Interpreting gradable adjectives: Rational reasoning or simple heuristics?

Alexandre Cremers

Abstract Gradable adjectives can be categorized into relative adjectives (‘tall’, ‘far’, ‘happy’), which are vague and context-dependent, and absolute ones (‘dry’, ‘dirty’, ‘full’), which are much less context-dependent and can receive a strict interpretation. Different explanations have been proposed in the literature to explain this split, most saliently: a lexical approach, where the category is determined by properties of the scale on which the adjectives measures entities (Kennedy & McNally 2005), and a pragmatic approach, which refers to properties of the distribution of measurements in the comparison class (Lassiter & Goodman 2013: a.o.). A related debate concerns the nature of the cognitive processes responsible for integrating contextual information: simple heuristics or sophisticated rational reasoning? Pragmatic approaches are split between theories which assume rationality at the speaker’s level and evolutionary theories which instead focus on long-term optimality, while lexicalist approaches tend to rely on heuristics. The experimental literature has established an effect of the comparison class on the interpretation of relative adjectives, but it is still unclear whether it can determine an adjective’s category, and rational models have not been directly compared with simpler heuristics. We present an experiment using nonce adjectives (to control for lexical information and world knowledge), in which the range of the scale is always closed. Comparison classes vary in the probability mass they place at scale boundaries, a factor which probabilistic pragmatic accounts take to be the determining factor. We found that simple heuristics perform as well as the best rational model, and that the degree distribution within the comparison class can lead to categorical distinctions in the interpretation of nonce adjectives, although it remains unclear whether the resulting categories constitute genuine absolute and relative meanings.

Keywords gradable adjectives · degree semantics · vagueness · probabilistic pragmatics

A. Cremers, Vilniaus Universitetas, `https://alexandrecremers.com/`

In Gabriela Bîlbîie, Berthold Crysmann & Gerhard Schaden (eds.), Empirical Issues in Syntax and Semantics 14,
31–60. Paris: CSSP. `http://www.cssp.cnrs.fr/eiss14/`

1 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) argued 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, with a strict minimum, maximum, or both. Whether the closed scale is upper- or lower-bound further distinguishes between minimum and maximum-standard absolute adjectives. Scales that are fully closed tend to give rise to maximum-standard adjectives.

For Kennedy & McNally (2005), these distinctions are a matter of lexical semantics, in that 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).

A point often raised against the lexical approach is that lexically-encoded scales do not always match the scale we would intuitively associate to an adjective. For instance, the cost scale associated with ‘cheap/expensive’ should have 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 (Kennedy 2007). 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 determine 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, where linguistic interpretations are first mapped to non-linguistic thoughts, which in turn can be assigned truth-values. She argues that such non-linguistics representations are needed anyway, and can in principle be tested independently. In the case of 'expensive', the assumption that our (non-linguistic) concept of price excludes 0 could indeed be independently motivated (see 'zero-price effect', Shampanier & Mazar & Ariely 2007).

In the same vein, McNally (2011) proposes that the difference between relative and absolute adjectives corresponds to different ways of categorizing objects. Relative adjectives would cluster them according to similarity with one another (which requires a comparison class), while absolute adjectives would use rule-based categorization. In this view, the crucial feature is not the scale structure anymore, but background knowledge about the dimension encoded by the adjective, which decides whether a rule can be derived or whether a comparison class is needed.

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: henceforth L&G) 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). These accounts therefore claim to make lexical stipulations superfluous.

This debate raises about a secondary question about the exact nature of the process that maps comparison classes onto adjective thresholds. According to McNally (2011), this process is a general similarity-based clustering and only plays a role for relative adjectives. There is no consideration of whether the resulting classification is optimal for language use. On the other hand, L&G and Qing & Franke (2014a: henceforth Q&F) assume that a single process, highly specialized for linguistic purposes, is responsible for both relative and absolute interpretations. They differ in that L&G see this process as explicit pragmatic reasoning occurring every time a gradable adjective is uttered, whereas Q&F adopt an evolutionary approach, in which the optimization of the mapping from comparison class to threshold happens at the level of a linguistic community, not internally for each agent (at the level of agents it could have solidified into a simple heuristics).

Previous experimental work (Syrett et al. 2004; Schmidt et al. 2009; Solt & Gotzner 2012; Qing & Franke 2014b, a.o.) shows that the distribution of degrees within a comparison class does affect the threshold of adjectives, but the link between closed scales and absolute interpretations has not been explicitly tested and simple heuristics have not been directly compared with Bayesian models. Xiang et al. (2022) recently showed that existing Bayesian models offer a good fit of the communicative effect of relative and absolute gradable adjectives, but fail to capture truth-value judgments, unless supplemented with semantic conventions. Meanwhile, most modeling work follows L&G in assuming that the prior distribution alone determines the class of the adjective (Q&F; Tessler & Lopez-Brau & Goodman 2017; Bennett & Goodman 2018).

We therefore propose a new experiment with two main goals. The first is 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). To do so, we strip all world knowledge and lexical information by using nonce adjectives and fictional measures, and observe whether a categorical distinction between relative and absolute adjectives can emerge nonetheless. The second goal is to test explicit quantitative accounts of gradable adjectives. Several different types of models have been proposed: simple heuristics (Schmidt et al. 2009; Schmidt 2009), rational reasoning (L&G), and evolutionary models (Q&F, Correia & Franke 2019). By collecting a large enough dataset, we will be able to systematically compare models from each category.

2 Experiment

2.1 Methods

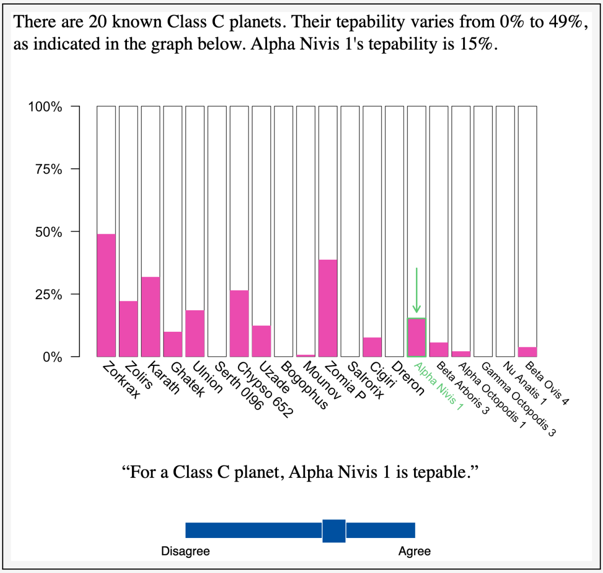
We tested the interpretation of nonce adjectives in the presence of explicit comparison classes which each comprised 20 planets, for which we gave fictional measurements of the dimension measured by the adjectives. The use of nonce adjectives ensured that only information about the scale and the comparison class was available to 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. The experiment was run on Alex Drummond's Ibex Farm.

After validating the consent form, participants received instructions which included the introduction text in (1) as well as three example items which drew their attention to the comparison class and the fact that sometimes a clearcut answer wasn't possible.

(1) An advanced alien civilization from a distant galaxy has explored all the planets in their star system as well as many planets orbiting neighboring stars. They have classified these planets into a number of categories and have measured different properties.

In this survey, you will see some of these measurements for some categories of planets, and we will ask you to tell us how much you agree with statements about individual planets.

After reading the instructions participants saw three training items similar to the examples to help them familiarize with the task. In each trial, they 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 participant, we created



There are 20 known Class C planets. Their tepability varies from 0% to 49%, as indicated in the graph below. Alpha Nivis 1's tepability is 15%.

"For a Class C planet, Alpha Nivis 1 is tepable."

Figure 1 Example trial with a Lower-bound comparison class (many items are at or close to 0%).

8 comparison classes by sampling parameters from uniform distributions with ranges given in Table 1 (with probability distributions corresponding to lower-, upper-, double-bound and unbound scales in the Bayesian pragmatics literature). Each comparison class was paired with a nonce-adjectives from Table 2. For each comparison class, we tested the applicability of a predicate to half the elements, and the applicability of its negation for the other half (randomly selected as odd and even quantiles). Three comparison classes were paired with bare (positive form) adjectives and four featured adjectives modified by 'very', 'extremely', 'absolutely', and 'quite'. Each of these constructions could appear in affirmative of negated form. The last comparison class appeared with 'a bit' in affirmative sentences and 'at all' in negative sentences. The 8 comparison classes were broken down into 16 blocks of 10 trials (affirmative and negative forms separated to avoid confusion). The blocks were presented in random order, and items within each block were also randomized. The association between scales, adjectives, and constructions was randomized and balanced.

Distribution	p_0	p_1	a	b	Support
Unbound 1	0	0	$[4, 40]$		$(0, 1)$
Unbound 2	0	0	$[3, 15]$		
Lower-bound 1	$[0.1, 0.7]$	0	$[0.7, 1]$	$[1, 6]$	$[0, 1)$
Lower-bound 2	$[0.2, 0.65]$	0	$[1, 2.5]$	$[1, 8]$	
Upper-bound 1	0	$[0.1, 0.7]$	$[1, 6]$	$[0.7, 1]$	$(0, 1]$
Upper-bound 2	0	$[0.2, 0.65]$	$[1, 8]$	$[1, 2.5]$	
Double-bound 1	$[0.1, 0.35]$		$[0.7, 1]$		$[0, 1]$
Double-bound 2	$[0.1, 0.25]$		$[1, 3.5]$		

Table 1 Parameter ranges of the inflated beta distributions used to generate comparison classes. p_0 and p_1 are the discrete probability mass at 0 and 1 respectively. a and b parametrize the beta distribution. The last column indicates the support of the distributions (which align with their name).

Bare form	Noun	Comparative	Superlative
roagly	roagliness	roaglier	roagliest
vibble	vibbleness	vibbler	vibblest
drok	drokth	drokker	drokkest
scrop	scropth	scropper	scroppest
plard	plardity	plarder	plardest
hif	hifth	hiffer	hiffest
tepable	tepability	more tepable	most tepable
plawic	plawicity	more plawic	most plawic

Table 2 Nonce adjectives used in the experiment, together with the derived noun for the measurement. The comparative and superlative forms were only used in examples and training items. We varied the morphology across the 8 adjectives, as some suffixes may be biased towards specific categories. However, as discussed in the results section, we did not observe any difference between the 8 adjectives.

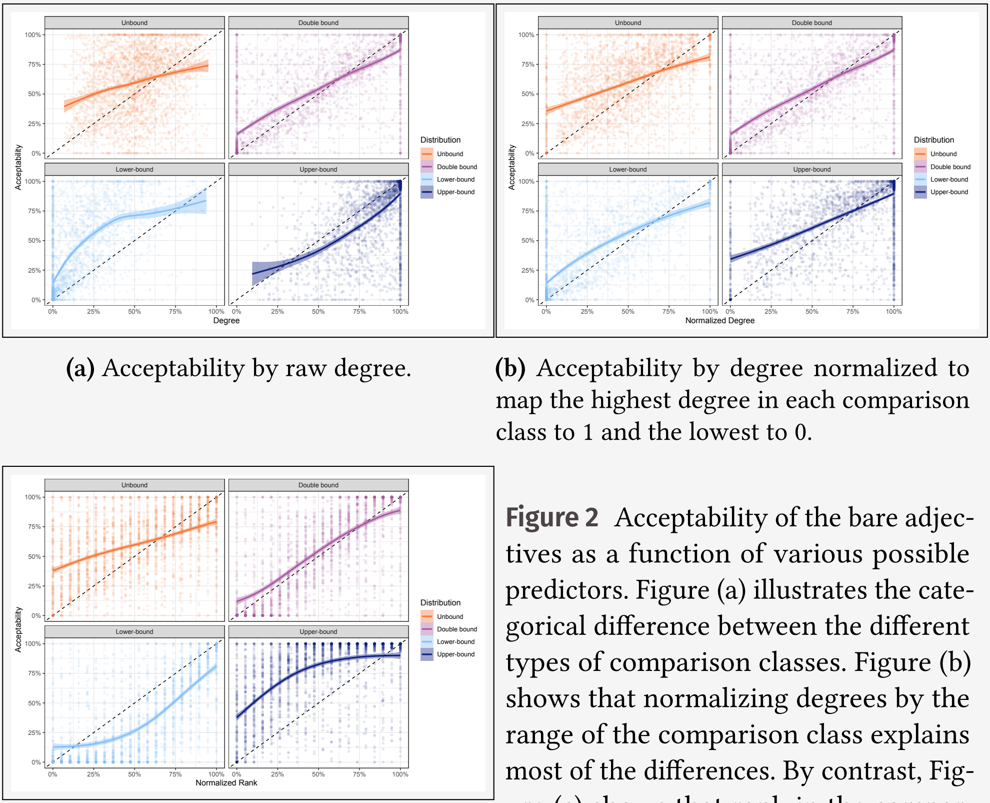
2.2 Participants

We recruited 222 participants on MTurk, paid $2 each (the survey took about 10min). We removed participants whose median RT was below 1s, or who had more than 50% duplicated responses (i.e., didn't move the slider between subsequent trials). 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: $-.36$), as well as participants who fell below the threshold on at least half of the blocks. The goal was to remove cases where participants missed a change of polarity between two blocks, which happened on 7% of affirmative blocks and 14% of negative blocks. In all, we filtered out 20% of the initial data set.

3 Results

The full data set and analysis scripts are available at `https://github.com/Alex-Cremers/nonce-gradable-adj`. 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 and steepness for pairs of affirmative and negative blocks (excluding the 'a bit/at all' cases). We found no significant differences (midpoint: $t(404) = -0.10$, $p = 0.92$; steepness: $t(456) = -0.85$, $p = 0.40$), 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 $\mathrm{Acc}(\neg S) = 1 - \mathrm{Acc}(S)$. From now on, we focus on the bare adjectives only.

In order to diagnose absolute interpretations, we computed the slope of acceptability as a function of degree at both ends of each scale. Min. std. adjectives would have a sudden increase in acceptability at the bottom of the scale, since the threshold should most likely be located right above the minimum degree. For max. std. adjectives, we expect a steep increase at the top of the scale, as the threshold should sit right below the maximum of the scale. By contrast, relative adjectives should be flat at both extremities of their degree distribution, since their vague threshold should be situated slightly above the middle of the scale (their acceptability should form a sigmoid). The slopes were computed using linear regressions on the first and last 3 degrees on each scale. Figure 3 displays the measured slopes.



(a) Acceptability by raw degree.

(b) Acceptability by degree normalized to map the highest degree in each comparison class to 1 and the lowest to 0.

(c) Acceptability by rank of the item in the comparison class, normalized to [0,1].

Figure 2 Acceptability of the bare adjectives as a function of various possible predictors. Figure (a) illustrates the categorical difference between the different types of comparison classes. Figure (b) shows that normalizing degrees by the range of the comparison class explains most of the differences. By contrast, Figure (c) shows that rank in the comparison class alone cannot explain participants' judgments.

For statistical analysis, we applied an inverse hyperbolic sine (IHS) transform on slopes, which gives good results on right-skewed data with negative values (Burbidge & Magee & Robb 1988). We ran two mixed-effects regressions on these IHS-transformed slopes—one for slopes at the bottom of the comparison class and one for slopes at the top—with Distribution type as a predictor (treatment-coded with Unbound as the reference level) and random by-subject intercepts. The detailed results, in Table 3, confirm the differences which are visible in the graph: acceptability ratings vary significantly more dramatically near closed boundaries than open boundaries.

At this point, it is still unclear what drives the difference between the distributions, and whether participants are sensitive to fine-grained distributional differences within each type of comparison class. As a post-hoc analysis, we tested the effect of p_0 (the probability mass at 0) on bottom slopes and p_1 on top slopes by adding them as predictor to the mixed-effects models described above. We found no effect of p_0 for bottom slopes ($\beta = .098$, $\chi^2(1) = 1.4$, $p = .24$) and a small but *negative* effect of p_1 for top slopes ($\beta = -.18$, $\chi^2(1) = 5.4$, $p = .02$). This suggests that the presence of items from the comparison class at a scale boundary can shift the threshold to this boundary, but *how many* items are at the boundary does not actually matter, even though we varied this number dramatically, from 2 to 14.

Finally, one may wonder whether the eight nonce adjectives differed with respect to our measure. We tried to vary the morphology among them, and it is possible that some suffixes were biased towards a specific category of gradable adjectives. To test this possibility, we updated the mixed models with a fixed categorical factor of adjective. This didn't improve the fit on either top ($\chi^2(7) = 4.54$, $p = 0.72$) or bottom ($\chi^2(7) = 4.49$, $p = 0.72$) slopes, suggesting that there is no significant differences between the 8 adjectives.

	Distribution	β	t	p
Bottom	(unbound)	0.44	5.7	$< .001$
	double-bound	0.27	2.6	.011
	lower-bound	0.31	2.9	.004
	upper-bound	−0.03	−0.3	.77
Top	(unbound)	0.30	4.6	$< .001$
	double-bound	0.34	3.8	$< .001$
	lower-bound	0.08	0.9	.37
	upper-bound	0.35	3.9	$< .001$

Table 3 Results of the mixed-effects model on IHS-transformed slopes. Unbound is the reference level (intercept), other parameters correspond to the difference. The Double-bound distributions have higher slopes than Unbound at both ends. The Lower-bound distributions have a steeper slope at the bottom of the scale, but not at the top. The Upper-bound distributions exhibits the opposite pattern.

4 Modeling

Our results confirm that participants are sensitive to the distribution of degrees in the comparison class, and further demonstrate that their response

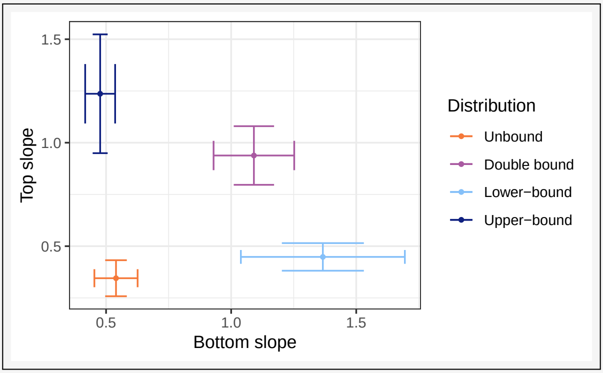


Figure 3 Slope computed on the three highest degrees as a function of slope on the three lowest degrees, by distribution type (mean and standard deviation).

patterns roughly fall into the usual categories of relative, min. std. and max. std. adjectives. However, the effect seems to come almost exclusively from the range of degrees in the comparison class, and whether this range reaches 0 or 1. This would suggest—against usual assumptions in the Bayesian pragmatic literature—that participants' behavior is not in fact sensitive to subtle distributional effects, and may be better described by a simple heuristic. Bayesian pragmatics is not out of the race yet however. For instance, the Speaker-Oriented Model (SOM) of Q&F allegedly switches to a minimum standard interpretation when the probability mass near the lower boundary reaches a tipping point. It would thus behave more categorically than the RSA model of L&G.

We now present quantitative models that have been proposed to capture effects of comparison class on adjectives and which we will test against our data. The first model is a very simple heuristic sensitive to the range of degrees only, while the second is a more complex one based on similarity-based clustering. We then present two implementations each of L&G's RSA model and Q&F's evolutionary SOM.

4.1 The RH-R model

Schmidt et al. (2009) tested 9 different descriptive models of gradable adjectives, which exploit various statistical properties of a discrete comparison class to predict the interpretation of an adjective. Of these 9, we will only consider the best two models. The first one, "Relative height by Range"

(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 allowed both parameters of the model (the proportion of degrees which validate the adjective and the fuzziness parameter) to vary independently by participant in a hierarchical model. For comparison with the Bayesian pragmatic models, we fitted the proportion of items for which the adjective is true as one minus the inverse logit of a "cost" parameter which was normally distributed among participants. The participants' fuzzyness parameters were sampled from a log-normal distribution. Each participant was also assigned a noise parameter (describing the error between the model predictions and the data), which was also log-normally distributed. For this model and all following models except CLUS, we also included a Gaussian random effect of nonce-adjective on cost.

4.2 The CLUS model

A second model proposed by Schmidt et al. (2009), CLUS, performed about as well as the RH-R on their data. It is a more sophisticated model based on a probabilistic clustering, and is therefore a good representative of what McNally (2011) assumes for relative adjectives. In detail, a Dirichlet process builds a probabilistic partition of the items in the comparison class, assuming that the degrees of items within the same cell follow the same normal distribution. In this model, the probability that an item counts as "tall" is the probability that it belongs to the same partition cell as the tallest item in the comparison class, conditional on the tallest and shortest items belonging to separate cells. Our detailed implementation of the model is given in Appendix A. The model has several free parameters governing the priors of the parameters of the Gaussian distribution for each cluster, and an α parameter tuning the model's bias towards few large clusters or many smaller clusters. A hierarchical fit was attempted, but turned out to be computationally too difficult. In the end, we fitted only the α parameter.

4.3 Lassiter and Goodman's RSA model

L&G build on the standard Rational Speech-Act model, where speakers are assumed to maximize the trade-off between informativity and cost for each utterance, but they assume that listeners also reason about which possible threshold θ could have led a speaker to use a gradable adjective. The literal listener L_0, parametrized by θ, is defined as:

(2) $$L_0(d|\text{adj}, \theta) \propto \varphi(d)[\![\text{adj}]\!]^{\theta}(d) \qquad \text{where } \varphi \text{ is the prior on degrees}$$

We adopt a non-strict semantics for gradable adjectives, except when $\theta = 0$. This means that $\theta = 0$ corresponds to a strict min. std. interpretation, and $\theta = 1$ to a strict max. std. interpretation:

(3) $$[\![\text{adj}]\!]^{\theta,d} = 1 \qquad \text{iff} \quad \big(d \geq \theta > 0 \quad \text{or} \quad d > \theta = 0\big)$$

As usual in RSA, the utility function U_1 represents a trade-off between informativity and cost, here parametrized by θ. We only consider two messages (the bare positive-form adjective and the null message) so the function describing the pragmatic speaker S_1 remains very simple:

(4) $$U_1(u|d, \theta) = \log L_0(d|u, \theta) - cost(u)$$

(5) $$S_1(\text{adj}|d, \theta) = \left| \begin{array}{ll} \dfrac{1}{1 + \mathrm{e}^{\lambda[\log \Phi^{c*}(\theta) + c_{\text{adj}}]}} & \text{if } [\![\text{adj}]\!]^{\theta}(d) = 1 \\ 0 & \text{otherwise} \end{array} \right.$$

Where $\Phi^{c*}(\theta)$ is the probability that $[\![\text{adj}]\!]^{\theta}(d)$ is 1 given a fixed θ and the prior distribution φ on d (i.e., the complementary cumulative distribution function of d, modulo the strict/non-strict adjustment). Note that, as a function of d, S_1 is simply a step function.

The pragmatic listener infers both d and θ, using a prior $P(\theta)$ on the threshold:

(6) $$L_1(d, \theta|\text{adj}) \propto \varphi(d)P(\theta)S_1(\text{adj}|d, \theta)$$

In order to make predictions regarding the acceptability of the adjective, we need to compute the posterior cumulative distribution function of θ (Lassiter & Goodman 2015: Eq. 32). We first marginalize over d:[1]

[1] The move from the first line to the second is valid because $S_1 = 0$ for $d < \theta$. The move to the third line is possible because S_1 is constant for $d \geq \theta$ and can therefore be factored out of the sum.

(7)
$$\begin{aligned} L_1(\theta|\text{adj}) &\propto \sum_d \varphi(d)P(\theta)S_1(\text{adj}|d,\theta) \\ &\propto \sum_{d\geq\theta} \varphi(d)P(\theta)S_1(\text{adj}|d,\theta) \\ &\propto \left(\sum_{d\geq\theta}\varphi(d)\right)P(\theta)S_1(\text{adj}|d\geq\theta,\theta) \\ &\propto \frac{\Phi^{c*}(\theta)P(\theta)}{1+\mathrm{e}^{\lambda[\log\Phi^{c*}(\theta)+c_{\text{adj}}]}} \end{aligned}$$

We can then derive the predicted acceptability. The normalizing constant is simply the integral up to the highest degree in the comparison class:

(8)
$$\text{Acc}(\text{adj}|d_i) \propto \int_0^{d_i} \frac{\Phi^{c*}(\theta)P(\theta)}{1+\mathrm{e}^{\lambda[\log\Phi^{c*}(\theta)+c_{\text{adj}}]}}\mathrm{d}\theta$$

We tested two priors on θ: a continuous uniform prior on [0,1], and recycling the discrete degree prior. The first option is what L&G proposed; the second corresponds to sampling the threshold from the comparison class. In the rest of the paper, we name these two models RSA-U and RSA-I (for **U**niform and **I**nformed priors, respectively).

An interesting property of the RSA model is that it cannot ever predict less than 100% acceptability on the maximal element in a comparison class. Indeed, the acceptability is meant to track posterior probability of θ after hearing an utterance of the adjective in its positive form. Since the adjective must be true of some degree to have been uttered truthfully, θ cannot exceed the highest degree in the comparison class. By contrast, the model does not necessarily prevent the adjective from receiving a tautologous interpretation, which would make even the lowest degree acceptable. The possible values for θ are further restricted by the prior however (in particular, our priors do not allow $\theta < 0$, so a degree of 0 is always unacceptable in our implementation).

The participants' rationality parameters followed a log-normal distribution while the costs followed a normal distribution. As with the RH-R model, participants were also assigned a noise parameter from a log-normal distribution.

4.4 Qing and Franke's SOM

The SOM proposed in Q&F is superficially similar to the RSA, in that it also works around a trade-off between communication success and cost. Conceptually, it is very different however: the threshold is assumed to be a convention among a community of language users, and the trade-off is optimized in the long term rather than for an isolated utterance.

Formally, the model assumes that there is a convention on a probabilistic distribution for θ, $\Pr(\theta)$, and that this distribution is an approximation of the threshold which would optimize a trade-off between expected success (the probability that communication is successful) and expected cost (which increases with the frequency of use of the adjective). For a discrete prior, the two components are defined as follow:

(9) Expected success:

$$\begin{aligned} ES(\theta) &= \sum_{u_1(d)=0} \varphi(d)\varphi(d|u_0,\theta) + \sum_{u_1(d)=1} \varphi(d)\varphi(d|u_1,\theta) \\ &= \sum_{d<\theta} \varphi(d)^2 + \sum_{d\geq\theta} \frac{\varphi(d)^2}{P(d\geq\theta)} \\ &= \sum \varphi^2 + \frac{P(d<\theta)}{P(d\geq\theta)} \sum_{d\geq\theta} \varphi(d)^2 \end{aligned}$$

To understand the definition of expected success, one first needs to consider that there is a $\varphi(d)$ probability that a speaker may want to communicate the degree d. If d is below the threshold, the speaker cannot use the adjective, and so the listener also has a probability $\varphi(d)$ to guess the correct degree (hence the $\varphi(d)^2$). If d is above θ, the speaker can and will use the adjective, so the listener can conditionalize on the information that $d \geq \theta$, therefore they will guess d with probability $\frac{\varphi(d)}{P(d\geq\theta)}$.

The expected cost is simply the cost of the adjective multiplied by the probability that the adjective will be used given θ:

(10) Expected cost:

$$EC(\theta) = c_{\text{adj}} \times P(d \geq \theta)$$

The utility of a fixed threshold θ is then defined as the difference between expected success and expected cost, and the conventional distribution of θ

is assumed to be a softmax over possible thresholds:

$$
\text{(11)} \qquad \begin{aligned} \Pr(\theta) &\propto \exp(\lambda[ES(\theta) - EC(\theta)]) \\ &\propto \exp\left(\lambda\left[\frac{P(d<\theta)}{P(d\geq\theta)}\sum_{d\geq\theta}\varphi(d)^2 - c_{\text{adj}}P(d\geq\theta)\right]\right) \end{aligned}
$$

In this case, the softmax is not meant to encode sub-rationality (as we're talking about optimization at the level of a whole community), but to represent various sources of noise (in particular uncertainty on the prior distribution) which lead to vagueness.

Unlike the RSA, the SOM does not impose any restriction on θ, and in particular does not exclude thresholds outside the comparison class (which would make the adjective trivially true or false). For comparison with the RSA model, we assume that θ can fall under the smallest degree if it is positive, but cannot exceed the highest degree in the comparison class (i.e., the adjective can be trivial, but it cannot be contradictory). The second difference is that the SOM is not sensitive to the actual degrees, only to the probability distribution over them. This means that if the comparison class comprises of n degrees, there are at most $n+1$ thresholds to consider (in practice, we'll only consider n or $n-1$, given the previous point).[2]

The model uses the same parameters as the RSA, so we adopted the same distributions. However, due to difficulty fitting the model, we adopted more restrictive priors on the variance of the random effects.

4.5 An update on the SOM

The notion of expected success in the SOM is only properly defined for discrete priors[3] and has some problematic mathematical properties. For instance, given a discrete comparison class where elements are equiprobable, the model always predicts the maximum-standard interpretation to be

[2] Things would be different if we considered all possible values of θ in [0,1] and decided to integrate $\exp \lambda U(\theta)$ (see fn. 5), but this would require picking a prior on θ, which goes contrary to the spirit of the SOM.

[3] Q&F propose a continuous generalization with $\int \varphi^2$ where φ is the density of a continuous prior, but with some continuous priors, this integral is not finite, so the distribution of θ is not defined. Even when finite, this quantity is not scale independent, while the expected cost is. As a result, the model predictions depend on the unit of measurement (i.e., the prediction for 'tall' are qualitatively different if heights are measured in cm or in).

optimal, unless the cost of the adjective is negative.[4]

For these reasons, we propose an update to the SOM to give it a better-behaved expression by replacing the problematic notion of expected success with expected informativity (in line with the RSA). We refer to this model as "SOM-EI" in what follows.

(12) Expected informativity:

$$\begin{aligned} EI(\theta) &= \sum_{u_1(d)=0} \varphi(d) \log \varphi(d|u_0, \theta) + \sum_{u_1(d)=1} \varphi(d) \log \varphi(d|u_1, \theta) \\ &= \sum_{d<\theta} \varphi(d) \log \varphi(d) + \sum_{d\geq\theta} \varphi(d) \left[\log \varphi(d) - \log(1 - \Phi(\theta))\right] \\ &= \sum \varphi \log \varphi - P(d \geq \theta) \log P(d \geq \theta) \\ &= -H(\varphi) - P(d \geq \theta) \log P(d \geq \theta) \end{aligned}$$

Where $H(\varphi)$ is the entropy of the prior. This notion of expected informativity can be generalized to the continuous case using the notion of differential entropy ($h(X) = E[\log X]$). While differential entropy is not scale independent, it only appears as an additive constant in the utility function $U(\theta)$, and therefore does not affect the predictions of the model:

(13) $$U(\theta) = -h(\varphi) - P(d \geq \theta) \left[\log P(d \geq \theta) + c_{\text{adj}}\right]$$

As this expressions makes clear, this version of the SOM is only sensitive to the proportion of elements in the comparison class that end up below or above the threshold. For instance, with $c_{\text{adj}} = 0$, the optimal θ is the value that makes the adjective applicable to $\mathrm{e}^{-1} \approx 37\%$ of the elements. In this regard, this model is similar to the PN "Percent Number" model of Schmidt et al. (2009), which also focuses on the proportion of items counting as "tall". The two models differ on the shape of the distribution around this mode however, since the PN is always Gaussian while the SOM can take more exotic forms and may have a different mean.[5]

The parametrization of the SOM-EI was identical to that of the SOM.

[4] Proof: if there are N items in the comparison class, of which k are at or above θ, utility reduces to $\frac{1}{N} + \frac{N-k}{N^2} - \frac{k}{N} c_{\text{adj}}$. Assuming that c_{adj} is positive or null, this is a strictly decreasing function of k, which means that θ should be as high as possible.

[5] Looking at equation (8), one might think that this is also true of the RSA-U, since the comparison class only shows up through the function Φ^{c*}, which encodes the probability

4.6 Methods and Results

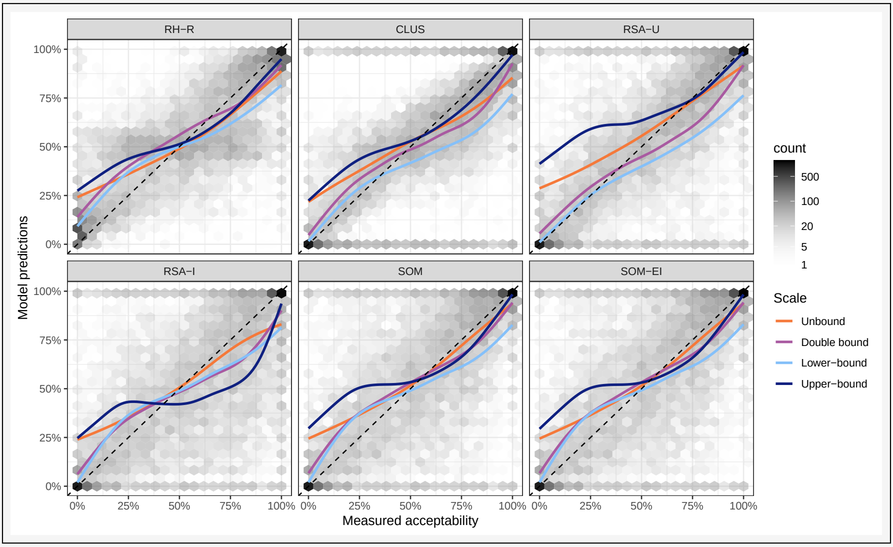


Figure 4 Model predictions against data for each of the models, with shade indicating the concentration of points in a given hex cell. The predictions of the best models fall along the diagonal since they tightly follow the data. The colored lines indicate the shape of predictions for each type of scale.

The models were fitted using Stan (Carpenter et al. 2017) on all data from bare adjectives. Participants' responses on sliders were assumed to follow a Gaussian censored at 0 and 1 (as in a tobit regression, Tobin 1958). Figure 4 presents each models predictions against the data.

We evaluated and compared the models through leave-one-out cross-validation (LOO-CV), using the PSIS technique from Vehtari & Gelman & Gabry (2017), as implemented in the package loo in R. LOO-CV consists in repeatedly removing one data point and refitting the model to generate predictions for the missing data point in order to evaluate the accuracy of a model's predictions. For complex models with a lot of data, this quickly

that the adjective is true (i.e. the proportion of the comparison class which falls above the threshold). Counterintuitively however, the distribution of degrees creeps back through the continuous prior on θ: because we need to integrate this continuous prior, the distance between the different degrees in the comparison class becomes relevant, and not just the proportion of items which falls below or above θ.

becomes impractical. The idea behind PSIS is to approximate the result of LOO-CV from the individual terms of the log-likelihood without having to refit the model for each data point. We treated each pair of participant and comparison class as a single data point for the cross-validation.

As the detailed results in Table 4 indicate, the RH-R, CLUS and RSA-U models best fitted the data (without significant difference between themselves). Surprisingly, sampling the threshold from the comparison class in the RSA-I model, leads to the worst predictions, in stark contrast with the RSA-U. The two versions of the SOM did not differ significantly, but were both much lower than the RSA-U.

Model	$\mathrm{elpd}_{\mathrm{loo}}$	Δ_{elpd}	$SE_{\Delta\mathrm{elpd}}$	p_{loo}	SE_p
RH-R	−1701	0	0	1287	44.2
CLUS	−1737	−35.8	160.8	1272	56.3
RSA-U	−1805	−103.5	178.0	1184	44.9
SOM-EI	−2612	−910.7	159.0	1011	45.6
SOM	−2696	−995.4	155.8	1267	70.8
RSA-I	−3774	−2073.3	161.4	1209	63.2

Table 4 Comparison of the different models using PSIS LOO-CV (Vehtari & Gelman & Gabry 2017). The first column indicates the *expected log pointwise predictive density*, which measures how well the model can generalize to unseen data. The next two columns, Δ_{elpd} and $SE_{\Delta\mathrm{elpd}}$, indicate the difference in elpd with the best model (RH-R in this case) and the estimated error on this difference. p_{loo} is the estimated effective number of parameters and SE_p the estimated error on this number.

We now turn to the posterior parameter estimates for each model, which are listed in Table 5. We first note that the RSA-U gives very reasonable parameters, with a median rationality of 2.4 and a mean cost of 1.6. In line with the results on top- and bottom-slopes which showed no differences between the nonce adjectives, the RSA-U also assigns the lowest variance of all models to by-adjective random effects.

By contrast, the SOM and SOM-EI require very negative costs of −2.8 and −4.9 respectively. As discussed above, the SOM is strongly biased towards maximum standard interpretations, and this bias can only be overcome by a negative cost or a large probability mass at the bottom of the scale. The SOM-EI does not usually place the mode of the distribution at the top of the scale, but it still puts more probability mass on high values, and no matter the distribution, it cannot predict the adjective to apply to more than 37% of the comparison class without a negative cost. To compensate for these large costs, the two models must assign very low rationality to the participants, with median 0.28 and 0.14 respectively.

5 Discussion

First of all, our results confirm the long-established effect of comparison class on the interpretation of gradable adjectives (Syrett et al. 2004; Schmidt et al. 2009; Syrett & Kennedy & Lidz 2010; Solt & Gotzner 2012; Qing & Franke 2014b; Xiang et al. 2022). They further establish that the distribution of degrees in the comparison class not only affects the interpretation of gradable adjectives, but is sufficient to see categorical contrasts reminiscent of the absolute/relative distinction emerge even in the absence of any world-knowledge or lexical information. On the face of it, one could think that this is enough to discard approaches which ground the absolute/relative distinction in lexical semantics (Kennedy 2007) or properties of the real-world scales denoted by the adjective (McNally 2011), but some limitations of the empirical design prevent us from drawing such a strong conclusion. Most importantly, all test sentences included a 'for'-phrase (e.g., "for a class B planet"), which has been argued to be incompatible with absolute adjectives (Siegel 1976: p155). This has led accounts which argue that absolute interpretations require more than a very biased comparison class (rule-based categorization for McNally 2011, a specific morpheme for Qing 2021) to assume that 'for'-phrases are incompatible with the source of absolute interpretations, and thus force a relative interpretation. McNally would further argue that information about the underlying real-world dimension is necessary to derive an absolute interpretations, as these are rule-based rather than similarity-based. Because our design intentionally stripped all pre-existing world-knowledge, it would prevent absolute interpretations. Participants may nevertheless be able to postulate simple rules by treating the 0 and 100% degrees as categorically different from the rest of the scale when some items reach them. This could in fact explain why we see categorical effects of distribution type but no dependency on the actual probability mass at scale ends. Note in passing that Qing's account of minimum-standard as zero-standard for the adjective 'profitable' may simply fall under McNally's account as a particular case of rule-based categorization.

Another potential limitation of the experiment concerns the use of percentages. The goal of the experiment was to make clear that all scales were closed on both sides, so that the comparison classes only varied in which part of these closed scales they populated. Percentages do not necessarily impose closed scales however. They can be used with dimensions we know to be associated with adjectives that encode open scales, as in (14) (Chris Kennedy, p.c.). Some scales also use percentages but can exceed 100%, as in (15), or even negative percentages (e.g., in a deflation situation). Conversely, a scale expressed in percentages may not actually reach its boundary, making it an open scale. For instance, the brightness of stars is defined in reference to the brightness of Vega.[6] A star can be arbitrarily faint, but it cannot reach 0% of the brightness of Vega, since all stars emit light.

(14) An optimum length is 50 percent of the length of the core

(15) In November 1923, inflation in Germany reached 29,525%

That being said, interpreting percentages as comparison with a reference, without the reference mentioned, does not seem very likely out of the blue, and the use fixed rectangles filled to various degrees in the images shown to participants reinforced the idea that the percentages were bound to 0–100%. The last point remains however: for comparison classes where no item reached 0% and/or 100%, it was entirely possible for participants to assume that these extreme values are physically impossible on the scale described by the nonce words. Note that for the results to remain compatible with a strict interpretation of Interpretive Economy, this would have to be the default assumption, otherwise we would have observed absolute interpretations for unbound comparison classes. On the other hand, if we accept that participants interpreted all nonce adjectives as denoting the same closed scale, our results would definitively establish that a closed scale can still give rise to a relative interpretation when the comparison class stays away from the boundaries. This would immediately capture the observation

[6] In practice, astronomers use the log of this ratio, but percentages are sometimes given for visible stars.

that ‘expensive/cheap’ are relative despite having a scale that includes 0.

Next, we must address the occasional ambiguity observed for lexicalized closed-scale adjectives between a min. std. and a max. std. interpretation. Kennedy (2007) cites in particular the case of ‘transparent’ which can alternatively mean “fully transparent” or “not fully opaque”. While these adjectives tend to favor a max. std. interpretation (see ‘full, closed’), McNally (2011) cites ‘familiar’ as a counter-example, which can receive a relative interpretation despite the availability of a “completely familiar” interpretation (see also debates around ‘likely’ in Lassiter 2011; Klecha 2012). Such examples would be prime targets for Bayesian pragmatic accounts: the variability may indicate sensitivity to a prior distribution which sometimes places more mass towards the top of the scale, sometimes towards the bottom. There is however good evidence against this probabilistic explanation in our data: for Double-bound distributions, not only did we not observe the usual max. std. interpretation that closed-scales adjectives such as ‘full’ typically receive, but we didn’t see any negative correlation between the top and bottom slopes (linear regression on IHS-transformed slopes: $\beta = 0.070, t = 0.94, p = .35$). If participants were split between min. std. and max. std. interpretations, we would expect such a correlation (they would place the threshold at one or another end of the scale). Instead, they all seem to assign a somewhat linear acceptability curve to the Double-bound cases. This would correspond to a uniform distribution for θ, as if they decided to remain fully agnostic on where the threshold might fall. This result suggests that when both scale ends are available, the distribution of degrees in the comparison class is not immediately relevant, and an external source is needed to disambiguate between relative, min. and max. std. interpretations. This could come from pure lexical idiosyncrasy, or—as McNally (2011) suggested—background knowledge about the physical dimension denoted by the adjective, which happens to be lacking here.

Given these potential caveats, in particular the point about for-phrases, our experiment likely did not reveal any genuine absolute interpretation, but only “absolute-like” relative interpretations. Remarkably, an important point of Q&F in favor of the SOM and against the RSA model was that the RSA failed to derive true absolute interpretations from extreme priors. Ironically, this inability to derive the true relative/absolute distinction from priors alone may be a strength of the model if these interpretations actually require more than probability mass near scale end. The SOM, on the other hand, is biased towards absolute interpretations—in particular maximum standard ones—, and may fare poorly on our dataset precisely because such extreme interpretations simply don't arise from priors alone. By contrast, the SOM outperforms the RSA in Xiang et al. (2022), who tested real English adjectives.[7] In short, the initial project of the probabilistic approach—to derive the absolute/relative distinction from distributional properties of the comparison class instead of stipulating it lexically—may have been doomed from the start because as soon as a comparison class is involved, we are dealing with a relative interpretation.

Our modeling results, together with those of Xiang et al. (2022), suggest that the RSA is a good model of relative interpretations, including "absolute-like" relative interpretations when the comparison class is concentrated at a scale end, but needs to be complemented with something else in order to capture actual absolute adjectives. We see two ways this could be done. The first is to accept Qing (2021)'s proposal that absolute interpretations involve a morpheme distinct from Kennedy and McNally's pos. We could even generalize this to include all rule-based interpretations, following McNally (2011). The second option would be to encode lexical knowledge in the θ-prior of the RSA. Indeed, the usual implementation of the RSA—the RSA-U which performed so well on our data—assumes a uniform prior on θ. This implies that the listener has no expectations whatsoever regarding the threshold before hearing a specific use of the adjective. Q&F already pointed out that this seems implausible. In practice, especially for frequent adjectives, the listener likely has quite specific expectations regarding where the threshold will fall. This effect would be particularly marked for absolute adjectives, since one can build more stable expectations regarding their threshold. It could even explain the idiosyncrasy of double-closed scale adjectives: over time, the prior on θ would concentrate on the side of the scale where comparison classes are most often concentrated, which can vary arbitrarily from one adjective to the next.

To conclude, we would like to come back to the debate between heuristics and explicit rational reasoning when it comes to computing the threshold

[7] It still struggled with min. std. adjectives though (presumably because of its bias towards max. std. interpretations).

of an adjective. We showed that the RSA and heuristics such as the RH-R and CLUS models were approximately equal when it comes to capturing participants' behavior in our experiment. However, our implementation of the RSA and modeling choices do not make it a great model of rational behavior either. First of all, affirmative and negative sentences clearly mirror each other in terms of acceptability, but in a model like the RSA, the negative sentence should be more costly, and therefore more informative. In other words, its threshold should be much lower. We avoided this problematic prediction of the model, and directly applied its prediction for an affirmative sentence to its negation by simply flipping the acceptability in $[0, 1]$. Second, the predictions for the affirmative sentence were computed based on only two messages: the bare form of the adjective and the null message. This ignores a lot of other messages a speaker might want to use to convey a given degree. In short, our implementation of the RSA makes it look more like an encapsulated heuristic than actual rational pragmatic reasoning. This actually seems like a better option when it comes to the computation of gradable adjectives thresholds, which doesn't seem to involve a lot of effort or conscious reasoning, appears to take place locally (e.g., within the scope of negation), and is acquired very early (Syrett & Kennedy & Lidz 2010), especially in comparison with implicatures. The relatively simple formula derived from the RSA-U model would only be a frozen heuristic, and the fact that it can be derived from a pragmatic model (reduced to its simplest form) could be a mere coincidence, or—more plausibly—would indicate that this heuristic mimics rational reasoning in bare positive uses of the adjective. As discussed in the introduction, there are two conceivable kinds of heuristics: the first would be the result of non-linguistic cognitive faculties clustering stimuli into categories which gradable adjectives can pick up. Such heuristics of non-linguistic origin would likely rely on categorizations useful for general cognition rather than specifically optimized for language use, and may not even be human-specific. The second kind would result from evolutionary processes shaping an optimal language (in terms of informativity, cost, learnability…), and would be specifically optimized for linguistic purposes.

Finally, the reliance on heuristics for the determination of the threshold would not mean, of course, that gradable adjectives cannot be involved in complex pragmatic reasoning. Leffel et al. (2019) present a puzzling example of interaction between vagueness and implicatures, and Cremers (2022) proposes an RSA model which explains this puzzle, but assumes that the distribution of the threshold has already been computed before genuine pragmatic reasoning can take place.

Acknowledgments I would like to thank Chris Kennedy and Ciyang Qing for in-depth discussions, as well as audiences at the UPenn Linguistics department, SURGE at Rutgers, and of course the audience of CSSP 2021. I also thank the anonymous reviewers of CSSP for valuable comments on the original abstract. I would like to thank Alex Drummond for the Ibex Farm. The experiment presented here made extensive use of the unique features of the last version of the Ibex Farm, including support for promises that Alex was kind enough to implement for this specific experiment. This research was funded by the European Social Fund under the measure 09.3.3-LMT-K-712 "Development of Competences of Scientists, other Researchers and Students through Practical Research Activities".

References

Bennett, Erin D & Goodman, Noah D. 2018. Extremely costly intensifiers are stronger than quite costly ones. *Cognition* 178. 147–161.

Bolinger, Dwight. 1972. *Degree words*. Van Schooneveld, Cornelis H. (ed.). Vol. 53 (Janua Linguarum). The Hague: De Gruyter Mouton.

Burbidge, John B & Magee, Lonnie & Robb, A Leslie. 1988. Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association* 83(401). 123–127.

Carpenter, Bob & Gelman, Andrew & Hoffman, Matthew D & Lee, Daniel & Goodrich, Ben & Betancourt, Michael & Brubaker, Marcus & Guo, Jiqiang & Li, Peter & Riddell, Allen. 2017. Stan: a probabilistic programming language. *Journal of statistical software* 76(1). 1–32.

Correia, José Pedro & Franke, Michael. 2019. Towards an ecology of vagueness. In Dietz, Richard (ed.), *Vagueness and rationality in language use and cognition* (Language, Cognition, and Mind 5), 87–113. Springer.

Cremers, Alexandre. 2022. A rational speech-act model for the pragmatic use of vague terms in natural language. In Culbertson, J. & Perfors, A. & Rabagliati, H. & Ramenzoni, V. (eds.), *Proceedings of CogSci 44*, 149–155.

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

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

Kennedy, Christopher. 2007. Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and philosophy* 30.

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

Klecha, Peter. 2012. Positive and conditional semantics for gradable modals. In *Proceedings of Sinn und Bedeutung 16*, 363–376.

Lassiter, Daniel. 2011. *Measurement and modality: the scalar basis of modal semantics*. New York University dissertation.

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

Lassiter, Daniel & Goodman, Noah D. 2015. Adjectival vagueness in a Bayesian model of interpretation. *Synthese*.

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

Lui, Arthur. 2021. *Dirichlet Process Gaussian mixture model via the stick-breaking construction in various PPLs*. Blog post. `https://luiarthur.github.io/TuringBnpBenchmarks/dpsbgmm`.

McNally, Louise. 2011. The relative role of property type and scale structure in explaining the behavior of gradable adjectives. In Nouwen, Rick & van Rooij, Robert & Sauerland, Uli & Schmitz, Hans-Christian (eds.), *Vagueness in communication*, 151–168. Berlin, Heidelberg: Springer Berlin Heidelberg.

Qing, Ciyang. 2021. Zero or minimum degree? Rethinking minimum gradable adjectives. *Proceedings of Sinn und Bedeutung* 25. 733–750.

Qing, Ciyang & Franke, Michael. 2014a. Gradable adjectives, vagueness, and optimal language use: a speaker-oriented model. In Snider, Todd & D'Antonio, Sarah & Weigand, Mia (eds.), *Proceedings of SALT 24*, vol. 24, 23–41. Washington, D.C.: LSA.

Qing, Ciyang & Franke, Michael. 2014b. Meaning and use of gradable adjectives: formal modeling meets empirical data. In Bello, Paul & Guarini, Marcello & McShane, Marjorie & Scassellati, Brian (eds.), *Proceedings of CogSci 36*, 1204–1209. Austin, TX: Cognitive Science Society.

Schmidt, Lauren A. 2009. *Meaning and compositionality as statistical induction of categories and constraints*. Massachusetts Institute of Technology dissertation.

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

Shampanier, Kristina & Mazar, Nina & Ariely, Dan. 2007. Zero as a special price: the true value of free products. *Marketing science* 26(6). 742–757.

Siegel, Muffy Emily Ann. 1976. *Capturing the adjective*. University of Massachusetts Amherst dissertation.

Sokal, RR & Rohlf, FJ. 1969. *Biometry: the principles and practice of statistics in biological research*. WH Freeman & Company Location San Francisco, CA.

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

Syrett, Kristen & Bradley, Evan & Kennedy, Christopher & Lidz, Jeffrey. 2004. Shifting standards: children's understanding of gradable adjectives. In Deen, K. U. & Nomura, J. & Schulz, B. & Schwartz, B. D. (eds.), *Inaugural conference on generative approaches to language acquisition-north america*, vol. 2, 345–352.

Syrett, Kristen & Kennedy, Christopher & Lidz, Jeffrey. 2010. Meaning and context in children's understanding of gradable adjectives. *Journal of semantics* 27(1). 1–35.

Tessler, Michael Henry & Lopez-Brau, Michael & Goodman, Noah D. 2017. Warm (for winter): comparison class understanding in vague language. In Gunzelmann, Glenn & Howes, Andrew & Tenbrink, Thora & Davelaar, Eddy (eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 1181–1186. Austin, TX: Cognitive Science Society.

Tobin, James. 1958. Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society* 26(1). 24–36.

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

Vehtari, Aki & Gelman, Andrew & Gabry, Jonah. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27(5). 1413–1432.

Wellwood, Alexis. 2020. Interpreting degree semantics. *Frontiers in Psychology* 10. 2972.

Xiang, Ming & Kennedy, Christopher & Xu, Weijie & Leffel, Timothy. 2022. Pragmatic reasoning and semantic convention: a case study on gradable adjectives. *Semantics and Pragmatics* 15. 9–EA.

Model	Parameter	Mean	$\mathrm{CI}_{\mathrm{low}}$	$\mathrm{CI}_{\mathrm{high}}$
RH-R	mean_cost	−0.46	−0.59	−0.31
	sigma_cost_adj	0.11	0.05	0.20
	sigma_cost_subj	0.71	0.61	0.81
	mean_log_eps	−0.70	−0.86	−0.56
	sigma_eps_subj	1.11	0.96	1.26
CLUS	alpha	1.18	1.17	1.20
RSA-U	mean_cost	1.63	1.27	2.00
	sigma_cost_adj	0.07	0.02	0.14
	sigma_cost_subj	1.79	1.42	2.18
	mean_log_lambda	0.87	0.74	0.99
	sigma_lambda_subj	0.68	0.57	0.80
RSA-I	mean_cost	−0.02	−3.37	3.15
	sigma_cost_adj	4.69	3.28	6.19
	sigma_cost_subj	12.53	9.92	15.27
	mean_log_lambda	0.10	0.00	0.20
	sigma_lambda_subj	0.37	0.29	0.46
SOM	mean_cost	−2.79	−4.47	−1.23
	sigma_cost_adj	1.74	0.86	2.89
	sigma_cost_subj	3.17	1.88	4.58
	mean_log_lambda	−1.29	−1.76	−0.82
	sigma_lambda_subj	2.03	1.69	2.38
SOM-EI	mean_cost	−4.91	−7.21	−2.73
	sigma_cost_adj	2.89	1.73	4.09
	sigma_cost_subj	6.10	4.52	7.70
	mean_log_lambda	−2.00	−2.37	−1.63
	sigma_lambda_subj	1.96	1.63	2.29

Table 5 Posterior parameters of the Stan models (mean posterior and 95% HDI confidence interval). The "mean_" parameters corresponds to fixed effects, the "sigma_" parameters correspond to the sd of random effects around these mean values (by subject or by adjective).

A Implementation of the CLUS model

The CLUS model assumes that the degrees of items in the comparison class were generated by an unknown number of normal distributions, and draws inferences about which items are likely to have their degrees likely coming from the same distribution. In practice, since the degrees in my experiment are bound to $[0, 1]$, I assume that the *arcsine transformed* degrees (Sokal & Rohlf 1969) follow an infinite Gaussian mixture.

For the details about the Dirichlet Process and the original implementation of the CLUS, see Schmidt (2009). My implementation follows the stick-breaking interpretation of the Dirichlet process, as described in Lui (2021). In practice, I set a cap at 10 clusters (exploration showed that the weights drop below 1% from the 7th cluster already).

Given Q the maximum number of clusters and $d_1, \ldots d_K$ the arcsine-transformed degrees in a given comparison, we can write the model:

$$
\begin{array}{rcll}
\alpha & \sim & \mathrm{Gamma}(2, 4) & \\
v_q|\alpha & \sim & \mathrm{Beta}(1, \alpha) & \\
w_1 & = & v_1 & w \text{ are the weights for the clusters} \\
w_q & = & v_q \prod_{r=1}^{q-1}(1 - v_r) & (1 < q < Q) \\
w_Q & = & \prod_{r=1}^{Q-1}(1 - v_r) & \\
z|w & \sim & \mathrm{Categorical}_Q(w) & \text{indicative vector of length K} \\
\mu & \sim & \mathcal{N}(\frac{\pi}{4}, 0.5) & \text{vector of means of Gaussians (length } Q) \\
\sigma & \sim & \mathrm{Gamma}(1.5, 4) & \text{vector of sd of Gaussians (length } Q) \\
d_k|z_k, \mu, \sigma & \sim & \mathcal{N}(\mu_{z_k}, \sigma_{z_k}) & \text{likelihood of each component} \\
d_k|w, \mu, \sigma & \sim & \sum_{q=1}^{Q} w_q \mathcal{N}(\mu_q, \sigma_q) & \text{marginal likelihood}
\end{array}
$$

We can already write the log-likelihood of the clustering (for a given scale):

$$\ell_{\text{cluster}} = \sum_k \underset{q}{\mathrm{LSE}} \left(\log w_q + \log f(d_k|\mu_q, \sigma_q)\right)$$

where LSE is the log-sum-exp operation.

For the participants' judgment, we assume that the acceptability tracks the probability of an item being in the same cluster as the largest item in the comparison class, conditional on the smallest and largest being in different clusters:

$$P(z_i = z_K | z_1 \neq z_K) = \left| \begin{array}{cl} 0 & \text{if } i = 1 \\ 1 & \text{if } i = K \\ \frac{P(z_i = z_K \neq z_1)}{P(z_K \neq z_1)} & \text{otherwise} \end{array} \right.$$

Let's decompose the numerator in cases where i is neither 1 nor K. The degrees of the different elements are assumed to be independent, so:

$$\begin{aligned} P(z_i = z_K \neq z_1) &= \textstyle\sum_q \frac{w_q f(d_i|\mu_q,\sigma_q)}{f(d_i|\mu,\sigma,w)} \frac{w_q f(d_K|\mu_q,\sigma_q)}{f(d_K|\mu,\sigma,w)} \left(1 - \frac{w_q f(d_1|\mu_q,\sigma_q)}{f(d_1|\mu,\sigma,w)}\right) \\ &= \textstyle\frac{1}{f(d_i|\mu,\sigma,w) f(d_K|\mu,\sigma,w)} \sum_q \exp a_{i,q} \end{aligned}$$

$$a_{i,q} = 2 \log w_q + \log f(d_i|\mu,\sigma,w) + \log f(d_K|\mu_q,\sigma_q) + \text{log1m} \frac{w_q f(d_1|\mu_q,\sigma_q)}{f(d_1|\mu,\sigma,w)}$$

Similarly for the denominator:

$$P(z_K \neq z_1) = \textstyle\frac{1}{f(d_K|\mu,\sigma,w)} \sum_q \exp b_q$$

$$b_q = \log w_q + \log f(d_K|\mu_q,\sigma_q) + \text{log1m} \frac{w_q f(d_1|\mu_q,\sigma_q)}{f(d_1|\mu,\sigma,w)}$$

Finally, we can write the predicted acceptability from which we can derive the likelihood of participant n's response y_i:

$$\log \text{Acc}(d_i) = \operatorname*{LSE}_q a_{i,q} - \operatorname*{LSE}_q b_q - \log f(d_i|\mu,\sigma,w)$$

$$y_i \sim \mathcal{N}^{[0,1]}(\text{Acc}(d_i), \epsilon_n)$$

ϵ_n is specific to participant n, and (μ, σ, w) are specific to n and a particular comparison class.