Title (required)

Provide the working title of your study. It may be the same title that you submit for publication of your final manuscript, but it is not a requirement.

The role of syntactic modification for comparison class inference

Description (optional)

Please give a brief description of your study, including some background, the purpose of the study, or broad research questions.

In order to interpret a sentence like “That’s a big Great Dane” listeners have to establish a comparison class - a reference group against which the referent is implicitly compared with respect to the property described by the adjective (e.g., size: compared to what is the Great Dane big?).

Yet, the intended comparison class usually goes unsaid and the listener is left to infer the correct comparison class from a variety of a priori possible comparison classes (e.g., big compared to other animals, to other dogs, to other Great Danes etc.). This study examines whether listeners are sensitive to syntactic modification of the noun by the critical adjective when trying to infer the comparison class.

In prior studies we showed that listeners flexibly adjust the inferred comparison class given simple *Subject Predicate* utterances which differed syntactically (e.g., “That Great Dane is big” vs. “That’s a big Great Dane”) and described a referent appearing in different contexts (Tessler, Tsvilodub, Snedeker and Levy, 2020, https://doi.org/10.31234/osf.io/n8eyj). In particular, in a set of experiments we investigated the role of the noun position in such sentences, hypothesizing that listeners interpret utterances by reasoning about rational speakers, who choose utterances optimally accomplishing their communicative goals - i.e., the goals of reference (helping the listener establish the correct referent) and predication (communicating a property of the referent). To this end, the position of the noun is a cue to the intended comparison class - when the noun appears in the predicate of the sentence (e.g., “That’s a big Great Dane”) it can naturally be explained as intended to constrain the comparison class; however, when the noun appears in the subject (e.g., “That Great Dane is big”), it can potentially be explained away as aiding the listener to establish reference (especially combined with the dectic ‘that’), and hence is a weaker cue to the intended comparison class (Tessler et al., 2020). Therefore, when interpreting the noun of an utterance listeners evaluate its role with respect to the comparison class by trading off the utility of the noun for reference and predication.

In this experiment, we aim to disentangle the effect of noun position (subject noun vs. predicate noun) from syntactic modification (direct modification: e.g., “big Great Dane”, vs. no modification: “The Great Dane is big”) by looking at *directly modified* nouns appearing either in the predicate or in the subject of the sentence. Consistent with our reference-predication trade-off hypothesis, we predict that even directly modified nouns appearing in the subject of the sentence are more likely to be produced by a speaker as intended for reference, and hence provide a weaker cue to the intended comparison class.

In contrast, directly modified nouns appearing in the predicate should constrain the comparison class more strongly.

Hypotheses (required)

List specific, concise, and testable hypotheses. Please state if the hypotheses are directional or non-directional. If directional, state the direction. A predicted effect is also appropriate here. If a specific interaction or moderation is important to your research, you can list that as a separate hypothesis.

In this experiment, participants have to infer the comparison class and provide it via a free production paraphrase task, given sentences containing an adjective describing a subordinate referent (e.g., a dog of a specific breed, like a Great Dane) in a basic-level context (a group of different members of the same basic-level category as the referent, e.g., a group of dogs of different breeds).

In the critical condition, the critical noun of the sentence is the subordinate-level label of the target referent, which is directly modified by a size adjective consistent with general expectations about the referent's size relative to its basic-level category (i.e., Great Danes are always described as big, and pugs are always described as small). This adjectival phrase appears either in the subject or in the predicate of the sentence (within-subjects). Further, the sentence contains a second noun describing a visually distinctive feature of the referent, appearing in the syntactic position, complementary to that of the critical noun. For example, the referents for the dog-category have prize-bows on them, marking them as 'prize-winners', supported by contextual set-up. The context picture contains other instances having the same feature.

Resulting critical sentences have the form "That ADJ N1 is N2" (for example "That big Great Dane is a prize-winner", Subject N condition), or "That N2 is a ADJ N1" (for example, "That prize-winner is a big Great Dane", Predicate N condition).

The filler condition is an exact replication of the basic-level context subordinate noun condition of Experiment 3 in Tessler, Tsvilodub, Snedeker and Levy (2020).

Participants will be asked to fill-in the inferred comparison class in the sentence "It is {big, small} relative to other __". Their responses will be classified into subordinate comparison classes (i.e. matching the critical label noun N1 modified by the adjective), and comparison classes non-matching the critical noun, i.e., subsuming superordinate, basic-level categories, as well as the categories communicated by the second noun N2 in the sentence (e.g., "prize-winners").

We hypothesize that nouns in the predicate position restrict the comparison class more strongly, so that in the critical condition utterances like "That prize-winner is a big Great Dane" are more likely to signal that the referent is big for a Great Dane than utterances like "That big Great Dane is a prize-winner". For latter sentences, comparison class inference may be driven by the second noun of the sentence, or pragmatic inferences driven by factors like the context or world knowledge.

Furthermore, we expect a similar effect in filler trials, such that sentence with a predicate noun like "That's a big Great Dane" are more likely to communicate the comparison class, compared to subject noun sentences like "That Great Dane is big".

Therefore, we hypothesize a higher proportion of responses matching the critical subordinate noun in the predicate noun condition, compared to the subject noun condition of the critical trials. We further predict a higher proportion of responses matching the noun in the predicate filler condition, compared to the subject filler condition.
Hence, we crucially predict an effect of syntax (subject vs. predicate) in the critical condition, spelled out as as the following Bayesian generalized linear mixed-effects model estimate being credibly different from zero (see below for coding scheme):
2*beta_syntax - 2*beta_syntaxXtrial_type > 0
We do not predict a main effect of trial type (filler vs. critical). We include an interaction between syntax and trial-type, but remain agnostic the direction of this estimate.

Study design (required)

Describe your study design. Examples include two-group, factorial, randomized block, and repeated measures. Is it a between (unpaired), within-subject (paired), or mixed design? Describe any counterbalancing required. Typical study designs for observation studies include cohort, cross sectional, and case-control studies.
This study is a two-by-two within-subjects free production task, conducted as a web-based experiment. Participants see a context picture and read a sentence about a referent which they have to paraphrase, completing a total of eight main trials, presented in two blocks of four trials each. The eight main trials consist of four critical and four filler trials, where each trial is a condition resulting from a unique combination of the noun position condition (subject N vs. predicate N) crossed with the size of the referent within its basic-level category (e.g., large vs. small subordinate dog category). Five different contexts sampled from four different basic-level categories are used: dogs, flowers, birds and trees; for each basic-level context, there are two possible targets representing a large-subordinate and a small-subordinate category, respectively (e.g., for the dog category, potential targets are a Great Dane and a pug).
The experiment consists of two blocks. In each block, participants first complete warm-up trials. In the first part of the block, participants first complete two rounds of labeling warm-up trials. A round consists of a demonstration trial where participants see two distinct subordinate members of a basic-level category used in this block and read their labels. For example, for the flowers-category they see pictures of a sunflower and a daisy next to each other and read “This is a sunflower” and “This is a daisy”, respectively. They can proceed after 3.5 seconds to the next trial where they have to label other instances of the same categories themselves. They also provide a common label for the pictures (i.e., “flowers”). The order of the pictures is randomized. They are provided feedback on their labels and can proceed only after correcting their labels, if they were incorrect. After two labeling warm-up rounds, participants complete two demonstration trials of at least 3.5 seconds each learning about the additional features of the referents described by the second noun of the critical sentences in the main trials. For example, participants see a picture depicting the sunflower and the daisy in pots with bows, and read: “These flowers are gifts. Notice the bow on the pots.” (‘gift’ being the additional noun). Finally, participants complete a comparison class paraphrase practice trial given the following scenario: *“Speaker A: ‘The Empire State Building is tall’. What do you think Speaker A meant? ”*; participants provide their answer in the blank of the sentence *“The Empire State Building is tall relative to other __”*. A correct

answer can be one of ‘buildings, skyscrapers, houses, constructions’. Participants have to correct their answer to one of these options before proceeding if it was incorrect.
The warm-up trials in the second experimental block are identical, but there is no paraphrase practice trial.
Then, participants complete four main trials in each block - two main and two filler trials, in randomized order, where a filler trial is always the first trial of the block. In the critical main trials, a subordinate referent with an additional feature (e.g., a gift bow) appears in the corresponding basic-level context. Participants read different context stories for each context. For example, for the flower context, they read “You and your friend are at their garden and you see the following:” above the context picture. The context picture consists of six different representatives of the same basic-level category as the target, including two other members of the same subordinate category as the referent, and two other individuals with the feature described by the second noun of the sentence, e.g., in the flower-context there are two other ‘gifts’. This set-up ensures that both nouns in the critical sentences have identical referential utility in the given context.
Below, on each trial they read “Your friend runs far ahead of you. You see your friend in the distance:”, followed by a depiction of the referent with the additional feature next to a person; to induce the illusion of distance, both are small relative to the context picture. Then they read “Your friend says:”, followed by the critical sentence (“That ADJ N1 is N2” or “That N2 is a ADJ N1”, within-subjects). Finally, they are asked: “What do you think your friend is saying it is {big, small} relative to?”, introducing the paraphrase template.
For a given basic-level category, one of the possible targets appears in this critical trial (e.g., the large-subordinate sunflower). The other possible target (i.e., the small-subordinate dandelion) then appears in a filler trial in the same block.
Filler trials are identical to main trials with basic-level contexts and subordinate nouns from experiment 3 in Tessler et al., 2020 (https://doi.org/10.31234/osf.io/n8eyj). Participants read “You and your friend see the following:” above a context picture, identical to the contexts used in filler trials, modulo the additional visual features (like gift bows). Below, the read “Your friend goes ahead of you. You see your friend in the distance:”, introducing the picture of the target referent next to a person. Again, the depiction is small relative to the context picture in order to induce the distance-illusion. Then, the critical sentence appears below “Your friend says:”. The paraphrase prompt is identical to the critical trials.

The size of the referent (i.e. large-subordinate vs. small-subordinate) is counterbalanced across syntactic conditions (subject vs. predicate N) and trial types (critical vs. filler) within-participant, resulting in 8 unique conditions. Each participant sees each condition once, resulting in eight main trials.

After the main trials, participants complete an optional post-test socio-demographic questionnaire.

Randomization (optional)

If you are doing a randomized study, how will you randomize, and at what level?

The order of the syntactic conditions (subject vs. predicate N), the order of the trial-types (critical vs. filler) and the size of the target (small-subordinate vs. large-subordinate) are

randomised within-subjects, as described above. The order of the basic-level contexts is also randomized.

Data collection procedures (required)

Please describe the process by which you will collect your data. If you are using human subjects, this should include the population from which you obtain subjects, recruitment efforts, payment for participation, how subjects will be selected for eligibility from the initial pool (e.g. inclusion and exclusion rules), and your study timeline. For studies that don't include human subjects, include information about how you will collect samples, duration of data gathering efforts, source or location of samples, or batch numbers you will use.

Participants will be recruited on Amazon's Mechanical Turk and compensated $1.00/participant. Only participants with US IP addresses are eligible. Subjects will be excluded if they report a native language other than English. Further, participants are required to answer a comprehension question. In the comprehension scenario, a named speaker (e.g., Alice) talks to a named listener (e.g., Bob) about the weather. Participants must then type in a box to whom the speaker is talking (i.e., Bob). Actual names are randomly sampled from a list of common names. Maximally three attempts are possible. Participants failing this "captcha" won't be able to continue to the main experiment.

Sample size (required)

Describe the sample size of your study. How many units will be analyzed in the study? This could be the number of people, birds, classrooms, plots, interactions, or countries included. If the units are not individuals, then describe the size requirements for each unit. If you are using a clustered or multilevel design, how many units are you collecting at each level of the analysis?

Based on a power analysis requiring a power of 0.85, 300 subjects passing our exclusion criteria will be recruited.

Measured variables (required)

Describe each variable that you will measure. This will include outcome measures, as well as any predictors or covariates that you will measure. You do not need to include any variables that you plan on collecting if they are not going to be included in the confirmatory analyses of this study.

We record the responses provided by the subjects in the main trials. Further, we record the syntactic condition (subject vs. predicate N position), the trial type (critical vs. filler) and the target of each trial. In the warm-up trials, we record the responses and the number of attempts needed per warm-up trial until the correct labels are entered.

Additionally, participants' native language is recorded (if provided).

Statistical models (required)

What statistical model will you use to test each hypothesis? Please include the type of model (e.g. ANOVA, multiple regression, SEM, etc) and the specification of the model (this includes each variable that will be included as predictors, outcomes, or covariates). Please specify any interactions, subgroup analyses, pairwise or complex contrasts, or follow-up tests from omnibus tests. If you plan on using any positive controls, negative controls, or manipulation checks you may mention that here. Remember that any test not included here must be noted as an exploratory test in your final article.

We will fit a Bayesian logistic mixed-effects regression model using the R brms package (Bürkner, 2017).

We will regress the response category (matching vs. non-matching the subordinate target-label of the stimulus sentence) against the fixed effects of syntax (subject vs. predicate N position), trial type (critical vs. filler) and their two-way interaction. We will include maximal random effects structure licensed by our experimental design: random intercepts, random slope effects of syntax, trial type and their interaction by-participant and by-target. The data will be collapsed across target referent size. Default priors will be used.

The model in R brms-syntax is:

response ~ syntax * trial_type + (1 + syntax * trial_type | subject_id) + (1 + syntax * trial_type | target)

Transformations (optional)

If you plan on transforming, centering, recoding the data, or will require a coding scheme for categorical variables, please describe that process.

Participants' responses will be categorized as matching or non-matching the subordinate target label in the stimulus sentence. All valid subordinate target labels (e.g. Great Dane, sunflower etc., identifiable as correctly referring to the target in the given context) will be categorised as 'matching'; compound responses wherein the head noun is the valid subordinate target label will also be categorized as 'matching' (e.g., "prize-winning Great Dane"). All other responses will be categorized as "non-matching" the critical noun. The noun number as well as misspellings will be disregarded.

The categorical predictors will be sum-coded (syntax: subject N = 1, predicate N = -1, trial type: filler = 1, critical = -1).

Inference criteria (optional)

What criteria will you use to make inferences? Please describe the information you'll use (e.g. specify the p-values, Bayes factors, specific model fit indices), as well as cut-off criterion, where appropriate. Will you be using one or two tailed tests for each of your analyses? If you are comparing multiple conditions or testing multiple hypotheses, will you account for this?

We will analyse the 95% credible intervals of the effects in the Bayesian model. If the CI excludes 0, we will consider this evidence that the respective effect is credibly different from 0.

Data exclusion (optional)

How will you determine which data points or samples if any to exclude from your analyses? How will outliers be handled? Will you use any awareness check?

Data of participants who report technical issues in the comments (e. g. distorted views) will be excluded from the analysis.

Data from participants who report a native language other than English will be excluded.

Participants who need more than 4 attempts on any warm-up trial will be excluded from the analysis.

Invalid responses (e. g., labels that cannot be identified as referring to the target referent) will be excluded.

Missing data (optional)

How will you deal with incomplete or missing data?

If data isn't recorded for some but not all of the trials for a participant, we will include the trials from that participant for which we have data.

Files:

https://github.com/polina-tsvilodub/refpred/blob/master/analysis/direct-modification/direct-modification-prereg-final.Rmd