STANFORD
UNIVERSITY

Qualifying Paper 1

by

Elisa Kreiss

Committee
Chair: Judith Degen
Dan Lassiter
Tobias Gerstenberg

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

INSERT DATE OF SUBMISSION

Abstract

INSERT SOME ABSTRACT

Contents

[ek: determine how to spell (non)colordiagnosticity]

1 Introduction

1.1 SOME SUBSECTION

2 Experiment: Norming

Previous research has established that many different factors can affect eye movement data. For this reason we have run an elaborate set of norming studies to account for a variety of possible artifacts. [ek: talk about the artifacts – why is nameability important? why colordiagnosticity? What is so difficult about it?]

Motivated through our experimental design [ek: more?], we aimed to find ten color diagnostic objects, evenly distributed over five different colors. To find the most ideal items, we started off with six colors (green, orange, pink, red, white, yellow), each with four possible colordiagnostic instances. Those items were then normed for colordiagnosticity and nameability.

2.1 Norming for color-diagnosticity

[ek: watch your tenses; define a common textit/quotation style]

The norming of color-diagnosticity is adapted from [ek: cite Tanaka and Presnell]. They claim that an object should only be considered color-diagnostic, if the color property centrally defined the object's identity. For example, the color *red* is considered very typical for a sportscar, as is the color *yellow* for a banana. However participants are more likely to use the modified utterance *red sportscar* than *yellow banana*. [ek: Tanaka and Presnell] can account for these differences in production data by considering how relevant the property *color* is for the definition of the object. For a sportscar, the color property was rarely mentioned as defining the object which is in clear contrast to the banana.

Since the probability to which an object is spontaneously modified is part of the crucial manipulation in this work, we will adopt the method used by [ek: T and P] to determine the color-diagnosticity of the potential stimuli. Participants are asked to list

three perceptual features of an object, which they entered into three free production text boxes. They could only proceed if they specified all three features or indicated that by a button press that they did not know the object.

Participants. We recruited 40 participants over Amazon's Mechanical Turk. All participants indicated that they were unfamiliar with the four nonce words we included as attention checks, but two participants were excluded because they rated more than eight objects as unknown to them.

Procedure. 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 more green thing), and 23 asked for presumably non-color diagnostic objects.

Results. We evaluated the results according to whether a color was mentioned at all in the features, a color was mentioned as a first feature, and if a color was mentioned did participants agree on a specific color. [ek: results ...]

2.2 Norming for nameability

Goal: Are the image depictions we chose nameable, the way we intended?

Task: "What is this?"

free production

50 trials; 26 depictions of presumably color diagnostic objects (same as in color diagnosticity norming + 1 more lettuce depiction) and 24 presumably non-color diagnostic ones (same as in color diagnosticity norming + 1 more (sports)car);

20 participants; exclusion: 2 participants because they indicated that they were confused or didn't do the HIT correctly; resulting number of participants: 18

We evaluated the results according to how many labels were used. If more than one label was used, we favored cohort competitors over entirely separate terms (e.g., bike and bicycle are more acceptable than traffic cone and cone); Wrt to lettuce we had romaine and iceberg lettuce depictions. The simple noun lettuce was more frequently used for the iceberg than the romaine lettuce which is why we favored the iceberg lettuce.; other things: zucchini was half of the time misclassified as cucumber, similarly

for pickle, traffic cone had a lot of different labels, such as simply cone, caution cone, hazard cone, safety cone

2.3 Norming for typicality

Goal: Does the color manipulation of the images show the desired difference in typicality ratings?

Task: "How typical is this object for a **NOUN**?"

slider rating, underlyingly coded as ranging from 0 to 100

45 trials; 11 color diagnostic objects, each in their typical color and 1-2 atypical colors (i.e., 25 stimuli); 20 non-color diagnostic stimuli

30 participants; exclusions: none; everyone thought they did the HIT correctly

Results: generally clear distinction between typical and atypical instance; From the three items that were normed in two atypical colors (carrot, corn, pumpkin), we see the biggest difference between the red and white pumpkin. Therefore, we should choose the white pumpkin and (following from that) the green carrot and red corn. There does not seem to be a big difference between the yellow egg and snowman, but the white egg is rated even more typical and its size fits better to the other stimuli. Therefore, we should choose the egg over the snowman (given that both are also nameable). Even though the orange banana is predominantly rated below 50, it is still not as atypical as other objects.; The non-color diagnostic objects are all rated as very typical instances.

2.4 Norming for free production

Goal: Are the image depictions we chose nameable, the way we intended?

Task: "What is this?"

free production

31 trials (22 cd – each participant saw one instance of each object at random, i.e., either typical or atypical; 20 non-cd)

Results: swan is often called a goose; two people identified the white carrot as parsnip

2.5 Final stimuli selection

In the end, we have 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. This counterbalance aims to reduce artifacts of salience such as red is generally more salient as a warn signal [ek: ref?] and blue is highly atypical for most objects. A full list of stimuli can be found in table [ek: add table and reference].

3 Experiment: Production

3.1 Method

Participants.

Procedure.

Materials.

Data Preprocessing and exclusion.

3.2 Results

4 Experiment: Comprehension

4.1 Method

Participants. We recruited 80 participants over Amazon's Mechanical Turk, 40 for each color competitor typicality manipulation. The study took on average 7 minutes and each of them were paid [ek: ...] for their participation. We restricted participation to workers with IP addresses in the US and a approval rate of previous work above 97%.

Procedure. This experiment is a one-player adaptation of the production study explained in [ek: ref to section] and follows the design of an incremental decision task [ek: cite Qing]. participants were put into the listener role of the reference game. That means, they needed to identify which object was the target given a referring expression placed above the grid. Crucially, they do not observe the complete referring expression at once, but instead the utterance is gradually revealed. After new information is

revealed, participants are required to make their best guess onto which object is the most likely target. The choices had to be made prior any disambiguating information (after observing "Click on the"), after observing an adjective ("Click on the yellow") and after observing the fully disambiguating noun ("Click on the yellow banana!"). The clicks before information are observed are useful to determine whether there are already strong priors in item selections, even before any information were observed. The clicks after revealing the adjective are the critical clicks that will affect our interpretation of inferences drawn from the adjective. After observing the noun there is only one possible referent left. These clicks are used as attention checks.

Participants completed 55 trials in total, 20 of which were critical trials and 35 were fillers. The filler trials were supposed to ensure that participants perceive the referring expressions as being generated by a natural speaker. Firstly, all target trials are color modified utterances. To avoid that participants learn that the color modifier is always part of the referring expression, we need utterances that only have the bare noun. Second of all, we need to make sure that targets are not logically derivable. In a context such as [ek: Fig ref!], the target can be determined by reasoning about the distractor choice. The target is the only object that shares the type (i.e., banana) with a distractor and its color (i.e., yellow) with another distractor. One can then derive that the object that is being distracted from will be the target. This regularity could also be learnt over the time course of the experiment. The filler trials therefore need to introduce primarily unmodified referring expressions that target other objects in the display. The exact trial structure is summarized in table 1.

Before participants proceeded to the main trials, they had to complete four practice trials constructed from the speaker perspective. In the speaker role, they saw a grid of four non-color diagnostic objects, one of which was marked as the target by a green border surrounding it. They were then asked to refer to the object such that a second player could identify it. The practice trials were introduced to familiarize the participants with the task.

The main trials were randomized with one restriction: Trials in which we expect a color

trial type	number	utterance	referent
critical	20	modified	target
filler	5	unmodified	color competitor (typical)
filler	5	modified	color competitor (atypical)
filler	5	modified	contrast
filler	20	unmodified	distractor (typical)

Table 1

Overview of the trial structure for the comprehension study.

modifier to be superfluous or even misleading only occurred after the 15th trial. These were contexts where there was no contrast and either both target and color competitor typical objects, or the target typical while the color competitor was atypical. This measure should minimize the risk that participants perceive the "utterance generator" as unnatural.

Materials. [ek: add choice of contexts (i.e., one per color,...); clarify what is within and between-subject manipulation with rationale]

The stimuli were the same as in the production study.

Data Preprocessing and exclusion. We excluded participants who indicated that they did the Hit incorrectly or were confused (7), who indicated that they had a native language other than English (4), who gave more then 20% erroneous responses (2) and who did the Hit multiple times (1). Overall, we excluded 14 out of 80 submissions (17.5%). An erroneous response is defined as a click to a non-target object after observing the fully disambiguating noun, i.e., participants are excluded who selected the wrong final object more than 11 times.

4.2 Results

5 Discussion

Summary.

5.1 Conclusion

6 Appendix

The appendix could not be included, due to file size restrictions. If you are interested in the full version of the thesis, I am happy to provide it.

7 Acknowledgments

First and foremost, I would like to thank Judith Degen for her scientific guidance in all respects and also for her personal support throughout the whole project.

I would also like to thank the Computation and Cognition Lab at Stanford University for hosting me, which enabled this project in the first place. I am especially thankful to Robert X. D. Hawkins and Noah D. Goodman who were always available for scientific and personal advice and discussions.

Finally, I'm grateful for the continuous support from Eleni Gregoromichelaki who provided place for project-related exchanges with other students and whose belief in me was a constant motivation.

8 References