Using contrastive inferences to learn about new words and categories

Claire Bergey and Daniel Yurovsky

S1 Experiment 1

In addition to the analyses reported in the main text, we ran a pre-registered linear mixed effects model predicting target choice from the presence of an adjective in the utterance, the adjective type (size or color), and the display type (unique target display or contrastive display). People were more likely to choose the target if there was an adjective in the utterance ($\beta_{adjective} = 2.21$, $t = 7.18$, $p = < .001$), and were more overall likely to choose the target on unique target trials ($\beta_{unique} = 3.81$, $t = 10.68$, $p = < .001$). There was an interaction between the presence of an adjective and the type of adjective, such that people were especially likely to choose the target when there was a size adjective in the utterance ($\beta_{adjective*size} = 0.95$, $t = 2.18$, $p = .029$). There was a three-way interaction between the presence of an adjective, the type of adjective, and the search type such that the contrastive strength of size over color was weaker in the unique target trials than in the contrastive trials ($\beta_{adjective*size*unique} = -3.06$, $t = -2.61$, $p = .009$),

Figure S1 shows referent choice in both the unique target display trials and the contrastive display trials. Unique target displays had one unique referent (the target) and two identical distractors that differed from it both in shape and the critical feature. Contrastive displays had a target, a contrastive pair which matched the target in shape but had a different critical feature, and a lure which matched the target on the critical shape but differed from it on the critical feature.

Table S1: Full model of target choice from Experiment 1.

term	estimate	std.error	statistic	p.value
	-2.07	0.26	-7.94	< .001
adjective (vs. no adjective)	2.21	0.31	7.18	< .001
size (vs. color)	-0.17	0.36	-0.46	.646
unique target display (vs. contrastive display)	3.81	0.36	10.68	< .001
adjective * size	0.95	0.44	2.18	.029
adjective * unique target	1.32	1.09	1.22	.223
size * unique target	-0.17	0.46	-0.37	.709
adjective * size * unique target	-3.06	1.18	-2.61	.009

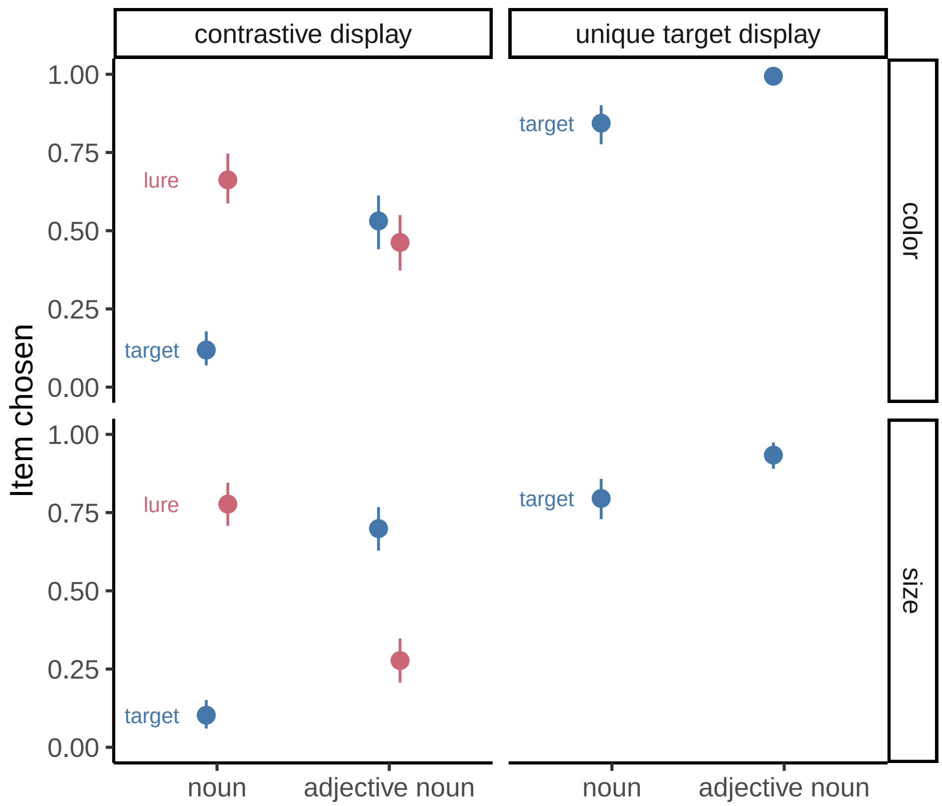


Figure S1: Referent choice in both the contrastive display trials and the unique target display trials.

S2 Experiment 2

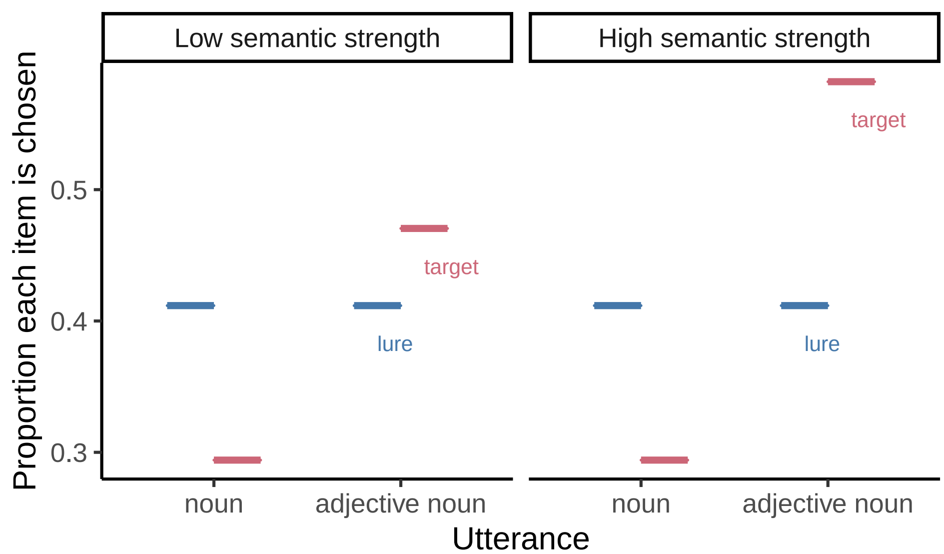


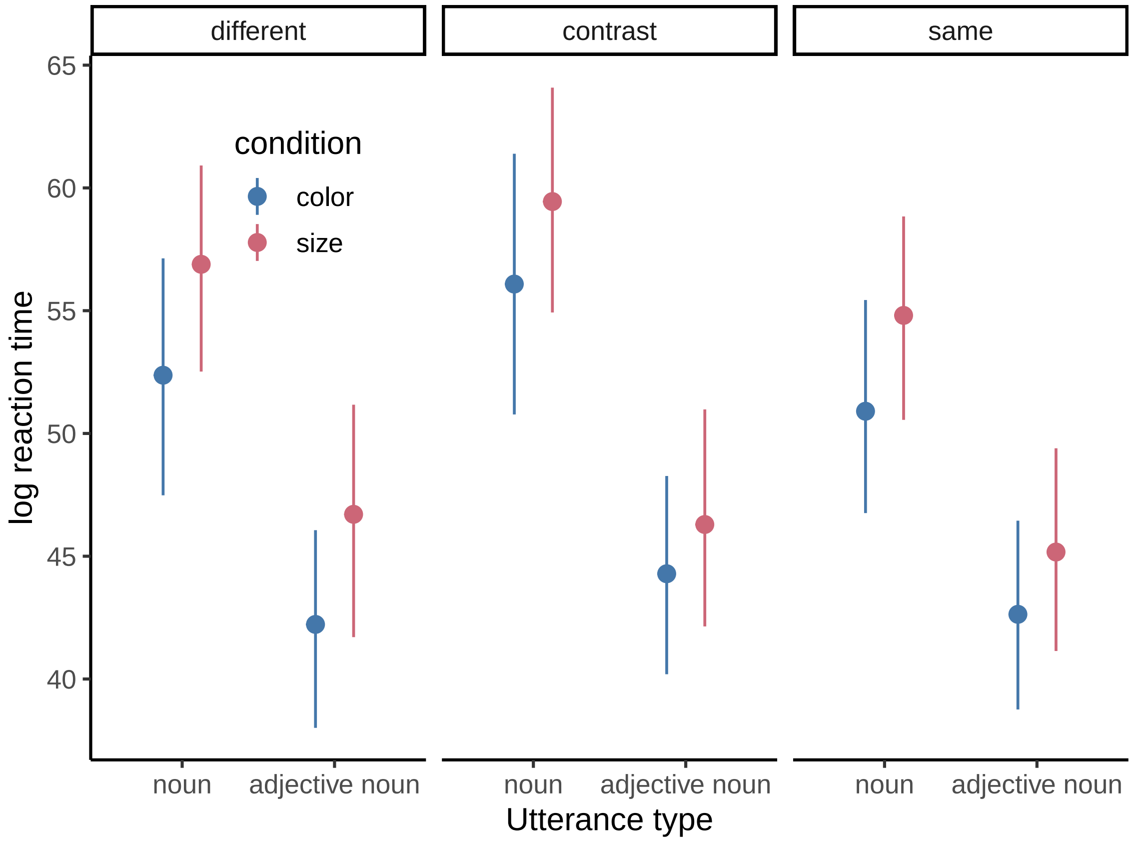
Figure S2:

Degen, Hawkins, Graf, Kreiss, & Goodman (2020) capture asymmetries in description of size and color by positing that different features have different semantic strength. They posit that color has stronger semantics than size, such that "red table" is a better literal description of a small red table than "small table" is. Under these assumptions, RSA using these continuous semantics explains people's tendency to mention color more often than size in a variety of tasks. Can their model explain the asymmetry we find between color and size in Experiment 1?

In Experiment 1, we found that people more consistently choose the target using contrastive inferences about size than color. We incorporated their continuous semantics into our RSA model of referent choice, which reasons over possible lexicons. In Figure S2 we show the difference in referent choice when a feature has low semantic strength (0.8) compared to high semantic strength (0.99). A feature with low semantic strength results in a weaker contrastive inference (reduced choice of the target in the *adjective noun* trials) compared to a feature with high semantic strength. Degen et al. (2020) find that color has stronger semantics than size, which would result in a stronger contrastive inference about referent choice when color adjectives are used. This is not what we find: people make stronger contrastive inferences about referent choice when *size* adjectives are used. Thus, while a model with continuous semantics could in principle explain the asymmetry we find, it would need to have stronger semantic values for size than color. We note that while the same continuous semantics do not explain both our data and the production data from Degen et al. (2020), neither does the model we propose explain the production data. We leave it to future work to form a more complete account of color-size asymmetries in both production and comprehension.

S2.1 Experiment 2 Reaction Times

Thus, participants treated all adjectives as marked, and inferred lower typicality, regardless of whether they could felicitiously be interpreted as contrasting between potential target referents. But were participants nonetheless sensitive to this information in their response times? We investigated this question by analyzing participants' time to advance after seeing the aliens' referential exchange. Though this task was not speeded, we hypothesized that participants would advance more quickly after seeing referential exchanges that were easier to process. After dropping all response times less than 1 second and longer than 10 seconds, and log transforming them because of the right skew in response time data, we predicted participants' time to advance on each trial of the experiment from utterance type, context type, critical adjective type, and the interaction between utterance type and context type (`log(rt)` $\sim$ `adjective * search + type + (1 |subj)`). This model showed a reliable effect of utterance type ($\beta_{adjective} =$, $t =$, p)–participants were faster when an a descriptor was provided despite having to process an additional word. There was no main effect of critical adjective type ($\beta_{size} =$, $t =$, $p =$), nor context type ($\beta_{contrast} =$, $t =$, $p =$; $\beta_{same} =$, $t =$, $p =$), but the interactions between utterance type and context type trended towards significance for both non-contrast searches ($\beta_{adjective*contrast} =$, $t =$, $p =$; $\beta_{adjective*same} =$, $t =$, $p =$). Directionally, these results indicate that participants took longer to process utterances which were under-described (within-category contrast trials with no adjective) than those with appropriately no description, and processed trials with an appropriate level of description (contrast trials with an adjective) more quickly than those with superfluous description.



S3 Experiment 3

In addition to the regressions reported in the manuscript, we two pre-registered, targeted regressions to test the effect of utterance type to more specifically in case these effects were unclear in the maximal models. First, we filtered to adjective and no adjective trials and fit a linear mixed effects model predicting prevalence judgment by utterance type with a random slope of utterance type by subject. Participants' prevalence judgments were significantly lower when an adjective was used in the utterance (β = -9.17, t = -7.09, p = $<$ .001). Second, we included all trials in a linear mixed effects model predicting prevalence judgment by utterance type with a random slope of utterance type by subject. Utterances without an adjective resulted in significantly higher prevalence judgments than alien utterances (β = 7.76, t = 4.91, p = $<$ .001), and utterances with an adjective did not result in significantly different prevalence judgments than alien utterances (β = -1.42, t = -0.91, p = .363).

References

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review*, *127*, 591–621.