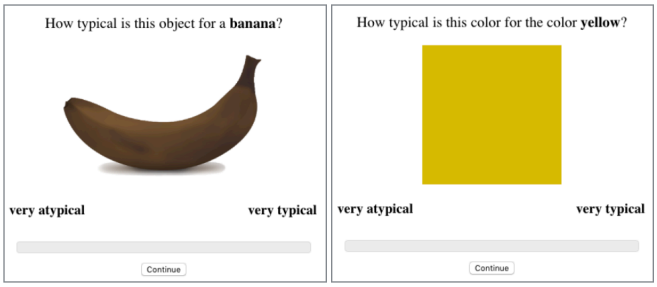


Figure 3: Typicality norming studies for object and color patch norming.

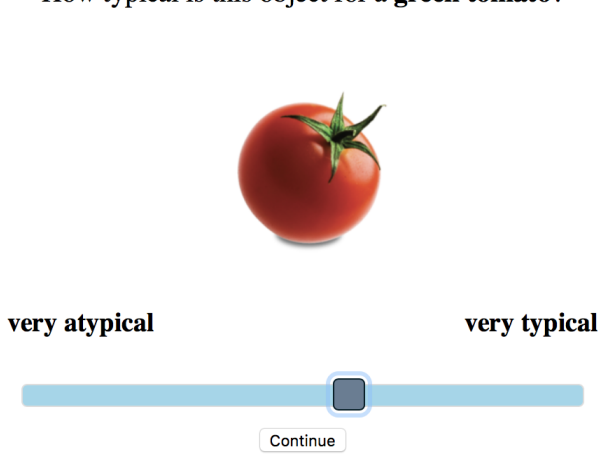


Figure 4: Typicality norming study.

**Annotation** After collecting the data, the different utterances had to be 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 (93.60% of the original set) had to be cleaned manually from 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.
Then the resulting utterances were categorized. Only five utterances (0.003%) had to be sorted into ”others”.

**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 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).

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).

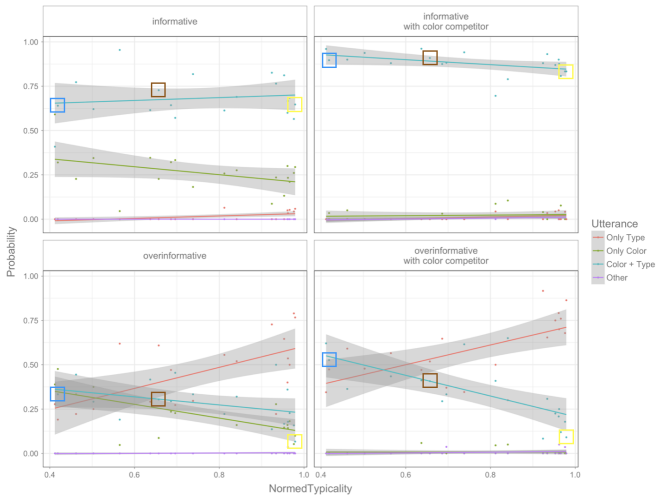


Figure 5: Proportion of Color (”blue”), Type (”banana”), and ColorAndType (”blue banana”) utterances as a function of mean object typicality ???for the Type utterance???, across conditions. ”COLOR banana” cases are circled in their respective color.

## Results

### Modeling referential expressions

In the baseline model, we assume deterministic semantics in which the typicality values can only be -Infinity, or 0. The simple model incorporates a non-deterministic semantics, i.e., the typicality values can range from 0 to 1. The ghost-context model is similar to the simple model, but adds uncertainty by taking the true context and creating possible variants from it in a random fashion. This contributes the possibility of a misperception. Possible contexts that are more similar to the true context appear with a higher probability than those with a lower one. Here, similarity is only calculated on basis of the whole object (color and type) being identical to its corresponding partner in the true context This model also has an extended version in which the similarity measure is more fine-grained, i.e., color is not a binary feature anymore (either true or false) but rather can be more or less similar to the color of the object in the true context.

### 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