STANFORD
UNIVERSITY

Qualifying Paper 1: Production expectations modulate contrastive inference

by

Elisa Kreiss

Committee
Chair: Judith Degen
Daniel Lassiter
Tobias Gerstenberg

A paper submitted in partial fulfillment
of the requirements for
Candidacy
in
Linguistics

04/30/2020

Contents

1 Desiderata for experimental items

Manipulating the typicality of color-diagnostic items for an experiment in the contrastive inference paradigm posited a number of requirements onto the items. First of all, the objects needed to be color-diagnostic (Section 2), since those objects show the highest difference in color modifier use dependent on their color typicality (Sedivy, 2003; Tanaka & Presnell, 1999; Westerbeek, Koolen, & Maes, 2015), our central manipulation. In addition, the items needed to be shape-diagnostic, such that they would still be recognizable when presented in a non-prototypical color. Plums, oranges, limes and lemons for example are objects with low shape-diagnosticity and are therefore hard to identify when presented in an atypical color (Section 5). Furthermore the items needed to be easily recognizable and nameable (Section 3 and 5). Previous literature suggests that unexpected utterances and labels might be a confound in eyetracking experiments (Qing, Lassiter, & Degen, 2018). To make the items further suitable for eyetracking studies, target and color competitor were never cohort competitors of each other (e.g., Cole and Jakimik (1980), Marslen-Wilson (1984)), i.e., they were always distinguishable at noun onset.

Furthermore, to our knowledge, previous experiments which manipulated the color typicality of objects did not counterbalance the colors used for typical and atypical instances. Certain colors (e.g., blue) primarily occurred as an atypical instance while other colors (e.g., green) occurred as a typical one. However, color hues vary in their a priori preference and elicit different emotional responses (see Elliot and Maier (2014); Palmer, Schloss, and Sammartino (2013) for reviews on color preference and psychology). The color hue imbalance for typical and atypical objects might therefore be a non-negligible confound to the typicality effect. Each color in our data set occurs twice as a typical instance and twice as an atypical instance to counteract this potential confound.

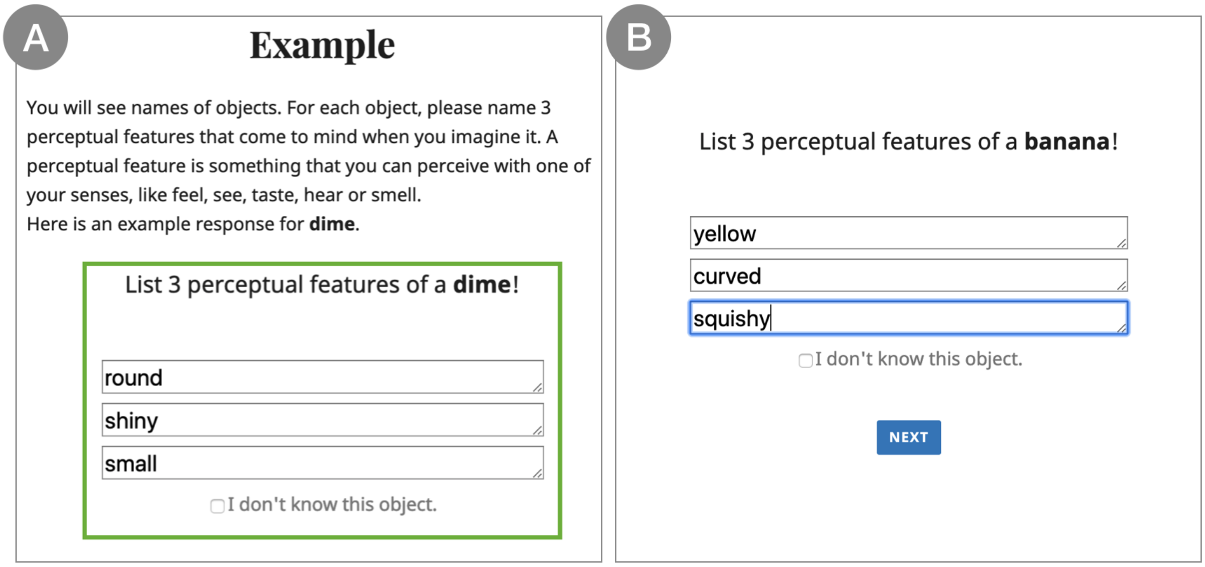
Finally, the typical and atypical instance of each object was normed to ensure that the color manipulation of the images shows the desired difference in typicality ratings (Section 4).

Since half of the conditions require the target and color competitor both to be (a)typical, we needed two (a)typical instances for each color. Motivated through this experimental design, the final set of stimuli consists of ten color diagnostic objects (each of them in a typical and atypical instantiation), evenly distributed over five colors. To determine the most suitable items, our initial set of potential stimuli comprised six colors (green, orange, pink, red, white, yellow), each with at least four presumably typical color-diagnostic instances (25 items in total).

2 Norming for color-diagnosticity

All potential stimuli were normed for their color-diagnosticity. Items were considered color-diagnostic if color was a perceptual property that was closely associated with the object, and if there was agreement on the specific color it was associated with (Tanaka & Presnell, 1999).

Participants. We recruited 40 participants over Amazon's Mechanical Turk. We restricted participation to workers within the US and a previous Hit approval rate of at least 98%. The study took about 20 minutes and we paid $4.00 for participation. All participants indicated that their native language was English.



Example

You will see names of objects. For each object, please name 3 perceptual features that come to mind when you imagine it. A perceptual feature is something that you can perceive with one of your senses, like feel, see, taste, hear or smell.
Here is an example response for **dime**.

List 3 perceptual features of a **dime**!

round
shiny
small

□ I don't know this object.

List 3 perceptual features of a **banana**!

yellow
curved
squishy

□ I don't know this object.

Figure 1. Example trial (in A) and critical trial (in B) for the color-diagnosticity norming experiment.

Materials and procedure. The experimental design is adapted from Tanaka and Presnell (1999). Participants were asked to list three perceptual features of an object, which they entered into three free production text boxes. They could proceed to the next trial if they either entered all three features or indicated by ticking a checkbox that they did not know the object. In the beginning of the experiment, participants saw an example trial where the term *perceptual feature* was defined and they were shown an example response for the object *dime*, which was described as *round*, *shiny*, and *small* (as used in Tanaka and Presnell (1999)).

Each participant saw 52 trials, four of which were control trials with nonce words. From the remaining trials, 25 asked for presumably color-diagnostic objects (four for each of the six colors and one additional green object), and 23 asked for presumably non-color-diagnostic objects. The example trial and one of the critical trials is displayed in Figure 1.

Analysis and exclusions. All participants indicated that they were unfamiliar with the four nonce words we had included as attention checks. Two participants were excluded because they rated more than eight objects as unknown to them, resulting in a total of 38 participants.

Results. An item was considered color-diagnostic if color was mentioned as the first feature, **and** participants agreed on the specific color the object was associated with.

As expected, color was in general more likely to be mentioned for the objects that were intended to be color-diagnostic and less likely for the non-color-diagnostic objects (see Figure 2). An exception to that was *bell pepper* which was expected to be non-color-diagnostic but for which participants still mentioned color as an important perceptual feature. However *bell pepper* was associated with more than one color. Although it was mainly described as *green*, several participants also mentioned *red* and to a small degree *black*, *yellow* and *orange*. For all items that were intended to be color-diagnostic, participants generally agreed on the color.

Figure 3 shows the proportion of color mention as the first perceptual feature, facetted by the color they were associated with. The dashed horizontal line marks the case when

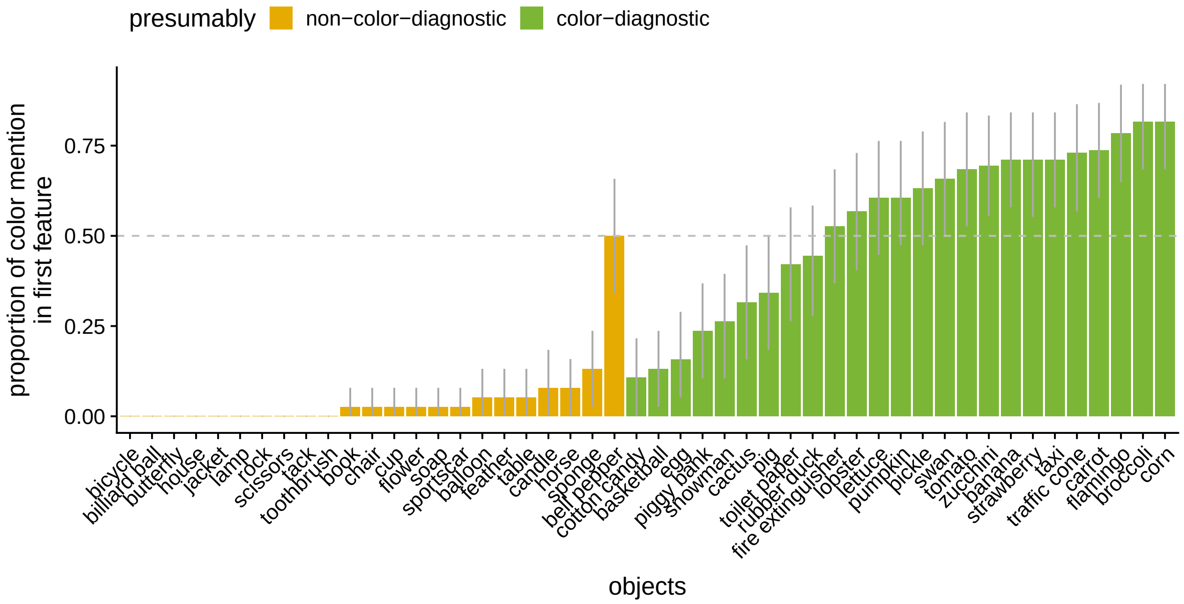


Figure 2. Proportion of color being mentioned as a first perceptual feature for presumably non-color-diagnostic objects (orange) and color-diagnostic objects (green). Color was generally mentioned more often for the color-diagnostic objects. An exception to that is bell pepper, which is categorized as non-color-diagnostic because it is associated with more than one color.

color was mentioned half of the time as the first perceptual feature. The items in the colors pink and white had the lowest proportion of color mention overall. This is the main reason why pink items were excluded from the final set of stimuli. Items with low proportions of color mention in each color were less likely to be chosen as stimuli, which is why we for example excluded *cactus*, *basketball* and *cotton candy* from the final set.

3 Norming for nameability

After we normed for color-diagnosticity, we chose an image depiction for each object and normed them for their nameability. Determining whether an object is nameable served two main purposes. Firstly, we ensured that the dominantly chosen label corresponds to the concept we had normed for in the color-diagnosticity norming (Section 2). Furthermore, participants should agree on the label that best describes the chosen depiction. Strong agreement among participants is an indication that the noun

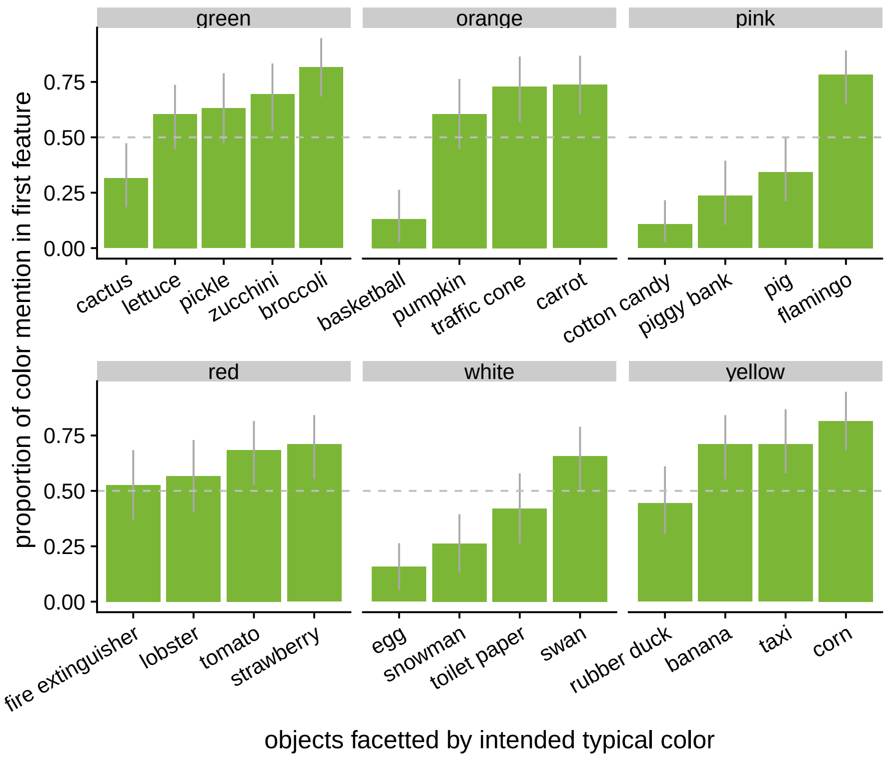
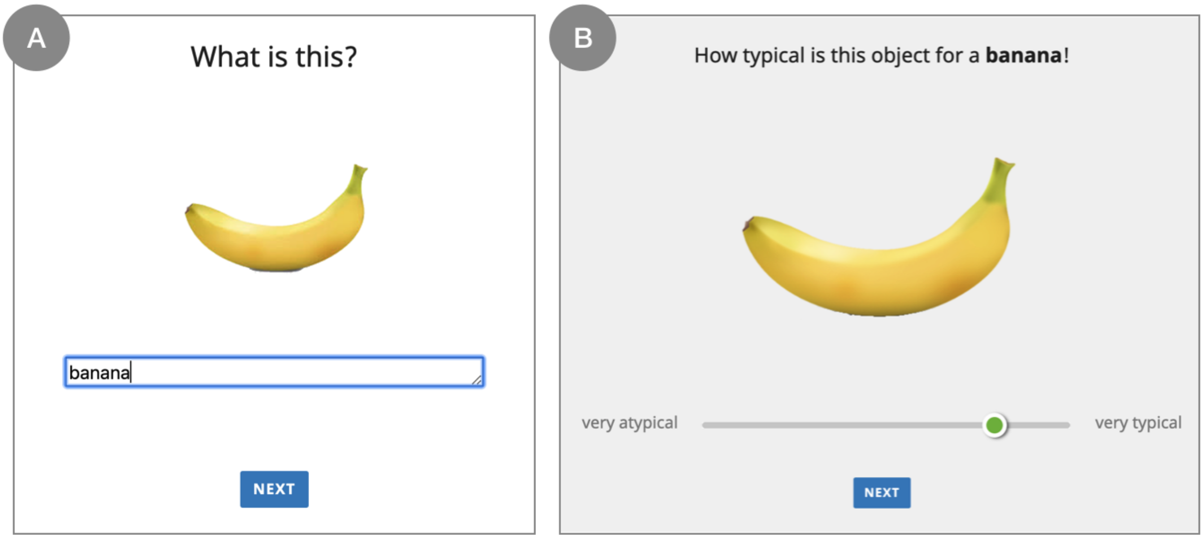


Figure 3. Proportion of color being mentioned as a first perceptual feature for presumably color-diagnostic objects, facetted by the assumed color association. The dotted line marks the point where half of the time, color was mentioned as the first feature.

label can be used reliably. Strong disagreement about the label might affect the production and comprehension of the referring expressions. If speakers are uncertain about an object's property (in this case, its type), they are more likely to overmodify (Horacek, 2005; Williams & Scheutz, 2017). Additionally, a listener's surprisal of hearing an unexpected label might affect for instance their fixations in eyetracking experiments, creating a potential confound in the analysis (Qing et al., 2018).

Participants. We recruited 20 participants over Amazon's Mechanical Turk. We restricted participation to workers within the US and a previous Hit approval rate of at least 97%. The study took about 7 minutes and we paid $1.30 for participation. All participants indicated that their native language was English.

Materials and procedure. Each participant saw 50 trials in which they were asked "What is this?" with a depiction of an object and a free production text field, as shown



What is this?

banana

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

Figure 4. Example trials of nameability and free production norming (in A) and typicality norming (in B).

in Figure 4A. We used the same objects as in the color-diagnosticity norming study (i.e., 25 color-diagnostic and 23 non-color-diagnostic objects). In two cases (lettuce and sports car) we chose two depictions to determine the most prototypical instance of these items.

Exclusions. Two participants were excluded because they indicated that they might have misunderstood the instructions, resulting in a total of 18 participants.

Results. We evaluated the results according to how many labels were used. If more than one label was used, we favored cohort competitors (e.g., *bike* and *bicycle* were more acceptable deviations than *traffic cone* and *cone*).

Overall, participants agreed on the labels, but there were some notable exceptions. Both, the pickle and zucchini, were called *cucumber* to a non-negligible degree. Given that they also have low shape-diagnosticity, both objects were excluded from the final set of stimuli. Other items that received a variety of labels were the traffic cone (e.g., *traffic cone, cone, caution cone, hazard cone*) and the rubber duck (e.g., *rubber duck, duck, duck toy*). These cases were problematic because the labels are not cohort competitors of each other, which makes them crucially distinct in real-time (auditory) setups such as eyetracking experiments.

Finally, we investigated which depiction best represented the generic lettuce: romaine or iceberg. We found that when participants saw the iceberg lettuce before the romaine lettuce, they simply called it *lettuce*. However, if they saw the romaine lettuce first, they called it *romaine lettuce* 20% of the time. This suggests that the iceberg lettuce is the more prototypical lettuce (in the MTurk community) and was therefore chosen for the final set of stimuli.

4 Norming for typicality

After selecting the most promising objects and creating atypical counterparts, we normed depictions of these items according to their typicality to ensure that they are in fact perceived as typical and atypical. Norming the atypical instances is especially necessary since in fact most of these items occur in different colors in the world. For example, there are red bananas, purple carrots, blue corn cobs, yellow tomatoes, varying colors of pumpkins and artificially colored eggs. A successful typicality manipulation should maximize the difference between typical and atypical ratings for each object.

Participants. We recruited 30 participants over Amazon's Mechanical Turk. We restricted participation to workers within the US and a previous Hit approval rate of at least 97%. The study took about 4 minutes and we paid $0.80 for participation. All participants indicated that their native language was English.

Materials and procedure. Each participant saw 45 trials in which they were asked "How typical is this object for a **NOUN**" with a depiction of an object in either its typical or atypical instance, as shown in Figure 4B, and where **NOUN** is the established label from the nameability norming experiment. Participants indicated their response on a continuous slider which was initialized in the center of the scale and was underlyingly coded as ranging from 0 to 100.

For the typicality norming, we selected 11 color-diagnostic objects from the set of the previous norming studies and presented each in their typical color and in one to two atypical colors. Overall, this resulted in 25 color-diagnostic and 20 non-color-diagnostic stimuli. The atypical depictions were created from the typical depictions by changing the typical to an atypical color hue[1]. This means that potential greenery around the item or stems were preserved to maximize the inherent naturalness of the item (as can be seen in the carrot or pumpkin items in Figure 7).

We only used the colors of the typical items (i.e., green, orange, red, white, and yellow) for the atypicality manipulation to create a counterbalanced final set of stimuli. Again, the colors were distributed evenly over all objects such that each color occurred as an atypical instance on exactly two objects.

Results. In the analysis, we assessed whether the color manipulation of the images showed the desired difference in typicality ratings.

As shown in Figure 5, there generally was a clear distinction between the typical and atypical instances for each object. Of the three items that were normed in two atypical colors (carrot, corn, and pumpkin), the red and white pumpkin showed the biggest difference. Therefore, we chose the white over the red pumpkin, and, following from that, the green carrot and red corn as atypical instances. Finally, even though the egg and snowman received similar ratings for their atypical instance, the white egg is rated slightly more typical than the depiction of the white snowman.

However we have to note that even though the orange banana is predominantly rated below 50, it is still not as atypical as other items. This might be due to the high similarity between the colors yellow and orange.

5 Norming for free production

Finally, we normed the stimuli chosen from the typicality norming study as to whether those depictions are nameable as intended.

Participants. We recruited 50 participants over Amazon's Mechanical Turk. We restricted participation to workers within the US and a previous Hit approval rate of at least 97%. The study took about 5 minutes and we paid $1.10 for participation.

[1] The exception to this are the red and yellow strawberries which were both created from a picture of a yellow-green strawberry.

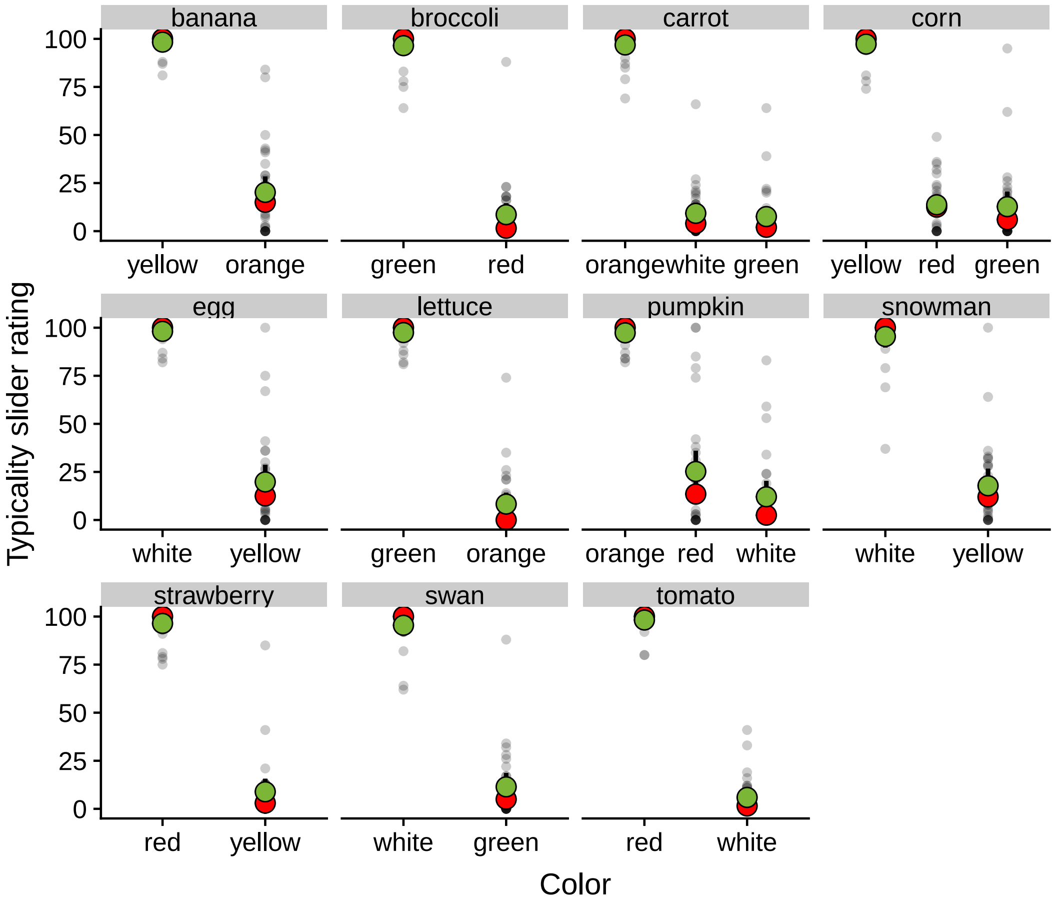


Figure 5. Typicality ratings for differently colored instances of the same object. Higher values indicate higher typicality. Individual data points are in gray, means in green and medians in red.

Materials and procedure. As stimuli, we used the objects and depictions selected from the typicality norming study, which was almost identical to the complete final set. The only redundancy at this point was the egg vs. snowman stimulus.

Participants saw each object exactly once and its typicality was randomized, resulting in 11 critical trials. This was done to prevent convergence effects on a label when the object occurred the second time (e.g., Clark and Wilkes-Gibbs (1986)). Each participant additionally completed 20 filler trials, where they saw non-color-diagnostic objects.

Analysis and exclusions. Three participants were excluded because they indicated that they did the experiment incorrectly or were confused, and another three participants because they used non-anticipated utterances in more than 1/3 of the trials.

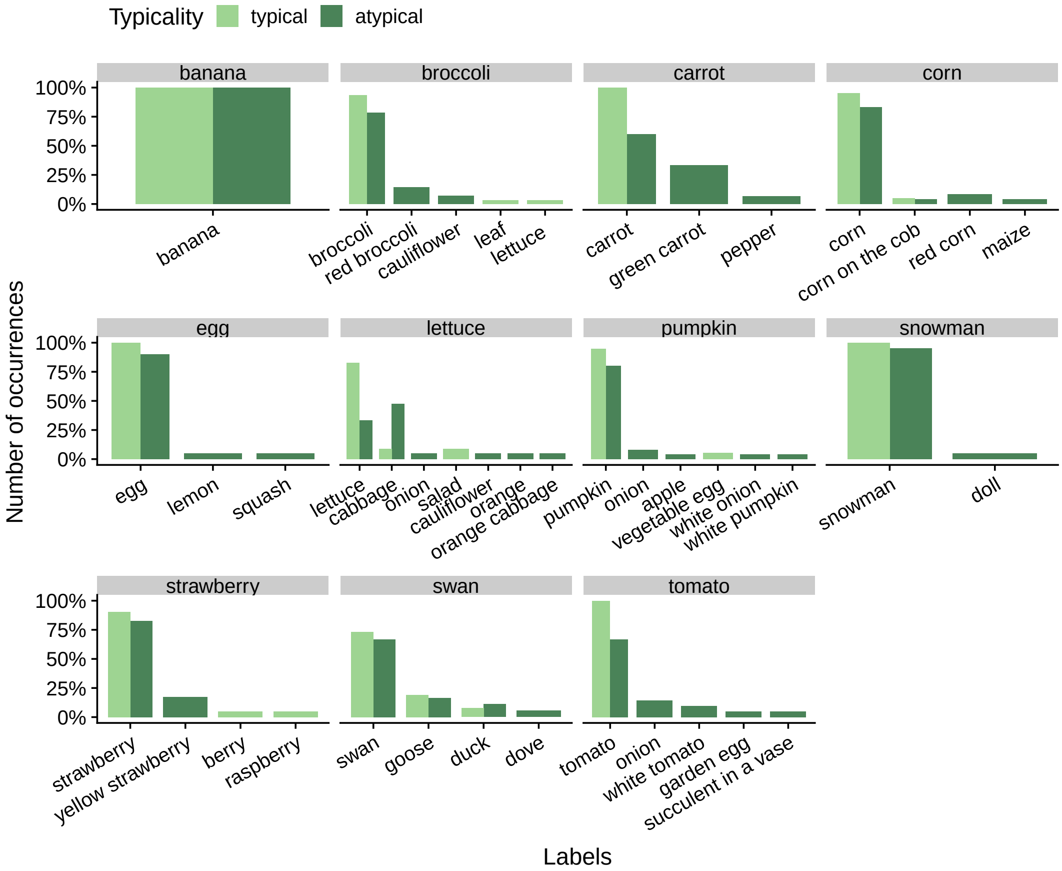


Figure 6. Labels produced for each item in a free production experiment for all remaining stimuli. Labels produced for typical items are in light green, labels produced for their atypical counterparts in dark green.

Results. As the results in Figure 6 show, participants generally gave the same label to both the atypical and typical instance. This supports the final stimulus selection. Since egg and snowman were both equally nameable, we chose the egg as the typical white instance due to its size compatibility with the other items.

There are three stimuli that are slightly less clearly nameable than the others: the atypical lettuce is often called *cabbage*, the swan is sometimes mistaken for a goose and the white carrot has been mislabeled as a parsnip by two participants. However we accept these deviations as negligible.

Lastly we observed that even in this highly simplified setup with written free text input, participants sometimes produced the color term spontaneously, but only for the atypical items (for example for the green carrot, red corn, and white tomato).

6 Conclusion

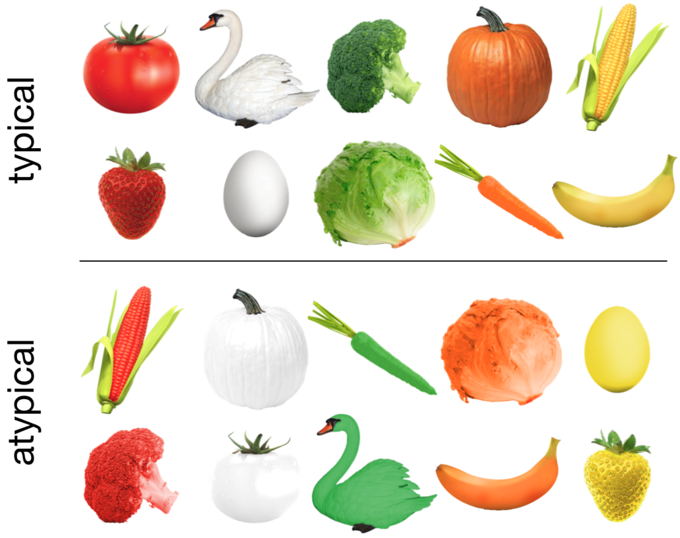


Figure 7. Final set of stimuli, ordered by color and typicality. Each object occurs in a typical and atypical color.

In the end, the final set of stimuli comprises 10 objects, each occurring in a typical and atypical color. Items can occur in the colors yellow, red, green, orange and white. Each color occurs twice as typical and twice as atypical. The objects are *banana, broccoli, carrot, corn, egg, lettuce, pumpkin, strawberry, swan,* and *tomato.* The full set of stimuli is displayed in Figure 7.

7 References

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Cole, R. A., & Jakimik, J. (1980). A model of speech perception. *Perception and production of fluent speech*, 133–163.

Elliot, A. J., & Maier, M. A. (2014). Color psychology: Effects of perceiving color on psychological functioning in humans. *Annual review of psychology*, *65*, 95–120.

Horacek, H. (2005). Generating referential descriptions under conditions of uncertainty. In *Proceedings of the tenth european workshop on natural language generation (enlg-05).*

Marslen-Wilson, W. D. (1984). Function and process in spoken word recognition: A tutorial review. In *Attention and performance: Control of language processes* (pp. 125–150). Erlbaum.

Palmer, S. E., Schloss, K. B., & Sammartino, J. (2013). Visual aesthetics and human preference. *Annual review of psychology*, *64*, 77–107.

Qing, C., Lassiter, D., & Degen, J. (2018). What do eye movements in the visual world reflect? A case study from adjectives. In *CogSci.*

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

Tanaka, J. W., & Presnell, L. M. (1999). Color diagnosticity in object recognition. *Perception & Psychophysics*, *61*(6), 1140–1153.

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*.

Williams, T., & Scheutz, M. (2017). Referring expression generation under uncertainty: Algorithm and evaluation framework. In *Proceedings of the 10th international conference on natural language generation* (pp. 75–84).