Meaning and Use of Gradable Adjectives: Formal Modeling Meets Empirical Data

Ciyang Qing (qciyang@gmail.com)  Michael Franke (m.franke@uva.nl)
Institute for Logic, Language and Computation, University of Amsterdam
Science Park, Kruislaan 107, 1098 XG Amsterdam, The Netherlands

Abstract

The meaning of gradable adjectives is highly context-dependent, and is notoriously difficult to capture precisely. Recent work in theoretical linguistics suggests that the way we use gradable adjectives can be explained in terms of optimal language use. To test this hypothesis we formulate a probabilistic speaker model that combines ideas from Bayesian approaches to pragmatic reasoning as social cognition with broader optimality considerations, as suggested by evolutionary linguistics. We demonstrate that, despite its simplicity, the model explains empirical data on the applicability of adjectives in context astonishingly well.

Keywords: gradable adjectives; context-dependence; natural language production/generation; Bayesian cognitive modeling

Introduction

The meaning and use of gradable adjectives like tall, big, dark or full is elusive in manifold ways and continues to inspire research in not only linguistics and psychology, but also machine learning and other related fields. It is notoriously difficult to pin down the meaning and use of gradable adjectives mainly because of their context-dependence and vagueness. Whether a sentence like “John is tall” is felt to be true or pragmatically appropriate depends on the context in which the sentence is used, in particular on a contrasting set of individuals against which John is to be compared (in terms of his height). The goal of the work presented here is to shed light on this immediate context-dependence of gradable adjectives.

Building on previous work in the same direction (Barner & Snedeker, 2008; Schmidt, Goodman, Barner, & Tenenbaum, 2009; Syrett, Kennedy, & Lidz, 2010), we investigate the dependence of intuitive judgements of applicability of gradable adjectives on statistical properties of a visually presented comparison class. The key novelty of the present proposal is a fully predictive probabilistic model that formalizes the idea that the applicability of gradable adjectives is determined by pragmatic reasoning about optimal language use in context. The model we present enriches previous Bayesian approaches to pragmatic reasoning in terms of social cognition (e.g. Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Lassiter & Goodman, 2013) with ideas borrowed from evolutionary game theory (Potts, 2008; Franke, 2012).

More concretely, the model presented here is superficially similar, but conceptually distinct from the interpretation-based model of Lassiter and Goodman (2013). In contrast to the latter, the model we present here makes straightforward predictions about both comprehension and production. The key assumption in our modeling is that the speaker’s pragmatic reasoning is anchored in considerations of global optimality of the conventional usage conditions. We argue that it is the production side that is responsible for judgements of pragmatic applicability, and it is this component of the model that we test empirically. We fit our model to data that was gathered by Solt and Gotzner (2012) (for a quite different purpose) and test model predictions against data from our own replication of their experiment. Despite its simplicity, our optimality-based model explains the data astonishingly well.

The next section first introduces our speaker-based probabilistic model. Thereafter we detail Solt and Gotzner’s (2012) experiment and our replication. We discuss the model’s fit to the data and conclude with a critical reflection.

Optimal Use of Gradable Adjectives

One of the most fundamental reasons of using gradable adjectives is to ensure communicative success when trying to describe a referent, e.g., such as to pick it out, or learn about its properties. To make this intuition more concrete, we adopt a degree-based approach to the semantics of gradable adjectives (Kennedy & McNally, 2005; Kennedy, 2007). Degree-based approaches hold that a sentence of the form “x is A” is true iff the degree $d_A(x)$ to which object x has property A exceeds a contextually given standard of comparison $\theta$, i.e., $d_A(x) \geq \theta$. However, exactly how $\theta$ is conventionally derived from the context is left open. Following the general tenet of evolutionary linguistics that language conventions are shaped to achieve optimal communicative success, we propose to fill this gap by defining which values of $\theta$ are optimal in a given context. The motivating idea behind our production model is then that speakers employ a standard of comparison $\theta$ with a probability proportional to the communicative efficiency of using $\theta$ as a general convention.

Specifically, we measure communicative efficiency of $\theta$ in terms of the extent to which the utterance “x is A” would help resolve the (possibly implicit and hypothetical) Question under Discussion “how is A?” against the background of a contextually given comparison class of objects with varying levels of A-ness. We use $p$ to denote the common knowledge about prior probability distribution of A-degrees in the comparison class. For example, if the conversation is about basketball players’ heights, then $p$ is the prior distribution of the height of a basketball player from one’s common world knowledge. (Here we only consider discrete degree scales but it is easy to adapt the model to continuous density functions.)

When the conventional standard of comparison $\theta$ is fixed, a literal listener $p_0$, upon hearing the utterance “x is A,” learns from its semantic truth that the actual degree $d_A(x) \geq \theta$. Thus his new belief about $d_A(x)$, denoted as $p_0(d_A(x) | A; \theta)$, is the conditional probability $p(d_A(x) | d_A(x) \geq \theta)$. If on the other hand the speaker says nothing, the literal listener has no additional information and thus his belief stays the same:
The communicative efficiency of using $\theta$ as a conventional standard of comparison for a comparison class, is then defined as the speaker’s expected chance of success in making the listener believe in the actual degree $d_A(x)$ of an individual $x$ randomly chosen from that comparison class. For example, to measure how efficient a standard $\theta$ of “tall” is for describing basketball players, we calculate on average how likely the listener can update his prior with the information $d_A(x) \geq \theta$, which increases his chance of believing in the actual degree.

This formula captures a general tradeoff between informativity and applicability (c.f. Lasitter & Goodman, 2013): when the threshold $\theta$ is high, $\rho_0(d_A(x) \mid A; \theta)$ is high when $d_A(x) \geq \theta$, so the positive form is very informative. E.g., if $\theta = 2.2m$, the listener would have a good sense of the true height of a basketball player when he is described as “tall.” However, the positive form will seldom be applicable, since few individuals will have degrees that exceed the threshold.

The first summand corresponds to individuals whose $A$-degrees are lower than the conventional standard $\theta$ and thus the positive form $A$ cannot be truthfully asserted. The second summand corresponds to individuals whose $A$-degrees are no less than $\theta$. The speaker can truthfully utter “$x$ is $A$” and the listener can update his prior with the information $d_A(x) \geq \theta$, which increases his chance of believing in the actual degree.

We further derive a pragmatical listener’s interpretation rule by applying Bayes’ rule:

$$\rho(d \mid A; \lambda, c) \propto p(d) \cdot \sigma(u \mid d; \lambda, c),$$

but we will focus here on the production rule (4).

**Empirical Data**

In order to test the predictive power of the above production model, we collected participants’ intuitive judgements of the pragmatic applicability of the positive forms of several adjectives when confronted with comparison classes of varying statistical composition. Our design is that of Solt and Gotzner (2012), with minor modifications. We will introduce our replication first and then mention these minor differences.

**Participants, Materials and Methods**

96 US participants were recruited via Amazon’s Mechanical Turk. Each of them received $0.25 for the experiment.

We tested intuitive applicability judgements for four gradable adjectives: big, dark, tall and full. For each adjective, we presented contexts of 36 items. Each item instantiuated the adjective in question to one out of 14 possible degrees (balls varying in size, grey rectangles varying in lightness, cartoon characters varying in height, glasses varying in water level; see Fig. 1a). We chose mostly abstract items so as to minimize the effect of participants’ background world knowledge.
Stimuli were designed to make all 13 differences between adjacent degrees perceptually uniform.

We included 4 kinds of contextual prior distributions in our experiment. Each context consisted of 36 items spanning over 11 out of the 14 degrees. The baseline, left-skewed and right-skewed priors span over the lower 11 degrees with different distributions, and the moved prior spans over the upper 11 degrees (4th–14th) and has the same shape of distribution as the baseline. Fig. 1b shows the number of items for each degree in the 4 distributions.

Each participant finished 4 trials. In each trial they saw a context corresponding to 1 of the 4 adjectives under 1 of the 4 priors and were asked to check all items for which they would use the adjective in the given context (Fig. 2). We used a Latin square design for adjective-prior combinations within the 4 trials and counterbalanced the order of adjectives.

Results of their experiment are shown in Fig. 3 (blue lines). We can see that the result of our replication is close to theirs in most conditions, except for the baseline and left-skewed conditions for tall. Since our main purpose here is to use these data to test our model, we skip reporting further statistical analysis of the data themselves in the interest of space.

**Parameters Learning and Model Validation**

Our model has free parameters: \( \lambda \) (rationality) and \( c_A \) (absolute applicability of adjective \( A \)). We will use Bayesian inference (MacKay, 2003) to learn likely values of these parameters from the data of Solt and Gotzner (2012), and then test the model’s predictions on our own replication.

We assume the following binomial process that generates data in both experiments: for each adjective \( A \) and prior \( p \),

\[
    n_i^{A,p} \sim \text{Binomial}(N_i^{A,p}, \sigma_p(A | d_i; \lambda, c_A)), \tag{6}
\]

where \( n_i^{A,p} \) is the number of items of degree \( d_i \) checked by participants in the condition with adjective \( A \) and prior \( p \), and \( N_i^{A,p} \) is the total number of items of degree \( d_i \) in this condition.\(^1\) Hence, for a given adjective, \( \lambda \) and \( c_A \), for each 1 of the 4 priors, our model makes predictions for all 11 degrees. Thus the model makes 44 predictions for each adjective.

**Parameters Learning** We assume that \( \lambda \) is a constant, while each adjective has its own parameter \( c_A \). This is because \( \lambda \) is the general degree of rationality in our sample population, whereas different adjectives could have different senses of general applicability \( c_A \) depending on their lexical properties. With this, we use the following priors:

\[
    \lambda \sim \text{Unif}(0, 100) \quad c_A \sim \text{Unif}(-1, 0), \tag{7}
\]

where \( A \in \{ \text{big, dark, tall} \} \). We draw 8000 samples (after a burn-in period of 9000 samples) from the posterior distribution \( P(\lambda, c | D_{SG}) \), i.e., we make a joint inference of \( \lambda, c_{\text{big}}, c_{\text{dark}}, c_{\text{tall}} \) from the dataset \( D_{SG} \) (Solt & Gotzner, 2012). For these posterior samples of parameters, we have

\(^1\)Note that we allowed participants to check none of the pictures, and some participants did not check all the pictures with the highest degree. In order to take these possibilities into account, we introduce an unobserved maximal degree \( d_{12} \) with prior probability \( p(d_{12}) = 0 \). Since this degree corresponds to a \( \theta \) according to which the positive form is never used, the utility associated with it is rather small. Nevertheless, the soft-max function will assign a small non-zero probability to it, and hence the model always predicts that \( d_{12} \) might have a small probability not to be checked.
\[ \lambda = 48.23, \quad \text{sd} = 1.14; \quad \bar{c}_{\text{big}} = -.064, \quad \text{sd} = .003; \quad \bar{c}_{\text{tall}} = -.024, \quad \text{sd} = .003; \quad \bar{c}_{\text{dark}} = -.054, \quad \text{sd} = .002. \]

Since Solt and Gotzner (2012) did not include full in their experiment, we cannot learn the parameters directly from their dataset. Instead, we use the posteriors from their dataset to constrain the parameter \( \lambda \):

\[ \lambda_{\text{full}} \sim \text{Norm}(48.23, 1.14) \quad c_{\text{full}} \sim \text{Unif}(-1, 0). \]  

We get \( \lambda_{\text{full}} = 46.67, \quad \text{sd} = .904; \quad \bar{c}_{\text{full}} = -.158, \quad \text{sd} = .066. \)

**Model Validation**  We validate our model in two ways.

First, we use *Bayes model averaging* (Hoeting, Madigan, Raftery, & Volinsky, 1999)

\[ \sigma(u \mid d; D_{SG}) = \sum_{\lambda, c} P(\lambda, c \mid D_{SG}) \cdot \sigma(u \mid d; \lambda, c), \]  

(9)

to compute the model’s predictions after it learns the free parameters from \( D_{SG} \). The predictions are shown in Fig. 3 (pink lines), and Fig. 4 shows the relation between model predictions and participants’ choices for each adjective on the replication dataset \( D_{\text{rep}} \). Model predictions correlate well with observations (\( R^2_{\text{big}} = .97, \quad R^2_{\text{dark}} = .98, \quad R^2_{\text{tall}} = .94, \quad R^2_{\text{full}} = .95, \)

with overall \( R^2 = .96 \) and \( p < .001 \) for all cases). Correlations remain highly significant even when we only keep those data points for which our model’s prediction is within the range of \((0.05, 0.95)\) (\( R^2_{\text{big}} = .93, \quad R^2_{\text{dark}} = .94, \quad R^2_{\text{tall}} = .88, \quad R^2_{\text{full}} = .90, \)

with overall \( R^2 = .90 \) and \( p < .001 \) for all cases). This suggests that our model does capture the general trend of participants’ choices, rather than by simply assigning extreme probabilities to extreme degrees.

Second, in order to better diagnose the model’s predictions for each data point, we investigate the posterior predictive distribution (c.f. Kruschke, 2011). Concretely, for each of the 8000 samples of parameters drawn from the posterior distribution described before, we use the binomial generative process (6) to generate a new dataset. Thus in the end we have 8000 simulated datasets. Then for each adjective, each prior and each degree, we look at the number of items checked in the actual dataset (either \( D_{SG} \) or \( D_{\text{rep}} \)) and record the frequency of this actual observation in the simulated datasets.

Finally, we calculate the posterior predictive credibility value as the sum of relative frequencies of all observations that occurs no more often than the actual observation in the simulated datasets. This posterior predictive credibility value
then captures the estimated maximal threshold on credibility thresholds under which the observed data would not contradict our model. Concretely, a value of .05 means that the observed data falls within a 95% HDI interval of the posterior predictive; a value of 1 means that the observation was the mode of our posterior predictive sampling.

The posterior predictive credibility values are shown in Table 1. We can see that the model’s predictions generally pass the predictive check. For those degrees where the model fails to meet a critical threshold of .05 on both data sets (marked in bold), we note two possible sources of bad fit: (1) The discrepancy between the two datasets due to noise. As a result, the model fitting the training set can fail to generalize to the test set. We also want to emphasize here that since the model needs to fit all degrees under all priors simultaneously, noises in one degree might influence performance on another degree as well. (2) The discrepancy between the two datasets due to differences in stimuli. For instance, the stimuli for tall generally have greater height-to-width ratio in the experiment by Solt and Gotzner (2012) than in our replication. As a result, the general applicability parameter is probably greater for their contextual comparison classes than for those in the replication dataset. This might explain why the model generalizes well in the moved condition but performs poorly on the baseline and left-skewed conditions.

The two validation methods both suggest that the model in general captures participants’ applicability judgements well.

**Need for General Applicability** It remains to be checked whether the data can be explained reasonably well in terms of contextually optimal language use alone, or whether there is reason to believe that there are further absolute criteria that regulate the use of adjectives, beyond the immediate statistical properties of the comparison class. To test this, we can use the Savage-Dickey method to calculate the Bayes factor in favor of a model $M_{c=0}$ which assumes that there is no absolute re-alignment of $\theta$ necessary with the model $M_{c<0}$ we fitted to the data (e.g. Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). By this method we compute the factor with which our posterior credence should shift towards $M_{c=0}$ as:

$$
P(c = 0 | M_{c<0}, D) / P(c = 0 | M_{c<0})
$$

The denominator of the right-hand fraction is 1, but the numerator is so small that our finite sampling procedure cannot
assign a non-zero value to it. This holds true for all adjectives involved. Consequently, we should conclude that the data suggests strongly that our model needs some absolute upward shifting of \( \theta \). But, as noted before, the estimated posterior values of \( c \) vary from adjective to adjective. As expected from formal semantic accounts that distinguish absolute adjectives like full from relative adjectives like tall, the data suggests that the absolute re-shifting needed to account for the applicability of full is substantially higher.

**Conclusion**

Combining the idea of pragmatic reasoning as social cognition and optimality considerations from evolutionary linguistics, the presented model is a fully generative cognitive model that successfully predicts intuitive applicability judgements of gradable adjectives in various contexts.

Despite the model’s noteworthy empirical success, it should be noted that everyday language use is far more complex and implicit than our highly simplified and controlled experiments suggest. Hence more work needs to be done to ultimately account for naturalistic language data. For example, we effectively assumed that participants are fully aware of the exact distribution of degrees in the comparison class, which is too idealized even for the artificial contexts in our experiment. A more comprehensive model would include participants’ latent representations of degrees and their estimated contextual distribution. Preliminary results from such modeling suggest that this improves predictive power. This is so because latent priors over degrees, estimated separately for each adjective-prior pair, can capture participants expectations about unrepresented degrees as well. Ever taller basketball players, though increasingly unlikely, are conceivable, while glasses will reach a maximally saturated degree of fullness. That’s why estimated latent priors are prone to improve predictive power because they can accommodate this kind of conceptual knowledge implicitly.

On a conceptual level, we advanced the hypothesis that the use of gradable adjectives is driven by optimality of descriptive language use. This contrasts with explanations based on optimal contextual categorization (e.g. Schmidt et al., 2009) or based on referential language use (e.g. Franke, 2012; Gatt, van Gompel, van Deemter, & Kramer, 2013). It is possible to think that the identification of \( x \)’s degree of A-ness is conceptually prior but subservient also to optimal categorization and the use of gradable adjectives in referential expressions, but more research is needed to explore this connection.

**Acknowledgements**

We thank Stephanie Solt and Nicole Gotzner for sharing their data, and Will Frager for help with our experiments. Thanks to Noah Goodman and Dan Lassiter for discussion. Michael Franke was supported by NWO-VENI-grant 275-80-004.

**References**


