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* θ, i.e., $d_A(x) \geq \theta$. However, exactly how θ 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 θ are optimal in a given context. The motivating idea behind our production model is then that speakers employ a standard of comparison θ with a probability proportional to the *communicative efficiency* of using θ as a general convention.

Specifically, we measure communicative efficiency of θ in terms of the extent to which the utterance "x is A" would help resolve the (possibly implicit and hypothetical) Question under Discussion "how A is x?" 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 θ is fixed, a literal listener ρ_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 $\rho_0(d_A(x) \mid A;\ \theta)$, is the conditional probability $p(d_A(x) \mid d_A(x) \geq \theta)$. If on the other hand the speaker says nothing, the literal listener has no additional information and thus his belief stays the same:

$\rho_0(d_A(x) \mid N;\ \theta) = p(d_A(x))$.

The communicative efficiency of using θ 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 θ of "tall" is for describing basketball players, we calculate on average how likely the speaker will manage to convey the height of a random basketball player by adopting that standard. Technically, we have:

$$ES(\theta) = \sum_{d_A(x)<\theta} p(d_A(x)) \cdot \rho_0(d_A(x) \mid N;\theta) + \sum_{d_A(x)\geq\theta} p(d_A(x)) \cdot \rho_0(d_A(x) \mid A;\theta)\ . \quad (1)$$

The first summand corresponds to individuals whose A-degrees are lower than the conventional standard θ and thus the positive form A cannot be truthfully asserted. The second summand corresponds to individuals whose A-degrees are no less than θ. 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.

This formula captures a general tradeoff between informativity and applicability (c.f. Lassiter & Goodman, 2013): when the threshold θ 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 $\theta = 2.2$m, 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. Thus such a θ is on average inefficient. For lower θ the positive form is often applicable, but this time $\rho_0(d_A(x) \mid A;\theta)$ does not improve much from the prior $p(d_A(x))$. E.g., if $\theta = 1.8$m, then a basketball player being described as "tall" would tell very little about his actual height, which is inefficient as well. Hence an optimal θ should strike a good balance between informativity and applicability.

If communicative efficiency is the only factor that matters, then conversational participants should strive for contextual standards of comparison that are optimal in this respect. However, theoretical linguists give good arguments that the lexical properties of a gradable adjective also set constraints on its general applicability (e.g. Kennedy & McNally, 2005; Kennedy, 2007). For instance, suppose there is a building whose windows are open to various extent, one might be inclined to describe a window as *open* even if it is actually the least open among all the windows. In order to capture this aspect, we define the utility of conventional threshold θ as:

$$U(\theta;\ c) = ES(\theta) + c \cdot \sum_{d_A(x)\geq\theta} p(d_A(x)), \quad (2)$$

where c is a "coverage parameter" that measures the extent to which a gradable adjective's absolute sense of applicability affects the standard of comparison. The higher c the more using the adjective is preferred over not saying anything, and vice versa. A positive c means that the gradable adjective is generally applicable to every individual in the context, like in the above example for *open*. Thus a lower θ is preferred, modulo the effect of contextual optimality. In contrast, a negative c means that the gradable adjective is generally inapplicable, so a higher θ is preferred. The absolute value of c reflects the interaction between communicative efficiency and absolute general applicability. If c is close to 0, it means that communicative efficiency is the dominant factor in determining θ, and if c is away from 0, then the absolute sense trumps communicative efficiency. We include this factor to assess whether an absolute sense of applicability of adjectives critically improves empirical predictions.

Using a standard soft-max function (e.g. Luce, 1959), we capture threshold choices in production as the probability:

$$\Pr(\theta;\ \lambda, c) \propto \exp(\lambda \cdot U(\theta;\ c))\,. \quad (3)$$

The intuition is that the higher the utility, the higher the probability with which speakers would adhere to standard θ, if they use language optimally, but actual speakers might make mistakes of various sorts and thus be sub-optimal, as captured by the degree of rationality parameter λ.

The production probability of using positive form A for degree d can be naturally defined as the sum probability of all threholds no greater than d (Lassiter, 2011):

$$\sigma(A \mid d;\ \lambda, c) = \sum_{\theta\leq d} \Pr(\theta;\ \lambda, c)\ . \quad (4)$$

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

$$\rho(d \mid A;\ \lambda, c) \propto p(d) \cdot \sigma(u \mid d;\ \lambda, c), \quad (5)$$

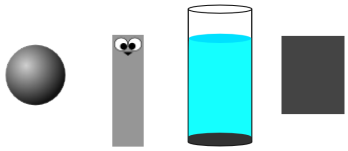
but we will focus here on the production rule (4).

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 instantiated 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.



(a) Example items

Prior	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}
baseline	1	2	3	4	5	6	5	4	3	2	1
left-skewed	2	5	6	6	5	4	3	2	1	1	1
right-skewed	1	1	1	2	3	4	5	6	6	5	2
moved	1	2	3	4	5	6	5	4	3	2	1

(b) Number of items for each degree in each prior condition

Figure 1: Stimuli used in our replication of Solt & Gotzner's study

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.

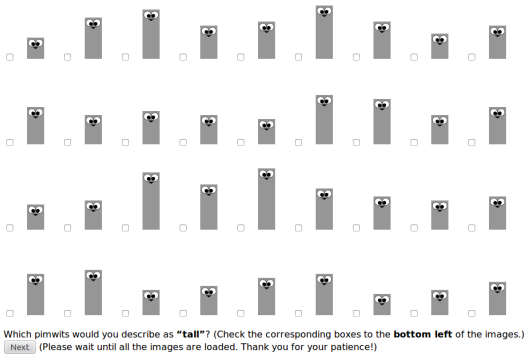


Figure 2: A sample trial

Qualitative Results The results are shown in Fig. 3. As expected, proportions of intuitive applicability judgements followed an S-shaped curve rising from lower to higher degrees. More importantly, the statistical distribution of the contextual comparison class had an apparent influence on the applicability judgements. E.g., when there are many high-degree items such as in the right-skewed condition, smaller proportion of low-degree items were chosen.

The Original Dataset The experiment by Solt and Gotzner (2012) had 194 participants in total (47 – 50 participants in each condition). Test items included *big*, *tall*, and *dark*, but also *pointy* instead of *full*. We chose *full* primarily because *pointy* is a rather unusual word and it is hard to construct items with uniformly spaced degrees of "pointiness."

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: λ (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,

$$n_i^{A,p} \sim \text{Binom}(N_i^{A,p},\ \sigma_p(A \mid d_i;\ \lambda, c_A)), \tag{6}$$

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, λ 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 λ is a constant, while each adjective has its own parameter c_A. This is because λ 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 $A \in \{\text{big}, \text{dark}, \text{tall}\}$. We draw 8000 samples (after a burn-in period of 9000 samples) from the posterior distribution $P(\lambda, \mathbf{c} \mid \mathcal{D}_{SG})$, i.e., we make a joint inference of λ, c_{tall}, c_{dark}, c_{big} from the dataset $\mathcal{D}_{SG}$ (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 θ 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_{11} might have a small probability not to be checked.

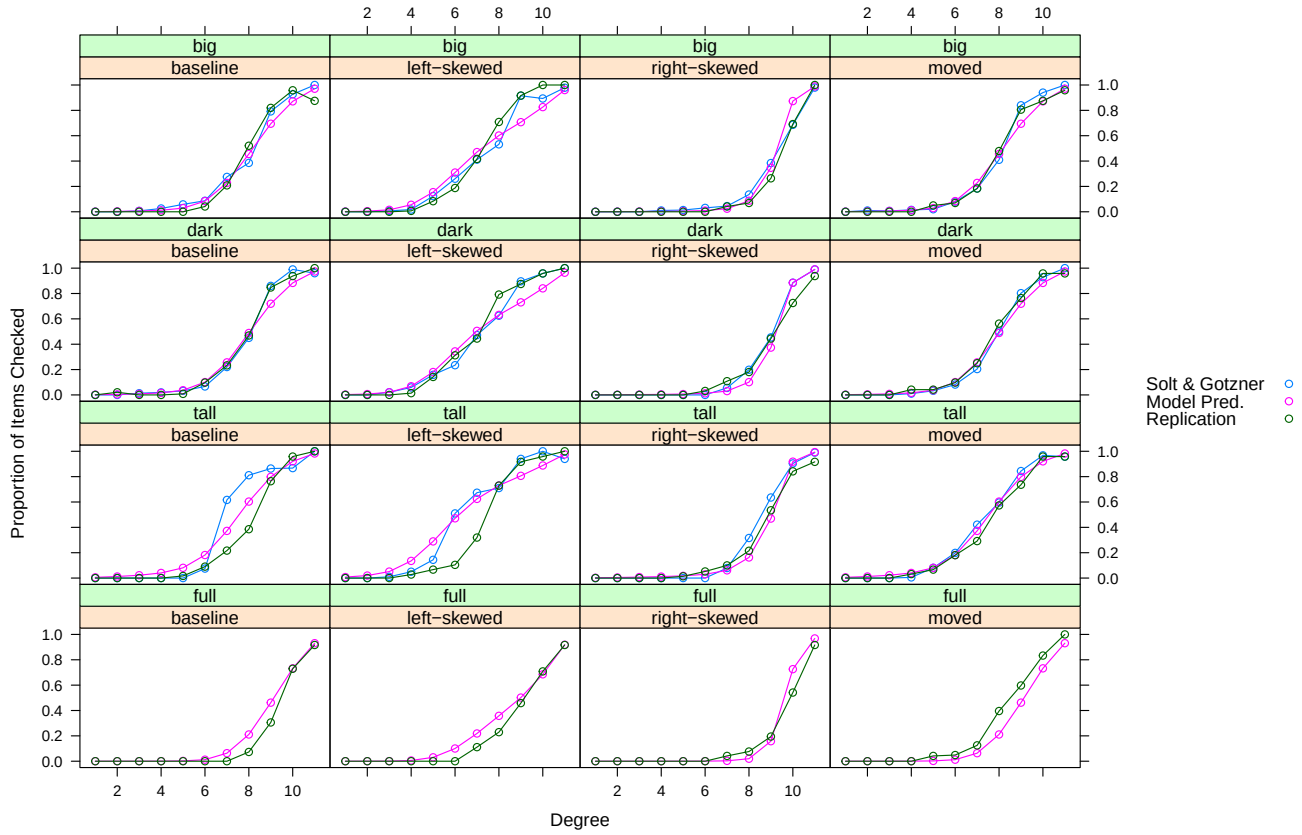


Figure 3: Observed and predicted applicability judgements for each degree for each adjective-prior pair. The blue and green curves show the observed proportions of items checked in each condition. The pink curve shows the mean posterior predictive values of the model when condition on the data by Solt & Gotzner.

$\bar{\lambda} = 48.23,\ \text{sd} = 1.14;\ \bar{c}_{\text{big}} = -.064,\ \text{sd} = .003;\ \bar{c}_{\text{tall}} = -.024,\ \text{sd} = .003;\ \bar{c}_{\text{dark}} = -.054,\ \text{sd} = .002.$

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 λ:

$$\lambda_{\text{full}} \sim \text{Norm}(48.23, 1.14) \quad c_{\text{full}} \sim \text{Unif}(-1, 0). \tag{8}$$

We get $\bar{\lambda}_{\text{full}} = 46.67,\ \text{sd} = .904;\ \bar{c}_{\text{full}} = -.158,\ \text{sd} = .006.$

Model Validation We validate our model in two ways.

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

$$\sigma(u \mid d, \mathcal{D}_{SG}) = \sum_{\lambda, \mathbf{c}} P(\lambda, \mathbf{c} \mid \mathcal{D}_{SG}) \cdot \sigma(u \mid d;\ \lambda, \mathbf{c}), \tag{9}$$

to compute the model's predictions after it learns the free parameters from $\mathcal{D}_{SG}$. 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 $\mathcal{D}_{\text{rep}}$. Model predictions correlate well with observations ($R^2_{\text{big}} = .97$, $R^2_{\text{dark}} = .98$, $R^2_{\text{tall}} = .94$, $R^2_{\text{full}} = .95$, 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 $(0.05, 0.95)$ ($R^2_{\text{big}} = .93$, $R^2_{\text{dark}} = .94$, $R^2_{\text{tall}} = .88$, $R^2_{\text{full}} = .90$, 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 $\mathcal{D}_{SG}$ or $\mathcal{D}_{rep}$) 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

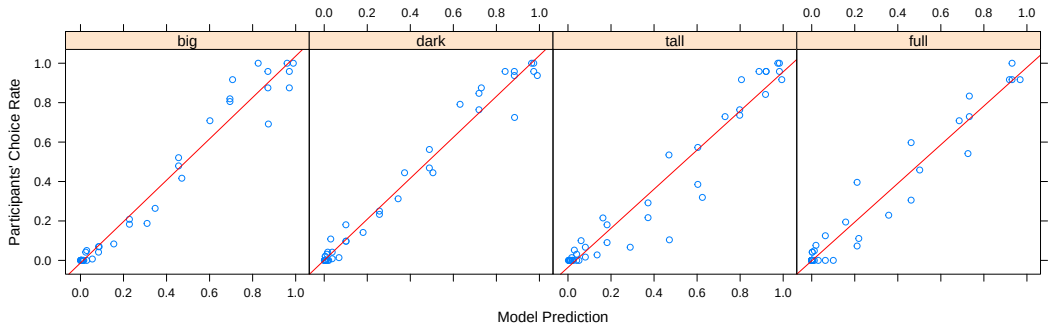


Figure 4: The relation between model predictions and participants' choices on the replication dataset

Table 1: Posterior predictive credibility values. The left value is for Solt and Gotzner's data, the right for our replication. Values in bold are those where the test values fall below a critical value of .05 for both data sets.

Adj	Prior	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}
big	baseline	1/1	1/1	1/1	.08/.63	.04/.06	.92/.08	.10/.59	.06/.32	.01/.06	.22/.21	.41/.03
big	left-skewed	1/1	.64/1	.33/.19	**.01/0**	.16/.02	.19/.01	.12/.42	.12/.24	**0/.02**	.25/.01	1/1
big	right-skewed	1/1	1/1	1/1	1/1	.17/1	.01/1	.06/.23	.01/.66	.24/.04	**0/0**	.10/1
big	moved	1/1	.27/1	.60/1	.52/.64	.58/.16	.92/.64	.12/.20	.21/.68	**0/.05**	.04/.83	.40/.17
dark	baseline	1/1	1/1	1/1	.57/.42	.87/.02	.05/.89	.19/.53	.30/.67	**0/.04**	0/.37	.39/1
dark	left-skewed	1/1	.43/1	1/.08	.65/.01	.37/.34	0/.59	.33/.24	.84/.02	.02/.16	.08/.41	.42/1
dark	right-skewed	1/1	1/1	1/1	1/1	1/1	.19/.12	.09/0	**0/0**	.01/.13	1/0	.24/.01
dark	moved	1/1	1/1	.64/1	.39/.07	.62/.80	.30/.89	.07/1	.83/.15	.04/.52	.21/.36	.64/.13
tall	baseline	1/1	.64/1	.08/.43	**0/.03**	**0/.01**	**0/.01**	**0/0**	**0/0**	.09/.48	.09/1	1/1
tall	left-skewed	1/1	.01/.18	**0/0**	**0/0**	**0/0**	.31/0	.36/0	.58/1	.01/.30	.01/1	.12/1
tall	right-skewed	1/1	1/1	1/1	.64/1	.19/1	.01/.37	.23/.08	0/.10	0/.19	.41/.01	.12/0
tall	moved	1/1	.65/1	.08/.41	0/1	.91/.74	.46/1	.17/.08	.77/.53	.30/.14	.09/1	.05/.07
full	baseline	–/1	–/1	–/1	–/1	–/1	–/.45	–/0	–/0	–/0	–/1	–/.68
full	left-skewed	–/1	–/1	–/1	–/1	–/.06	–/0	–/.02	–/.07	–/.43	–/.83	–/1
full	right-skewed	–/1	–/1	–/1	–/1	–/1	–/1	–/0	–/0	–/.38	–/0	–/.06
full	moved	–/1	–/1	–/1	–/1	–/0	–/.01	–/.02	–/0	–/.05	–/.25	–/.41

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 θ necessary with the model $M_{c\leq 0}$ 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:

$$\frac{P(D|M_{c=0})}{P(D|M_{c\leq 0})} = \frac{P(c=0|M_{c\leq 0},D)}{P(c=0|M_{c\leq 0})}.$$

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 θ. 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

Barner, D., & Snedeker, J. (2008). Compositionality and statistics in adjective acquisition: 4-year-olds interpret tall and short based on the size distributions of novel noun referents. *Child development*, *79*(3), 594–608.

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

Franke, M. (2012). On scales, salience & referential language use. In M. Aloni, F. Roelofsen, & K. Schulz (Eds.), *Amsterdam colloquium 2011* (pp. 311–320). Springer.

Gatt, A., van Gompel, R. P. G., van Deemter, K., & Kramer, E. (2013). Are we bayesian referring expression generators? In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of CogSci 35.*

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling lanuage understanding as social cognition. *Topics in Cognitive Science*, *5*, 173–184.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, 382–401.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, *30*, 1–45.

Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, *81*(2), 345–381.

Kruschke, J. E. (2011). *Doing bayesian data analysis*. Burlington, MA: Academic Press.

Lassiter, D. (2011). Vagueness as probabilistic linguistic knowledge. In R. Nouwen, R. van Rooij, U. Sauerland, & H.-C. Schmitz (Eds.), *Vagueness in communication* (pp. 127–150). Springer.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In T. Snider (Ed.), *Proceedings of the 23rd semantics and linguistic theory conference (SALT 23)* (pp. 587–610).

Luce, D. R. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Potts, C. (2008). *Interpretive Economy, Schelling Points, and evolutionary stability.* (Manuscript, UMass Amherst)

Schmidt, L. A., Goodman, N. D., Barner, D., & Tenenbaum, J. B. (2009). How tall is tall? compositionality, statistics, and gradable adjectives. In *Proceedings of CogSci 31.*

Solt, S., & Gotzner, N. (2012). Experimenting with degree. In A. Chereches (Ed.), *Proceedings of the 22nd semantics and linguistic theory conference (SALT 22)* (pp. 166–187).

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

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–189.