Context, scale structure, and statistics in the interpretation of positive-form adjectives

Daniel Lassiter and Noah Goodman, Stanford Psychology

Abstract of paper to be presented at *Semantics & Linguistic Theory 23*, UCSC, May 2013

Relative adjectives (RAs) in the positive form exhibit vagueness and context-sensitivity. We suggest that these phenomena can be explained by the interaction of an unsaturated threshold variable in the positive form with a probabilistic model of pragmatic inference. We describe a formal model of utterance interpretation as coordination, which jointly infers the value of the threshold variable and the intended meaning of the sentence. We report simulations exploring the effect of background statistical knowledge on adjective interpretation in this model. Motivated by these simulation results, we suggest that this approach can account for the correlation between scale structure and the relative/absolute distinction while also allowing for exceptions noted in previous work. Finally, we argue for a probabilistic explanation of why the sorites paradox is compelling with RAs even though the second premise is false, and show that this explanation predicts [K07]'s observation that the sorites is less plausible with absolute adjectives.

1. Relative adjectives and coordination. The meanings of RAs in the positive form are highly context-dependent: a cheap house is more expensive than an expensive book, and what counts as 'cheap' for a house is sensitive to the distribution of costs in a reference class. RAs are also *vague*, as evidenced by borderline cases and susceptibility to the sorites paradox.

(1) a. A house that costs \$10,000,000 is expensive (for this neighborhood).

b. A house that costs \$1 less than an expensive house is also expensive.

c. ∴ A house that costs \$1 is expensive (for this neighborhood).

We adopt a degree semantics in which adjectives relate individuals to a threshold value, schematically: $[\![A]\!] = \lambda x \lambda \theta_A[\mu_A(x) > \theta_A]$, where θ_A is a degree on A's scale and $\mu_A(x)$ is the measure of x on this scale. We assume that the lexical entry of a gradable adjective contains only a specification of the relevant scale — an ordered set of degrees, with antonyms related by reversal of the ordering. The positive form is not marked by a covert *pos*; instead, when no degree operator is present θ_A is left unsaturated and (if necessary) type-shifted to allow composition to proceed. So, for example, *Al is tall* denotes the function $\lambda \theta_{tall}[\mu_{tall}(\mathbf{Al}) > \theta_{tall}]$. Pragmatic inference is then required to determine a value for the threshold variable so that the sentence can be assigned a propositional meaning. These innovations are modeled on the treatment of free anaphors in variable-free semantics [J99], but could be replaced with a *pos*-based analysis with some small complications (cf. [K07:7]).

Building on previous game-theoretic [L69,C96,F08] and Bayesian [B12,FG12] accounts of linguistic coordination, we propose that speakers and listeners maintain models of each others' utterance planning and interpretation processes and use these models to infer the values of unsaturated variables using Bayesian inference. We define the *literal Bayesian listener* L_0 to be one who responds to an utterance u by conditioning on its truth without reasoning pragmatically: $P_{L_0}(A|u) \propto \delta_{u\ is\ true} \times P_{L_0}(A)$. E.g., if $u = Al\ is\ tall$ and $\theta_{tall} = 6'$, the posterior is $P_{L_0}(\cdot \mid \mu_{tall}(\mathbf{Al}) > 6')$.

Let the QUD be *How tall is Al?*, and assume that the speaker knows the true answer. Speaker and listener share the goal of coordinating utterance and interpretation so as to maximize the probability that the listener will choose the correct answer to the QUD. We thus define the utility of u for a *reflective speaker* S_1 to be proportional to its informativity to the literal listener L_0 about the true answer $\mathcal{A}$, minus a non-negative cost $C(u)$. Informativity is quantified as surprisal [S48,FG12].

(2) $$\mathbb{U}_{S_1}(u; \mathcal{A}, \theta_{tall}) \propto \log(1/P_{L_0}(\mathcal{A} \mid u, \theta_{tall})) - C(u)$$

We assume that agents sample actions using a soft-max rule with parameter $\alpha > 0$ [L59,SB98].

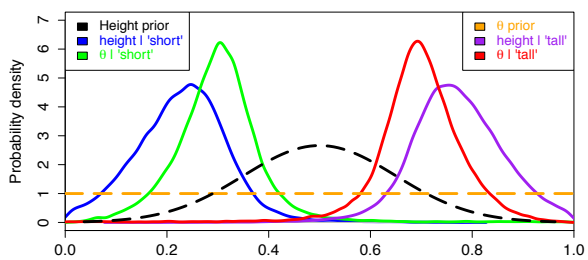
$$P_{S_1}(u) \propto \exp(\alpha \times \mathbb{U}_{S_1}(u; \mathcal{A}, \theta_{tall})) \tag{3}$$

The *reflective listener* L_1 interprets utterances using Bayesian inference, taking into account what S_1 would be likely to say given θ_A and $\mathcal{A}$ as well as the prior probability of each. (The technique of passing variables between speaker and listener to evaluate candidate interpretations is due to [B12].)

$$P_{L_1}(\mathcal{A}, \theta_A \mid u) \propto P_{S_1}(u \mid \mathcal{A}, \theta_A) \times P_{L_1}(\mathcal{A}) \tag{4}$$

(4) is derived using the model assumption that $\mathcal{A}$ and θ_A are independent and that $P_{L_1}(\theta_A)$ is uniform — i.e., the listener has no prior information about the meaning of *tall* except the structure of its scale. Background knowledge about heights is captured by $P_{L_1}(\mathcal{A})$. We treat overt comparison classes (*tall for an adult man*) as specifying a prior on $\mathcal{A}$ to be used for the purpose of interpreting u.

Suppose now that u = *Al is tall*. If θ_{tall} is extremely low — say, 1 inch — then u is completely uninformative, and $P_{S_1}(u \mid \mathcal{A}, \theta_A)$ is very low. The informativity of u increases with θ_{tall}, and would lead to a preference for extremely high values if it were not balanced out by low prior probability of the utterance's truth under such values. The posterior reflects this balancing process: in effect, interpretations are preferred which make Al fairly tall, but not implausibly so.



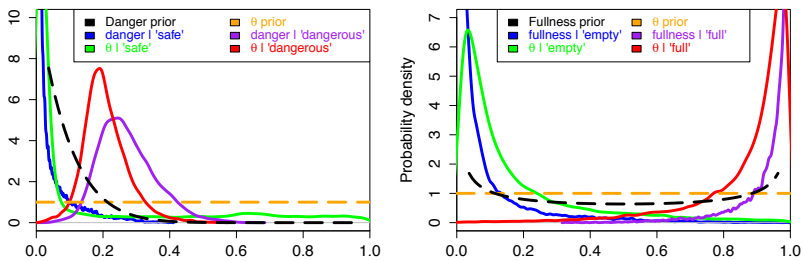
We simulated the joint posterior of $\mathcal{A}$ and θ_A given u = *Al is tall/short*, assuming that the heights of adult men are normally distributed.[1] Results are plotted in the figure to the left. The interpretations depend on the statistics of heights: the inferred meaning of *tall*, for instance, is a distribution centered around 1.3 standard deviations above the prior mean. Background knowledge thus interacts with lexical meaning and the pragmatic preference for informativity to yield a context-sensitive probabilistic meaning.

2. Absolute adjectives. A topological property of scales has been claimed to influence the meanings of positive-form gradable adjectives dramatically: those on closed scales are usually anchored to (one of) the endpoints(s) [KM05]. [K07] points out that norm-based theories of the positive form predict that relative meanings should be generally available. He argues instead that the meaning of a positive-form adjective associated with a bounded scale is *always* anchored to one of the endpoints. The proposed theoretical explanation invokes a pragmatic principle of *Interpretive Economy* (IE): 'Maximize the contribution of the conventional meanings of the elements of a sentence the the computation of its truth conditions'. [K07] uses IE to account for the fact that the sorites is less plausible with absolute adjectives (AAs). He argues that AAs are not vague, though they do admit of a qualitatively different kind of uncertainty — imprecision [L99].

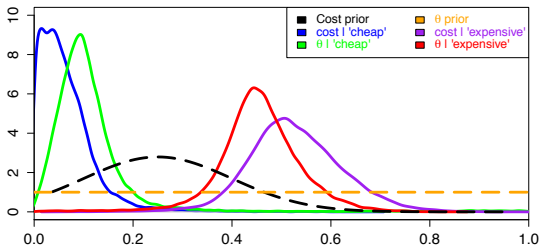
This analysis is insightful, but there are lingering problems. First, the theoretical status of IE is unclear: it is 'an optimization principle left unsupported by a theory of optimization' [P08:5]. In addition, a number of apparent counter-examples have been pointed out, including the RAs *bald* [K07], *likely/probable* [L10], and *(in)expensive* [K07,L10], all of which appear to live on bounded scales; [M11] also points to relative uses of *full*. It is possible to exclude the intuitive endpoints from these scales lexically (cf. [K07:34-5]), but the explanatory power of the theory is diminished if counter-examples can be eliminated in this way, and this would not account for relative uses of *full*.

1 All simulations reported here used Markov Chain Monte Carlo techniques to draw 30,000 samples from $P_{L_1}(\mathcal{A}, \theta_A \mid u)$ with $\alpha = 4$, $cost(u) \propto 2 \times length(u)$, a burn-in of 5000 samples, and a lag of 100. Plots show the kernel density of the relevant variables. The alternative actions considered are to say nothing or to use the adjective actually uttered or its antonym.

Our approach, in contrast, is embedded in a general pragmatic theory and is able to explain the correlation in question while also allowing for deviations. We suggest that the relevant mathematical property is a measure-theoretic one which generalizes the topological property of boundedness: prototypical relative interpretations arise with distributions which fall steadily to zero, while prototypical absolute interpretations arise with measures for which most of the prior mass is skewed toward some particular point, e.g., an upper or lower bound. The following figures illustrate using a prior on degrees skewed toward a lower bound (left) and one which is skewed equally toward both endpoints and away from the center (right). We suggest that the left panel captures key qualitative features of min/max adjective pairs such as *dangerous*/*safe*, and the right does the same for max/max pairs such as *full*/*empty*. This also accounts for the intuition that the meanings of AAs are less uncertain than those of RAs: the simulated posteriors of θ_{safe} and θ_{full} have much lower variance than that of θ_{tall}.



Our model predicts that adjectives on closed scales will receive relative interpretations if the prior has an appropriate shape. We illustrate with adjectives of cost. We suggest that the prior density of the lower bound should be small but non-zero; this corresponds to the observation that some things are free, but the cost of items in a reference class will usually be shifted substantially away from the zero point and trail off slowly. The figure to the left shows simulation results for a prior of this form. Even though the scale is lower-bounded in our extended sense (prior density > 0 at the minimum point), an object must have cost substantially greater than zero in order to have significant probability of counting as 'expensive'.



3. The sorites. Our theory captures the core insight of previous probabilistic accounts of vagueness [E99,FB10,L11], but improves on them by making precise predictions about the contextual meaning of the positive form. Our account of the sorites is related to these and to [K07]'s.

We return to the 'tall' simulation reported in §1. With $\mu_{tall}(x)$ sampled from $P(\cdot \mid u = \text{'}x \text{ is tall'})$ and $\mu_{tall}(y) = \mu_{tall}(x) - \varepsilon$, the probability that $\mu_{tall}(x) > \theta_{tall} \geq \mu_{tall}(y)$ is small when ε is small enough. In other words, with a small gap between individuals in a sorites sequence ($\varepsilon = .01$) there is approximately a 95% chance that a randomly chosen 'tall' individual will make the inductive premise true in our simulations. This feature makes the inductive premise (1b) intuitively compelling; but since its probability is less than 1, repeated use as a premise does not preserve high probability.

Absolute adjectives are different, essentially because their thresholds have lower posterior variance. In our simulations using the 'safe' prior, for example, the inductive premise has much lower probability ($\approx .65$ with $\varepsilon = .01$). Our model suggests that the difference in sorites susceptibility between RAs and AAs may be a difference in the *degree* of uncertainty remaining in $P(\theta_A \mid u)$, rather than a qualitative difference between two kinds of uncertainty.

References [B12] Bergen et al., *CogSci 2012* [C96] Clark, *Using Language* [E99] Edgington in *V'ness: A Reader* [F09] Franke, ILLC diss. [F12] — in *Ams.Colloq.2011* [FB10] Frazee & Beaver in *Ams.Colloq.2010* [FG12] Frank & Goodman, *Science* 336 (6084) [J99] Jacobson, *L&P* 22(2) [K07] Kennedy, *L&P* 30(1) [KM05] — & McNally, *Lg.* 81(2) [L59] Luce, *Individual Choice Behavior* [L69] Lewis, *Convention* [L99] Lasersohn, *Lg.* 75(3) [L10] Lassiter in *SALT 20* [L11] — in *V'ness in Communication* [M11] McNally in *V'ness in Communication* [P08] Potts ms., IE, Schelling points & evolutionary stability [S48] Shannon in *Bell Sys.Tech.J.* 27(3) [SB98] Sutton & Barto, *Reinforcement Learning*