Degree distribution in the comparison class explains the absolute-relative distinction for gradable adjectives

Alexandre Cremers
Vilniaus Universitetas

Gradable adjectives, scales, and comparison classes

The class of gradable adjectives can be divided between *relative* adjectives, such as 'tall' or 'far', which are highly context-dependent and vague, and *absolute* adjectives, the meaning of which is much more rigid (Unger, 1971; Bolinger, 1972). Absolute adjectives are further divided between *minimum standard*, such as 'dangerous', and *maximum standard*, such as 'dry'. Informally, the former conveys that an object presents at least some danger, while the latter conveys that an object is fully dry.

Kennedy & McNally (2005) argue that these distinctions stem from differences in the structure of the scales to which these adjectives refer. Relative adjectives map individuals onto *open* scales, with no definite boundaries, while absolute adjectives map individuals onto *closed* scales. Whether the closed scale is upper- or lower-bound further distinguishes between minimum and maximum-standard absolute adjectives. Scales that are fully closed (have both an upper and lower bound), give rise to maximum-standard adjectives.

For Kennedy & McNally (2005), these distinctions are a matter of lexical semantics: the adjective encodes the type of scale to which it maps entities (as the range of the measure function it denotes), thereby determining its class. The class, in turn, affects other lexical properties of the adjective, such as the modifiers it can combine with. For instance, only maximum-standard adjectives can combine with adverbs such as 'completely' or 'almost' (which make reference to an endpoint). Note that if an adjective is minimum-standard, its antonym will be maximum-standard, while relative adjectives do not combine with these adverbs, and neither do their antonyms (since they are relative too).

The lexically-encoded scale does not always match the scale we would intuitively associate to an adjective. For instance, the cost scale associated with 'cheap/expensive' has a clear minimum, free items. Yet, 'completely cheap' sounds deviant and, to the extent that we would accept it, would not intuitively mean 'free'. This suggests that the underlying scale determined by the adjective is not our intuitive notion of cost, lower-bound by zero, but a more abstract scale with no lower end (e.g., a logarithmic scale). While this observation can at first be seen as an argument in favor of the lexical semantics idea, it may actually threaten the whole enterprise. If apparent exceptions to the rule that scale boundaries determines the class of the adjective can be circumvented by postulating *ad hoc* scales, the whole proposal may become circular. Wellwood (2020) proposes to save the lexical semantics approach from unfalsifiability using a two-stage system of semantic interpretation.

Alternatively, recent Bayesian pragmatics accounts of gradable adjectives, while drawing much of their inspiration from Kennedy & McNally (2005) and subsequent work, offer a competing view in which the comparison class, rather than the lexical semantics of the adjective, fixes the properties of the scale, and thereby determines the class of the adjective (on a case-by-case basis). The central idea of Bayesian pragmatics is that listeners interpret utterances by updating their prior beliefs with the information provided by the speaker (Frank & Goodman, 2012). Lassiter & Goodman (2013) propose to model scale boundaries with prior beliefs where significant probability mass is located at one or the other end of the range of degrees. The adjective only provides a measure function (i.e. a function from entities to degrees), and prior beliefs about these entities is what ultimately determines

whether the resulting scale is open or closed. Coming back to the 'expensive/cheap' example, if we are discussing the purchase of a new fridge, we would typically consider the range of prices for new fridges, which clearly does not extend all the way down to the theoretical lowest price of zero. In this framework, theoretical boundaries on a scale (the range of the measure function denoted by the adjective) are irrelevant; what matters is the distribution of degrees in the comparison class (i.e. the image of the comparison class by the measure function).

Previous experimental work (Schmidt et al., 2009; Solt & Gotzner, 2012; Qing & Franke, 2014b) shows that the distribution of degrees within a comparison class indeed affects the threshold for the adjective, but the explicit link between closed scales and absolute adjectives remains untested. Meanwhile, most modelling work follows Lassiter & Goodman (2013) in assuming that the prior distribution alone determines the class of the adjective (Qing & Franke, 2014a; Tessler et al., 2017; Bennett & Goodman, 2018). We propose an experiment to adjudicate between the Bayesian pragmatics view, in which the distribution of degrees in the comparison class is sufficient to determine the class of an adjective, and the lexical semantics view, which stipulates that the theoretical boundaries of the scale are the deciding factor (even if the comparison class does not reach these boundaries).

Experiment

We tested the interpretation of nonce adjectives in the presence of explicit comparison classes which each comprised 20 planets, for which we gave fictional measurements of the dimension measured by the adjectives. The use of nonce adjectives ensured that only information about the scale and the comparison class could determine whether an adjective is absolute or relative. All measurements were expressed in percentages (thereby fixing clear theoretical boundaries for all scales), and the 20 planets in the comparison class corresponded to the 21-quantiles of a beta-distribution with possible inflation in 0 or 1 to represent closed scales. We tested 4 types of comparison classes (with probability distribution corresponding to lower-, upper-, double-bound and open scales).

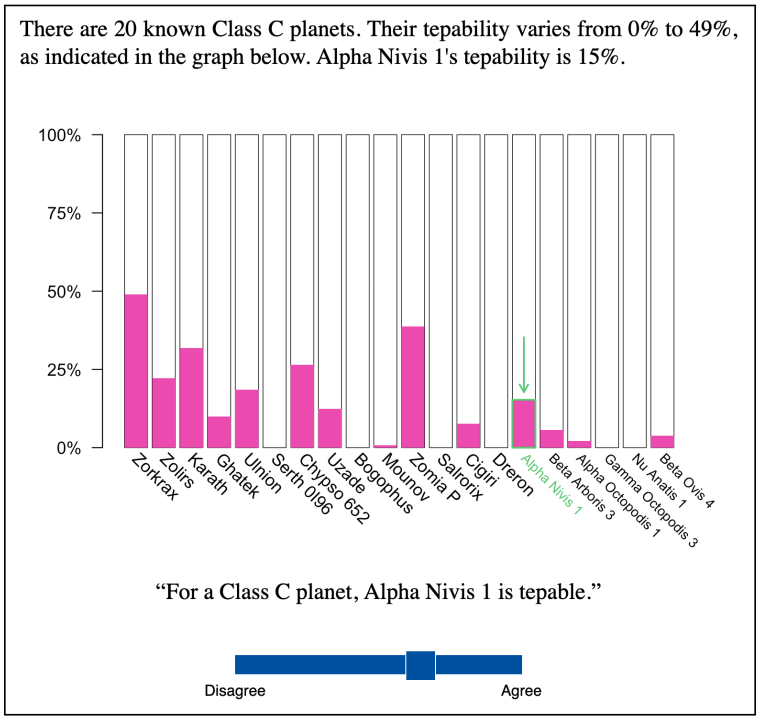

Figure 1: Example trial with a lower-bound scale (many items are at or close to 0%).

In each trial, participants were asked to judge the applicability of a predicate containing an

adjective to an element from the comparison class, using a continuous slider as shown in Figure 1. The slider followed the cursor and its position was recorded on the first click to make the task less tedious. For each comparison class, we tested the applicability of the predicate to half the elements in the comparison class, and its negation for the other half (randomly selected as odd and even quantiles). We tested 3 bare (positive form) adjectives, as well as 4 adjectives modified by 'very', 'extremely', 'absolutely', and 'quite'. Each construction could appear in affirmative of negated form. One more adjective appeared with 'a bit' in affirmative sentences and 'at all' in negative sentences. Each participant saw 8 scales (with a random combination of the nonce-adjectives and constructions), presented in 16 blocks of 10 items (affirmative and negative forms separated to avoid confusion). The blocks were presented in random order, and items within each block were also randomized. The parameters of the degree distributions were randomly drawn so that each participant saw 2 unique scales of each type. The association between scales, adjectives, and constructions was randomized.

We recruited 110 participants on MTurk, paid $2 each (the survey took about 10min). 10 participants whose median RT was below 1s were excluded. We further removed the 5% fastest responses (threshold: 908ms). Finally, we fitted linear regressions of acceptability by degree (flipped for negative sentences) and removed blocks where the regression coefficients was more than 1SD below the mean (threshold: $-.45$). The goal was to remove cases where participants missed a change of polarity between two blocks, and resulted in the removal of 6% remaining affirmative trials and 10% negative trials. In all, we filtered out 20% of the initial data set.

Results

We first tested how negation affected the results by fitting sigmoid functions with optional censoring at scale ends to each block, and compared the midpoints for pairs of affirmative and negative blocks (excluding the 'a bit/at all' cases). We found no significant differences ($t(461) = 0.50, p = .62$), confirming that negation does not shift the threshold for the adjective but only flips acceptability, in line with previous empirical findings (Hersh & Caramazza, 1976; Leffel et al., 2019). In the rest of the analyses, we pool data from affirmative and negative blocks under the assumption that $\text{Acc}(\neg S) = 1 - \text{Acc}(S)$. From now on, we focus on the bare adjectives only.

In order to diagnose absolute interpretations, we calculated the slopes of the acceptability increase between the two lowest and between the two highest degrees in each scale. Minimum-standard adjectives would have a clear gap in acceptability at the bottom of the scale, since the threshold should most likely be located right above the minimum degree. For maximum-standard adjectives, we expect a gap at the top of the scale, as the threshold should sit right below the maximum of the scale. Relative adjectives should be flat at both extremities of their degree distribution, since their threshold should be vague and most likely situated slightly above the middle of the scale. When a scale had more than one item at a given degree (i.e., on closed scales boundaries), the slope was computed based on the mean acceptability across these items. Negative slopes were filtered out. Figure 2 shows the bottom and top slopes for each category of scale we tested. Mixed-effects regressions on log-slopes with random subjects intercept and open scales as the reference revealed a significantly higher bottom slope for lower-bound and double bound scales (both $p < .001$), but not upper-bound ($p = .25$, flatter than relative if anything), while upper-bound and double bound scales had a significantly higher top slope than open scales (both $p < .001$) but lower-bound scales didn't ($p = .80$).

Lastly, we tested different quantitative models that have been proposed to capture effects of comparison class on adjectives. We included the best two models from Schmidt et al. (2009) (RH-R and CLUS), which predate the Bayesian pragmatics approach, the RSA model of Lassiter & Goodman (2013) and the Speaker-Oriented Model (SOM) of Qing & Franke (2014a). The RH-R is a very simple model, which assumes that the adjective is true of a fixed proportion of the degree range, with Gaussian noise around the degree that realizes this proportion. We assumed that both parameters of

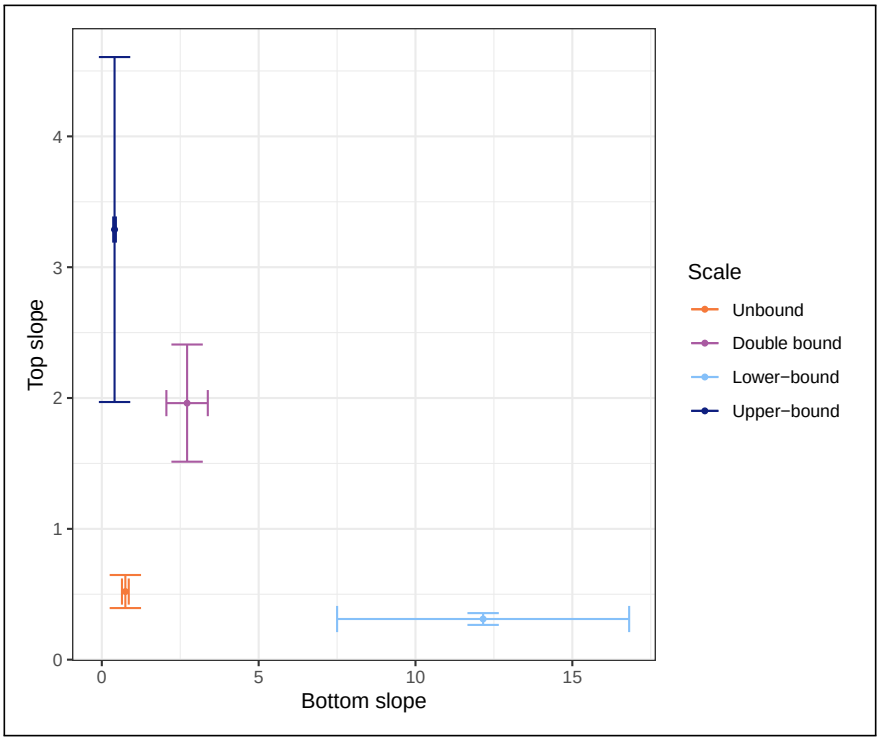

Figure 2: Slope between the two highest degrees as a function of slope between the two lowest degrees, by scale type (mean and standard deviation).

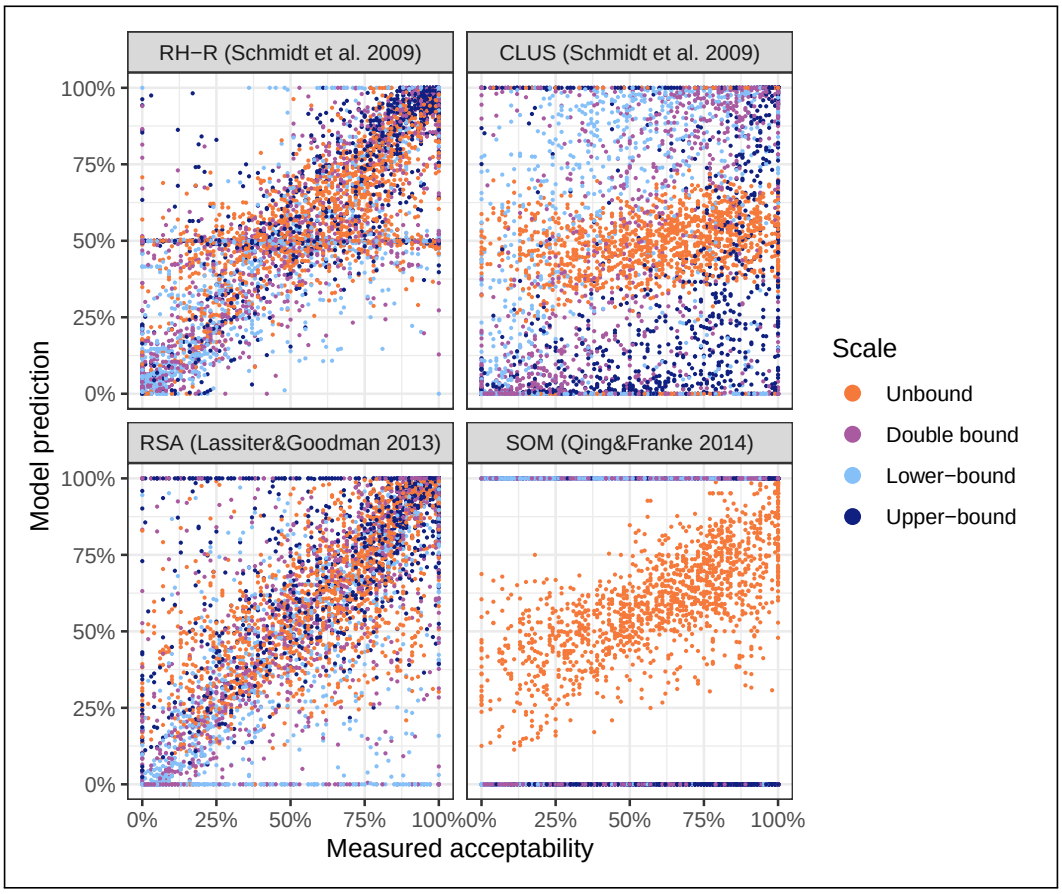

Figure 3: Predictions of each model against the data.

the model (the proportion of degrees which validate the adjective and the noise parameter) varied by participant and scale.[1] The CLUS model is based on a dirichlet process which tries to cluster the items in the comparison class. In the original paper, the probability that an item count as "tall" is the probability that it belongs to the same cluster as the tallest item, conditional on the tallest and shortest items belonging to separate clusters. We adapted this model to our data, which contains both affirmative and negative utterances, by making the probability also conditional on the item patterning with either the tallest or the shortest item in the comparison class (thereby enforcing the correct behavior for negation). We used beta distributions for the kernel and the parameters of the model were kept constant (using uninformative priors).[2] For the two Bayesian models, we allowed the cost of each adjective to vary within participant, but kept the rationality parameter constant across scales. All models were fitted by maximum likelihood under the assumption that slider data are normally distributed around the values predicted by the model, with censoring at both ends of the slider.

Figure 3 shows the predictions of the different models against the data, and Table 1 gives the likelihood broken down by scale type, overall likelihood, and the AIC and BIC. The RH-R and the RSA are in close competition for the best fit overall (the former performs slightly better, but with more parameters). The SOM gives a slightly better fit for open scales but trails behind on all closed scales, where it systematically predicts strict absolute interpretations. The CLUS model tends to overestimate the slope on closed scales, and underestimate it on open scales (where it struggles to create distinct clusters).

Model	ℓ_{unbound}	$\ell_{\text{lower-bound}}$	$\ell_{\text{upper-bound}}$	$\ell_{\text{double-bound}}$	ℓ_{TOTAL}	AIC	BIC
RH-R	−5584	−4494	−4326	−4360	−18765	☞38909	43451
CLUS	−6146	−5065	−5059	−4691	−20961	42512	44454
RSA	−5581	−4600	−4410	−4454	−19044	39078	☞42336
SOM	−5459	−5555	−5558	−5388	−21961	44592	46797

Table 1: Log-likelihood broken down by scale type, overall log-likelihood, AIC and BIC for each model.

Discussion

Our results with nonce adjectives clearly establish that comparison classes alone can make an adjective absolute or relative, even though the scale on which the adjective measures entities has natural lower and upper bounds (as indicated in our experiment by the use of percentages). These results cast doubt on the idea that the class of an adjective would be lexically encoded, or derived from lexically-encoded features of the adjectives, and supports the proposal that distributional properties of the comparison class are the deciding factor. Lexical content of the adjective remains relevant in that it provides the measure function which maps entities to degrees. For a given comparison class, it is conceivable that different measure functions would return degree distributions with different properties. Nevertheless, our data suggests that it is not necessary to postulate introspectively inaccessible scales or an additional layer of interpretation (as proposed by Wellwood, 2020) to account for apparent counter-examples to the generalization of Kennedy & McNally such as 'expensive/cheap'. Statistical properties of the comparison class are sufficient to explain these examples, and our results confirm that naive speakers do in fact make full use of these properties. Finally, we note that lower-bound scales gave the most

[1] A hierarchical model would have been ideal, but it turned out to be computationally impractical for all other models, so we stuck to a simpler architecture where parameters for all participants are considered independent.

[2] The hyperparameters of the model were not adjusted to best fit the data due to computational limitations. Nevertheless, the results suggest that doing so would not have helped in any way, as the model both underestimates the number of clusters for open scales and overestimates it for closed scales.

robust absolute meanings, and that double bound scales were most similar to upper-bound scales, as Kennedy & McNally (2005) observed in naturally occurring adjectives (e.g., ‘full’).

References

Bennett, Erin D & Noah D Goodman. 2018. Extremely costly intensifiers are stronger than quite costly ones. *Cognition* .

Bolinger, Dwight. 1972. *Degree words*, vol. 53 Janua Linguarum. The Hague: De Gruyter Mouton.

Frank, Michael C & Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336(6084). 998–998.

Hersh, Harry M & Alfonso Caramazza. 1976. A fuzzy set approach to modifiers and vagueness in natural language. *Journal of Experimental Psychology: General* 105(3). 254.

Kennedy, Christopher & Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* .

Lassiter, Daniel & Noah D Goodman. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In Todd Snider (ed.), *Proceedings of SALT 23*, 587–610.

Leffel, Timothy, Alexandre Cremers, Nicole Gotzner & Jacopo Romoli. 2019. Vagueness in implicature: the case of modified adjectives. *Journal of Semantics* .

Qing, Ciyang & Michael Franke. 2014a. Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In *SALT 24*, LSA.

Qing, Ciyang & Michael Franke. 2014b. Meaning and use of gradable adjectives: Formal modeling meets empirical data. In Paul Bello, Marcello Guarini, Marjorie McShane & Brian Scassellati (eds.), *Proceedings of the cogsci36*, 1204 – 1209. Austin, TX: Cognitive Science Society.

Schmidt, Lauren A, Noah D Goodman, David Barner & Joshua B Tenenbaum. 2009. How tall is tall? compositionality, statistics, and gradable adjectives. In Niels Taatgen & Hedderik van Rijn (eds.), *Proceedings of the 31st annual meeting of the cognitive science society (CogSci-2009)*, vol. 3, 3151–3156. Austin, TX: Cognitive Science Society.

Solt, Stephanie & Nicole Gotzner. 2012. Experimenting with degree. In Anca Chereches (ed.), *Proceedings of salt*, vol. 22, 353–364.

Tessler, Michael Henry, Michael Lopez-Brau & Noah D. Goodman. 2017. Warm (for winter). In *Cogsci 29*, .

Unger, Peter. 1971. A defense of skepticism. *The Philosophical Review* 80(2). 198–219.

Wellwood, Alexis. 2020. Interpreting degree semantics. *Frontiers in Psychology* 10. 2972. doi: 10.3389/fpsyg.2019.02972.