Proceedings of SALT 24: 23–41, 2014

Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model *

Ciyang Qing
ILLC, University of Amsterdam

Michael Franke
SfS, University of Tübingen

Abstract This paper addresses two issues that arise in a degree-based approach to the semantics of positive forms of gradable adjectives such as *tall* in the sentence *John is tall* (e.g., Kennedy & McNally 2005; Kennedy 2007): First, how the standard of comparison is contextually determined; Second, why gradable adjectives exhibit the relative-absolute distinction. Combining ideas of previous evolutionary and probabilistic approaches (e.g., Potts 2008; Franke 2012; Lassiter 2011; Lassiter & Goodman 2013), we propose a new model that makes exact and empirically testable probabilistic predictions about speakers' use of gradable adjectives and that derives the relative-absolute distinction from considerations of optimal language use. Along the way, we distinguish between vagueness and loose use, and argue that, within our approach, vagueness can be understood as the result of uncertainty about the exact degree distribution within the comparison class.

Keywords: gradable adjectives, vagueness, absolute and relative adjectives, evolutionary linguistics, probabilistic models, language production

1 Introduction

According to the degree-based approach to the semantics of gradable adjectives (e.g., Kennedy & McNally 2005; Kennedy 2007), the denotation of a gradable adjective such as *tall* is a function that maps individuals to *degrees* on an abstract *scale structure*, e.g., $[\![\text{tall}]\!] = \lambda x.\mathbf{height}(x)$. The meaning of the *positive form* of a gradable adjective, such as *tall* in the sentence *John is tall*, is taken to be the composition of the gradable adjective with a silent morpheme *pos*.

(1) $$[\![pos\ \text{tall}]\!] = \lambda x.\mathbf{height}(x) \geq \theta$$

θ is the contextual *standard of comparison* (also referred to as *threshold*).

* Many thanks to Daniel Lassiter and Noah Goodman for patiently answering our questions about their work. We are grateful also to Leon Bergen, Itamar Francez, Chris Kennedy, Malte Willer and the other participants of SALT and the Chicago semantics colloquium for insightful discussion. Thanks also to three anonymous reviewers for very helpful feedback. Michael Franke gratefully acknowledges support by NWO-VENI grant 275-80-004.

In order to fully understand the meaning and use of gradable adjectives, we need to understand how the threshold θ is determined. Clearly, θ is *context-dependent* in the sense that different thresholds are in place for *tall* when we talk about men or trees, i.e., for different so-called *comparison classes*. But the picture is even more subtle. On the one hand, for many gradable adjectives, θ appears to be vague in the sense that there can be uncertainty about θ despite perfect knowledge of the comparison class. Such adjectives are called *relative* adjectives, and their vagueness manifests itself in *borderline cases*. For example, in the expression "a tall basketball player," even though the comparison class, i.e., the set of basketball players, is explicit, one may still hesitate when deciding whether it is true for a player who is 2m tall. On the other hand, as Kennedy (2007) observes, some gradable adjectives, such as *full* and *dry*, have positive forms which are arguably not vague. For instance, a glass of water is full only when it is totally filled with water.[1] These gradable adjectives are called *absolute* adjectives. In sum, a complete theory of the meaning and use of gradable adjectives must spell out the contextual resolution of the threshold and in particular correctly predict the difference between absolute and relative adjectives.

Kennedy (2007) argues that conventional lexical semantic properties of gradable adjectives and contextual factors play a role in determining standards of comparison. He distinguishes between *open* and *closed* scale structures underlying relative and absolute gradable adjectives respectively and proposes the *Interpretive Economy* principle to spell out how lexical semantic properties determine the meaning of absolute adjectives. Subsequent evolutionary approaches (Potts 2008; Franke 2012) have tried to derive *Interpretive Economy* from more basic assumptions about goal-oriented language use, but these approaches make no concrete predictions about actual language use, in particular, the resolution of θ for relative adjectives. This is partially compensated by Lassiter & Goodman (2013) who propose a *Rational Speech-Act* (RSA) model to give precise quantitative predictions about the contextual *interpretation* of gradable adjectives that are derived from statistical properties of the contextual comparison class. However, since the RSA model is *listener-oriented*, a predictive speaker model is missing, as we will argue here. We will also argue that the RSA model might not satisfactorily explain the relative-absolute distinction.

We therefore propose a *speaker-oriented* probabilistic model that is inspired by the RSA model but also adopts the idea of an *optimal linguistic convention* from evolutionary approaches. The model makes concrete and empirically testable probabilistic predictions about the use of gradable adjectives and, by doing so, explains the robust dichotomy between absolute and relative adjectives. The following

1 In reality we often use these positive forms *loosely*, e.g., one may use *full* to describe a glass of water that is not absolutely full. We will discuss the relation between such *imprecision* and vagueness in later sections.

section 2 introduces our speaker-oriented model (SOM) in more detail. Section 3 illustrates the predictions of SOM for different scale structures and contextual priors and exposes our explanation of the absolute-relative distinction. Section 4 compares SOM with the closely related RSA model (Lassiter & Goodman 2013), before we conclude by discussing the implications of the new model and some open issues.

2 Optimal descriptive use of gradable adjectives

One common theme in both evolutionary and probabilistic pragmatics is to address language use in the broader context of social interactions between goal-oriented language users. We will adopt this functional view as well and focus on the descriptive use of positive forms, whose purpose is to convey information about the degree of some designated individual.[2] As a working example, in the following we assume that the possibly implicit *question under discussion* (QUD) is how tall John is.

The semantic problem that we are facing is whether the positive form can be applied. Thus we assume that the speaker can only choose between using the positive form (u_1) and saying nothing (u_0). This is certainly a radical simplification of real life communication. Alternatives include taking into account the antonym, compositional expressions (e.g., *neither tall nor short*) or explicit measure terms, if applicable. Nevertheless, as a first step we will adopt this minimalist setting to illustrate the main idea. We will discuss some alternatives in later sections.

As introduced before, the meaning of a positive form is relative to a contextual comparison class. Another shared feature of recent evolutionary and probabilistic accounts is to exploit the statistical information of a comparison class, in the form of a probability distribution over degrees on the scale, rather than to treat a comparison class simply as a set of individuals. This view provides a specific explanation of how a comparison class influences the meaning of positive forms, capturing the interaction between our background world knowledge and language use. For instance, when we talk about John's height, we not only know that we are comparing him against the set of male individuals, but also have some prior world knowledge about the distibution of adult male heights, $\phi(h)$. Of course, such prior knowledge is usually imprecise, since the comparison class can be implicit, and we usually do not have perfect world knowledge. Nevertheless, in this section we will consider the ideal case in which speaker and listener have an exact prior distribution $\phi(h)$ and it is common knowledge between them. In the next section we will argue that vagueness is closely related to the violation of this assumption in reality.

Let us summarize the setting so far, we have assumed that the goal is to convey the height of John. There is a commonly known prior distribution of male heights,

2 Of course, this is not the only purpose of using positive forms. For instance, positive forms can be used *referentially* to help the listener pick up the intended referent in the context (e.g., Franke 2012).

$\phi(h)$. In addition, the speaker knows John's height h_0 but the listener does not. The speaker can either use the positive form *tall* (u_1), or say nothing (u_0). Our task is to predict how likely the speaker will use the positive form to describe John, i.e., $\sigma(u_1 \mid h_0; \phi)$.

If the speaker's probabilistic knowledge of the threshold, $\Pr(\theta)$, is already known, then a natural production rule, proposed by Lassiter (2011), predicts that the probability that the speaker would use the positive form is the probability that the threshold θ is no greater than h_0.

$$\sigma(u_1 \mid h_0, \Pr) = p(\theta \leq h_0) = \int_{-\infty}^{h_0} \Pr(\theta)\, \mathrm{d}\theta \tag{2}$$

This rule can be intuitively understood as that the speaker randomly samples a threshold from the distribution $\Pr(\theta)$, compares h_0 with this θ, and uses the semantics to decide whether the positive form should be used. The remaining question is, of course, how such probabilistic knowledge of the threshold, $\Pr(\theta)$, is derived from the prior degree distribution $\phi(h)$.

Here we adopt the evolutionary perspective that $\Pr(\theta)$ is the conventional[3] linguistic knowledge formed under evolutionary pressure to efficiently communicate an individual's degrees in the comparison class $\phi(h)$. Specifically, for positive forms, the optimization problem that a linguistic community faces is to choose a threshold for the comparison class so that on average the positive form can be used to most successfully convey the degree of an individual from that comparison class.

We will use a hypothetical literal listener to evalute the *communicative effiency* of each semantic convention of the threshold θ. Recall that according to the semantics *John is tall* is true iff $h_0 \geq \theta$. We have two cases to consider. On the one hand, if $h_0 < \theta$, the speaker can say nothing, since *tall* is not true in this case. As a result, the literal listener can only use the prior information to infer John's height, so his belief about John's height is the same as the prior distribution.

$$\phi(h \mid u_0, \theta) = \phi(h) \tag{3}$$

In particular, the probability of him believing in the correct height is $\phi(h_0)$. On the other hand, if $h_0 \geq \theta$, the speaker can use *tall* truthfully to describe John and the literal listener can do an update by conditioning on its truth, which yields a new distribution.

$$\phi(h \mid u_1, \theta) = \phi(h \mid h \geq \theta) = \begin{cases} \frac{\phi(h)}{\int_\theta^\infty \phi(h)\, \mathrm{d}h} & \text{if } h \geq \theta\,, \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

3 Note that such a convention need not be *explicit*. Many conventions are formed implicitly and inductively, via reasonable generalizations from past experiences.

In particular, the probability of the literal listener believing in the correct height is $\frac{\phi(h_0)}{1-\Phi(\theta)}$, where $\Phi(\theta) = \int_{-\infty}^{\theta} \phi(h)\,\mathrm{d}h$ is the *cumulative probability* of the prior distribution $\phi(h)$ at θ.

Recall that we want to measure the communicative success of a threshold in the long run. Here John is taken to be a random individual from the comparison class, so the probability of John's height being h_0 is $\phi(h_0)$. Hence, on average we have the *expected success* rate of θ.

$$ES(\theta) = \int_{-\infty}^{\theta} \phi(h_0)\phi(h_0|u_0,\theta)\,\mathrm{d}h_0 + \int_{\theta}^{\infty} \phi(h_0)\phi(h_0|u_1,\theta)\,\mathrm{d}h_0 \tag{5}$$

The left summand corresponds to situations where the speaker has to stay silent because $h_0 < \theta$ and the literal listener can only use the prior knowledge, and the right summand corresponds to heights to which *tall* is applicable to induce a more accurate belief. Since h_0 is a bound variable in the above formula, we will simply rewrite it as h.

$$ES(\theta) = \int_{-\infty}^{\theta} \phi(h)\phi(h|u_0,\theta)\,\mathrm{d}h + \int_{\theta}^{\infty} \phi(h)\phi(h|u_1,\theta)\,\mathrm{d}h \tag{6}$$

If communicative success is all that we care about, then $ES(\theta)$ already provides us with a measure. In reality, however, language users have other goals such as reducing the speaking effort. Following Lassiter & Goodman (2013), also for better comparison (see section 4), we introduce a cost parameter of the positive form, c, to capture these other factors, and define the utility of a threshold as its expected success with the cost subtracted.

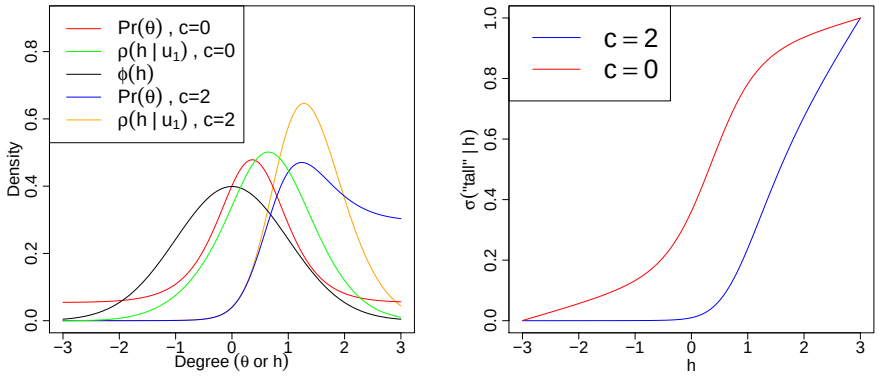
$$U(\theta) = ES(\theta) - \int_{\theta}^{\infty} \phi(h)\cdot c\,\mathrm{d}h \tag{7}$$

Note that the integral of cost starts from θ because the positive form is used only when $h \geq \theta$.

Now we are finally able to address where the linguistic knowledge $\Pr(\theta)$ comes from. From evolutionary considerations, we predict that the greater the utility of a threshold, $U(\theta)$, the more likely that people are going to use it as the convention.

$$\Pr(\theta) \propto \exp(\lambda \cdot U(\theta)) \tag{8}$$

Here we use a standard soft-max function to select the threshold sub-optimally (Luce 1959; Sutton & Barto 1998). The intuition is that people are more likely to select thresholds with higher utilities, but since they are not perfectly rational, they might make mistakes occasionally and end up with less optimal ones. The parameter $\lambda \geq 0$ is used to quantify the degree of rationality, i.e., the extent to which



(a) $\Pr(\theta)$ and $\rho(h \mid u_1)$ (b) $\sigma(u_1 \mid h)$

Figure 1 SOM predictions for Gaussian prior N(0,1), with $\lambda = 4$.

people stick to the strictly optimal threshold. If $\lambda = 0$, then $\Pr(\theta)$ reduces to a uniform distribution, meaning that there is no optimality consideration at all. When $\lambda \to \infty$, the soft-max function strictly maximizes the utility. We should assume that λ takes some value in between for actual language use. For instance, if we have three thresholds whose utilities are 1, 2, and 3, with $\lambda = 4$ we will have $\Pr(\theta_1) = \frac{\exp(4\times1)}{\exp(4\times1)+\exp(4\times2)+\exp(4\times3)} = .0003$ and similarly $\Pr(\theta_2) = .018$, $\Pr(\theta_3) = .9817$. We can see that the optimal threshold θ_3 has the greatest probability, and even the least optimal threshold θ_1 has a small probability to be selected.

Combining (2), (7) and (8), we have a full production model at our disposal, and the corresponding interpretation model for listeners can be derived by applying Bayes' rule.

$$\rho(h \mid u_1) \propto \phi'(h) \cdot \sigma(u_1 \mid h, \Pr') \tag{9}$$

Note that $\Pr'(\theta)$ and $\phi'(h)$ are correlated the same way as before, but in general the listener's prior world knowledge $\phi'(h)$ need not be the same as the speaker's. Nevertheless, as mentioned earlier, in the simplest case, we assume that prior world knowledge is in the common ground, i.e., $\phi'(h) = \phi(h)$.

Fig. 1 shows predictions by the SOM for the Gaussian distribution $\phi(h) \sim$ N(0,1), with all parameters the same for both the speaker and the listener. We can see from Fig. 1(a) that the SOM predicts that the distribution of the threshold, $\Pr(\theta)$, peaks slightly to the right of the average height, and that the posterior of height after

hearing *tall*, $\rho(h \mid u_1)$, is shifted from the height prior to the right. This corresponds well to our intuition that someone needs to be sufficiently taller than average to be described as *tall*. Also, we can see from Fig. 1(b) that the production rule of the SOM gives sensible predictions. The probability of describing someone of height h as *tall*, $\sigma(u_1 \mid h)$, roughly has an S-shaped curve.[4] Note that our model gives reasonable predictions even when the cost $c = 0$.[5]

Since $\Pr(\theta)$ is the core component of the SOM, in the next section we will focus on $\Pr(\theta)$ to better illustrate the SOM's predictions for different prior distributions. We will further show how the SOM accounts for the difference between absolute and relative adjectives observed by Kennedy (2007).

3 The absolute-relative distinction

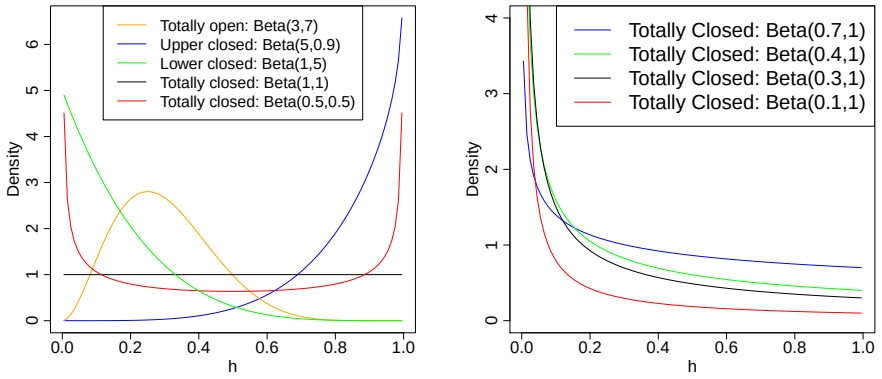
According to degree semantics (Kennedy & McNally 2005; Kennedy 2007), the crucial difference between absolute and relative adjectives is whether their underlying scale structures have accessible endpoints (i.e., lower or upper bounds). Previous accounts (e.g., Franke 2012; Lassiter & Goodman 2013) interpret this difference as a constraint on the type of probability distribution of the degrees. More specifically, probability distributions on open and closed scales differ in whether there can be significant probability mass on the endpoint. For instance, a relative adjective such as *tall* corresponds to a scale that has no maximal element because degrees of height are in principle unbounded (and this would be reflected via world knowledge in any prior even for a contextually fixed comparison class) and thus the probability must asymptotically fall to 0. In contrast, an absolute adjective such as *open* is associated with a scale that has a maximal element, and the occurrence probability of maximally open objects is usually non-negligible.

We adopt this view and apply the SOM to various distributions within the beta distribution family, which not only has a wide range of distributions that help us explore the exact boundary between absolute and relative adjectives, but also has nice closure properties that facilitate analytic derivations. A beta distribution is defined on $[0,1]$ and has two positive shape parameters α, β. Its densitiy function is defined as follows.

$$\phi(d;\alpha,\beta) = Kd^{\alpha-1}(1-d)^{\beta-1} \tag{10}$$

4 For $c = 2$ the curve is shifted far to the right, so its shape is not as obvious. It will be clear later that this value is too large, but we use it mainly to illustrate the realm of possible predictions.

5 Given the prevalence of gradable adjectives, we do not expect it to be very costly to utter them. In particular, the cost should not be the main factor that drives model prediction. Thus we take $c = 0$ as an approximation of the relatively small cost of the positive form in the next section when we discuss the absolute-relative distinction.



(a) Beta(α, β) (b) Beta($\alpha, 1$)

Figure 2 Correspondence between beta distributions and scale structures.

$K = 1/B(\alpha, \beta)$ is a normalization constant. Indeed, there is a tight correspondence between parameters of the beta distribution and scale types (Fig. 2). If $\alpha, \beta > 1$, both endpoints have zero probability mass, which corresponds to open scales. If $\alpha > 1, \beta \leq 1$, the lower endpoint has zero probability mass and the upper endpoint has nonzero probability mass, which corresponds to upper closed scales. Similarly, $\alpha \leq 1, \beta > 1$ corresponds to lower closed scales. Finally, if $\alpha, \beta \leq 1$, both endpoints have nonzero probability mass, which corresponds to totally closed scales.

The remainder of this section is dedicated to demonstrating that the SOM predicts the following correspondence between endpoint probability mass and the optimal threshold:[6]

(11) (i) If there is a sufficient amount of probability mass at the upper endpoint, then the maximal threshold is always optimal and so we obtain a maximum-standard reading, as in *The line is straight*, which is true, strictly speaking, only when the line is completely straight.

(ii) If (i) is not the case and the probability mass at the lower endpoint is sufficiently larger than elsewhere, then the non-minimal threshold[7] is optimal and so we obtain a minimum-standard reading, as in *The line is*

6 For now we always assume $c = 0$, and in the end we will show that this assumption is not crucial.

7 By non-minimal threshold we mean the one that corresponds to the non-minimal reading. In discrete cases this simply means the second minimal degree. In continuous scales, it means the utility function is decreasing on non-minimal degrees.

bent, which is true, strictly speaking, as soon as the line is not perfectly straight.

(iii) Otherwise the optimal threshold is highly sensitive to $\phi(h)$ and we obtain a relative reading.

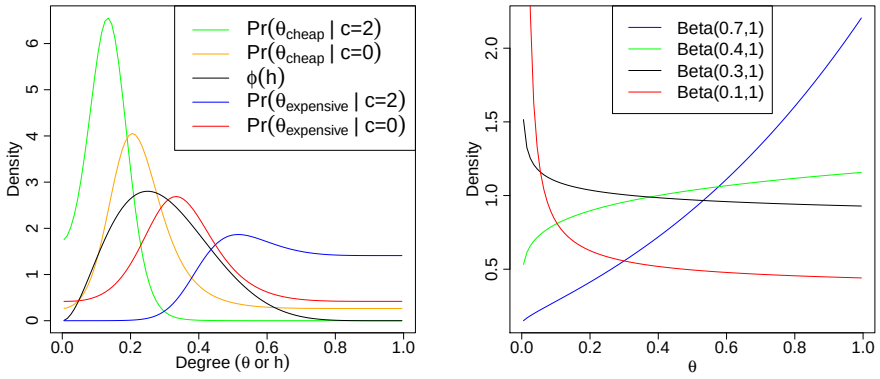
Let us first briefly explain how this correspondence correctly predicts the difference between absolute and relative adjectives observed by Kennedy (2007). Recall that when we formulate the model in the previous section, we assume that the prior distribution $\phi(h)$ is the speaker's exact knowledge about the comparison class. As already pointed out there, in reality, besides the fact that comparison classes are often implicit, the speaker almost always has uncertainty about the exact distribution $\phi(h)$ due to imperfect world knowledge. Typically, speakers only know the type of probability distribution for each adjective; they do not know the exact distribution. Even if the speaker directly observes the set of entities that forms the comparison class, due to noisy perception and the unavoidable blend from past experiences, his knowledge of $\phi(h)$ will be imprecise. Nevertheless, in the case of absolute adjectives, the speaker does not need to know the exact $\phi(h)$ in order to know where the optimal threshold is. According to (i) and (ii), as long as there is sufficient probability mass at either endpoint, the optimal threshold will be there.[8] This stability of optimal threshold explains why absolute adjectives are semantically not vague. In contrast, for open-scale adjectives, the optimal threshold is highly sensitive to $\phi(h)$ according to (iii) and the speaker cannot be sure where the optimal threshold is. Thus, the vagueness of relative adjectives is the result of such sensitivity of the optimal threshold when there is uncertainty about the exact prior.

In the following we try to show that (11) holds through representative examples in the beta distribution family. We start from the relatively simple part, (iii). For open scales, the corresponding beta distribution has parameters $\alpha, \beta > 1$. For example, Beta(3,7) is a distribution on an open scale that roughly corresponds to *cheap* and *expensive*.[9] Fig. 3(a) shows the SOM's prediction of $\Pr(\theta)$. Indeed, we can see that as the prior probability mass shifts to the left in Fig. 3(a), the optimal threshold also shifts to the left. (Compare the red and blue lines with those in Fig. 1(a).) We can also see that the optimal threshold is sensitive to the cost. Higher cost will drive the optimal threshold to a greater degree (recall that the ordering for *cheap* is reversed).

Let's look at closed scales next. We will first focus on cases where $\beta = 1$,

8 The speaker actually chooses the threshold sub-optimally via soft-max, reflecting the loose use of language, but if they are forced to, they can confirm that semantically the threshold is not vague because it is always at either endpoint.

9 Technically, we use Beta(3,7) for *degrees of expense* and Beta(7,3) for *degrees of cheapness*, as they have inverse orderings on the degrees, and put the predictions of both models in the same plot. The predictions do not change much for a prior distribution that has small non-zero probability mass at 0 used by Lassiter & Goodman (2013).



(a) $\Pr(\theta)$ predicted by the SOM

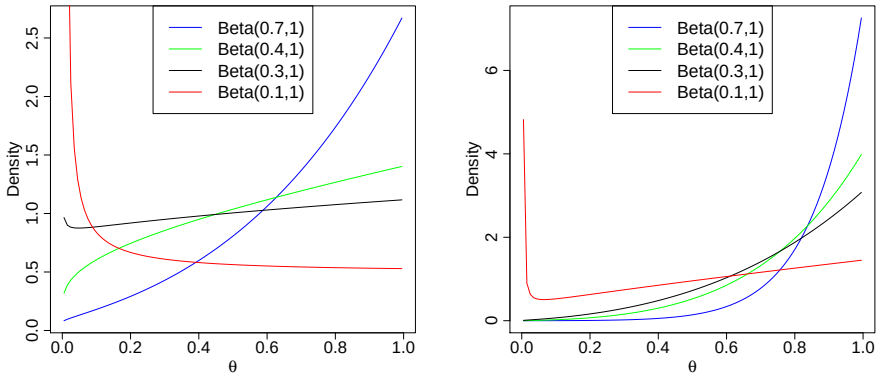
(b) $\Pr(\theta)$ predicted by the SOM, with $c = 0$

Figure 3 Predictions by the SOM for Beta(α,β), with $\lambda = 4$.

as other cases will be straightforward thereafter. From (10) it can be proved that Beta$(\alpha, 1)$ has density function $\phi(h; \alpha, 1) = \alpha h^{\alpha-1}$, and specifically the probability mass at the end point $h = 1$ is α. Fig. 3(b) shows the SOM's prediction of $\Pr(\theta)$. We can see that when α is high (0.4 or 0.7), $\Pr(\theta)$ is always increasing, which means the upper endpoint is always the optimal standard. This corresponds to (i). Meanwhile, when α is low (0.3 or 0.1), $\Pr(\theta)$ is always decreasing on (0,1], which means that a non-minimal standard is always optimal.[10] This corresponds to (ii). In fact, it can be shown that $\alpha = \frac{1}{3}$ is when a "phase transition" takes place, i.e., $\Pr(\theta)$ is increasing when $\alpha > \frac{1}{3}$, uniform when $\alpha = \frac{1}{3}$ and decreasing when $\alpha < \frac{1}{3}$. Hence we know that optimal thresholds for absolute adjectives are stable under a wide range of priors, which means slight uncertainty in $\phi(h)$ will not affect the speaker's knowledge about the optimal standard.

Fig. 4 further shows the robustness of the SOM's predictions with respect to costs. We already know that a higher cost will drive θ to the right (greater degree), which means higher costs will not affect maximal readings. Hence we only need to focus on thresholds for non-minimal readings such as Beta(0.3,1) and Beta(0.1,1). We can see that when the cost is relatively low for the prior, the prediction is almost unaffected, as shown in Fig. 4(a) for Beta(0.1,1). Meanwhile, when the cost becomes

10 The plot does not show the result of $\theta = 0$. Note that $\theta = 0$ means the positive form is always true, which is effectively the same as staying silent all the time. Thus $\theta = 0$ has very low utility and can never be optimal if $c > 0$.



(a) $c = 0.2$ (b) $c = 2$

Figure 4 Predictions by the SOM for Beta(α,1), with $\lambda = 4$.

relatively high, the upper endpoint also becomes a local optimum, as reflected by the v-shaped curves in Fig. 4(a) for Beta(0.3,1) and in Fig. 4(b) for Beta(0.1,1).[11] Nevertheless, the maximal threshold is only a local optimum. In fact, it can be proved that for $\alpha < \frac{1}{3}$, $U(\theta)$ always goes to infinity when θ approaches 0 (remember that $\theta = 0$ is excluded), regardless of the cost, so the non-minimal threshold is always globally optimal. Finally, when $\beta < 1$, the density $\phi(h)$ goes to infinity at the upper endpoint, which means $\Pr(\theta)$ will be driven even faster to the upper endpoint. So, again we obtain the maximal threshold as predicted by (i) and this prediction is also robust with respect to the cost parameter.

In sum, we have shown that the SOM correctly predicts the difference between relative and absolute adjectives. We interpreted vagueness as the stability of the optimal threshold under uncertainty about the exact prior distribution of degrees in the comparison class. We also illustrated how degree scales, by constraining the type of priors, influence speaker's knowledge about the optimal threshold. Open scales by definition do not have probability mass on endpoints, thus optimal thresholds are sensitive to the exact prior. As a result, the speaker cannot be sure about where they are when she is uncertain about the exact prior. This accounts for the vagueness of relative adjectives. In contrast, closed scales can constrain the priors such that there is sufficient probability at either end point, which is enough for the speaker to be certain about the optimal threshold, even if she is not sure about the exact prior. This

11 In fact, it is also true for Beta(0.3,1) in Fig. 4(b), but the turning point is too close to 0 to be observable.

explains why absolute adjectives are semantically not vague.

It should be noted here that prior distributions are not only constrained by scale types, but also by general world knowledge and immediate contextual information. Scale types might provide default priors (abstracted away from general world knowledge) that can be seen as part of the lexical properties of gradable adjectives. Nevertheless, when the speaker has more specific world knowledge about the comparison class or even directly observes it from the immediate context, she will know more about the degree distribution and adjust the prior accordingly. Different gradable adjectives can have constraints with different strengths. Open scale adjectives have no endpoints in principle so the prior distribution is more contextually variable. Closed scale adjectives, on the other hand, will usually be associated with the knowledge that there is in principle a lower or an upper bound, but still the corresponding priors that a speaker entertains at any given moment can still be influenced by world knowledge (*full* for glasses of wine) or direct contextual information (see experimental evidence by Solt & Gotzner 2012; Qing & Franke 2014).

4 Comparison to the Rational Speech-Act (RSA) model

In previous sections we introduced our model and illustrated how it predicts the relative-absolute distinction. In this section we will compare it with the closely related *Rational Speech-Act* (RSA) model proposed by Lassiter & Goodman (2013). The RSA model provides a probabilistic account of the semantics of gradable adjectives, based on a series of work in Bayesian pragmatics (e.g., Frank & Goodman 2012; Goodman & Stuhlmüller 2013). The RSA model improves on previous probabilistic approaches in that it provides precise quantitative predictions about (the listener's beliefs about) the probability distribution of the threshold.

Our speaker-oriented model shares similar assumptions and formalisms with the RSA model, but is also different from it in both conceptual and technical aspects. Below, we will explain the RSA model[12] in detail and focus on the main differences. The RSA model is based on the same scenario of descriptive language use and has the same literal listener component, i.e., (3) and (4).

The difference mainly resides in the speaker component. In the RSA model, the speaker is assumed to have an exact threshold θ saturated. In addition, when the positive form is applicable, i.e., $h_0 \geq \theta$, the speaker needs to choose between the positive form and staying silent, according to the informativity and cost of either

12 To make comparison easier, the version of the RSA model presented here is slightly different from the original formulation in that we do not consider antonyms. This allows us to do analytic derivations that simplify the computation of the posteriors. The obtained predictions are not crucially different from the original version and we will further discuss this modification in the end.

choice. Technically, the speaker applies the soft-max function according to the utilities defined below.

$$\sigma(u \mid h_0, \theta) \propto \exp(\lambda U(u, h_0, \theta)) = \exp(\lambda(\text{Info}(u, h_0, \theta) - \text{Cost(u)})) \tag{12}$$

Informativity is measured as the negative *surprisal* of the literal listener's updated belief about h_0.

$$\text{Info}(u, h_0, \theta) = -\log(1/\phi(h_0 \mid u, \theta)) = \log \phi(h_0 \mid u, \theta) \tag{13}$$

Assuming the costs for silence and *tall* are 0 and c, respectively, from (3), (4), (12), and (13), we can write down the probability of uttering *tall* explicitly.

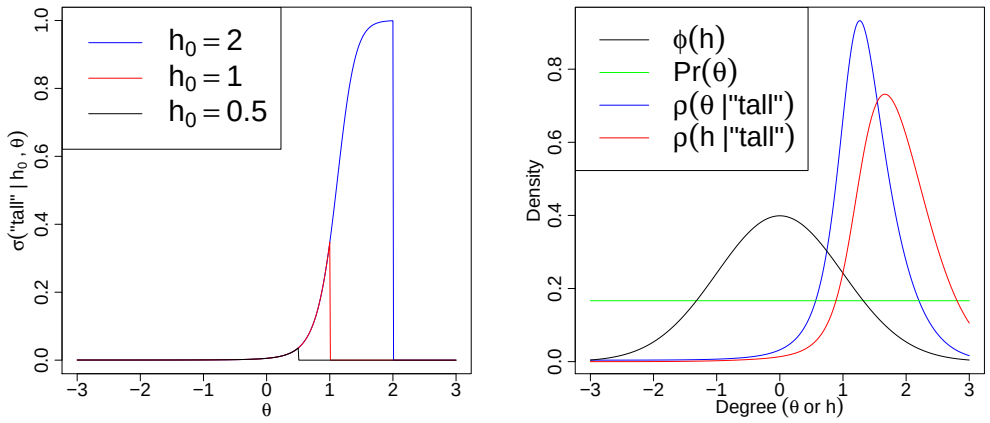
$$\begin{aligned} \sigma(u_1 \mid h_0, \theta) &= \frac{\exp(\lambda U(u_1, h_0, \theta))}{\exp(\lambda U(u_1, h_0, \theta)) + \exp(\lambda U(u_0, h_0, \theta))} \\ &= \frac{\exp(\lambda(\log \frac{\phi(h_0)}{\int_\theta^\infty \phi(h)\,\mathrm{d}h} - c))}{\exp(\lambda(\log \frac{\phi(h_0)}{\int_\theta^\infty \phi(h)\,\mathrm{d}h} - c)) + \exp(\lambda(\log \phi(h_0) - 0))} \\ &= \frac{1}{1 + e^{\lambda c} \cdot (\int_\theta^\infty \phi(h)\,\mathrm{d}h)^\lambda} \text{ if } h_0 \geq \theta, \text{ otherwise } 0 \quad . \end{aligned} \tag{14}$$

The predictions of the rule are shown in Fig. 5(a). We can see that $\sigma(u_1 \mid h_0, \theta)$ increases as θ increases, as long as $\theta \leq h_0$ holds (so that the positive form is semantically true). Another important feature is that once θ is fixed, $\sigma(u_1 \mid h_0, \theta)$ is the same for all $h_0 \geq \theta$. This shows that the RSA model does not have a convincing say on actual production of positive forms. First, it assumes that the speaker knows the exact threshold for the positive form, which is arguably not the case in reality. Second, even if the speaker does know (perhaps implicitly) the exact threshold θ, the RSA model would predict that the speaker uses the positive form with equal likelihood no matter what the actual degree h_0 is (as long as $h_0 \geq \theta$). This does not capture the actual production of positive forms, i.e., the greater the degree, the more likely that the positive form will be used (e.g., Schmidt, Goodman, Barner & Tenenbaum 2009; Solt & Gotzner 2012; Qing & Franke 2014).

In contrast, the SOM gives precise predictions about the speaker's uncertainty of the threshold and directly predicts the production probability of the positive form for each degree, which captures our intuition and can be further empirically tested.

Next we consider the derivation of the probability distribution of the threshold, which according to the RSA model is on the actual/pragmatic listener's level. Upon hearing the positive form, the pragmatic listener tries to make a joint inference about the threshold as well as the true value by using Bayes' rule.

$$\rho(h, \theta \mid u_1) \propto \phi(h) \cdot \Pr(\theta) \cdot \sigma(u_1 \mid h, \theta) = \frac{\phi(h) \cdot \Pr(\theta)}{1 + e^{\lambda c}(\int_\theta^\infty \phi(h)\,\mathrm{d}h)^\lambda} \tag{15}$$



(a) $\sigma(u_1 \mid h_0, \theta)$ (b) Posteriors of θ and h

Figure 5 RSA predictions for Gaussian distribution N(0,1), with $\lambda = 4, c = 2$.

$\Pr(\theta)$ is the prior linguistic knowledge about the threshold θ.

In order to obtain the posterior distribution of θ, we can marginalize over h.

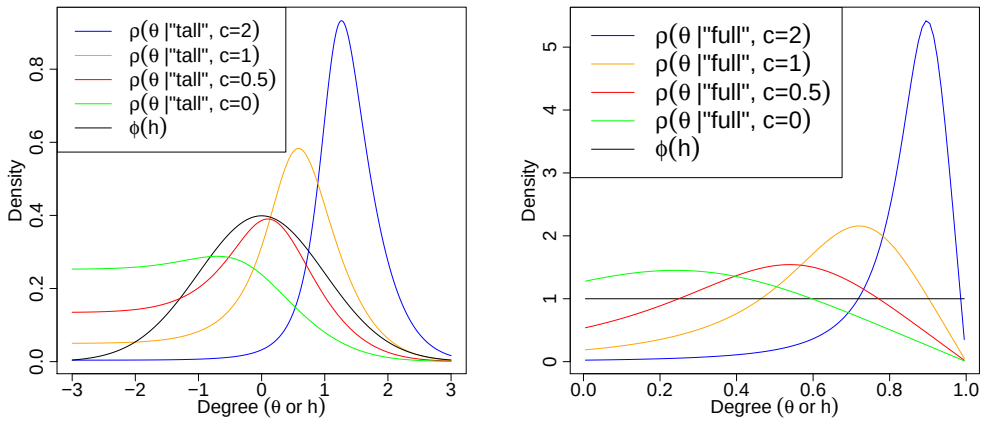
$$\rho(\theta \mid u_1) \propto \int_{-\infty}^{\infty} \phi(h) \cdot \Pr(\theta) \cdot \sigma(u_1 \mid h, \theta)\, \mathrm{d}h = \frac{\Pr(\theta) \cdot \int_{\theta}^{\infty} \phi(h)\, \mathrm{d}h}{1 + e^{\lambda c} \cdot (\int_{\theta}^{\infty} \phi(h)\, \mathrm{d}h)^{\lambda}} \tag{16}$$

Similarly, we can derive the posterior distribution of h.

$$\begin{aligned} \rho(h \mid u_1) &\propto \int_{-\infty}^{\infty} \phi(h) \Pr(\theta) \sigma(u_1 \mid h, \theta)\, \mathrm{d}\theta \\ &= \phi(h) \int_{-\infty}^{h} \frac{\Pr(\theta)}{1 + e^{\lambda c} \cdot (\int_{\theta}^{\infty} \phi(h)\, \mathrm{d}h)^{\lambda}}\, \mathrm{d}\theta \end{aligned} \tag{17}$$

If we take the prior of height $\phi(h)$ to be the normal distribution N(0,1) and use uniform threshold prior $\Pr(\theta)$, we can see from Fig. 5(b) that the RSA model's prediction of the threshold and the height posterior with $\lambda = 4$, $c = 2$. While the RSA model's posteriors of threshold and degree look very similar to the SOM's prediction in Fig. 1(a), we will argue below that the RSA model has certain features that can be problematic, especially for its account of the absolute-relative distinction.

Sensitivity to the cost parameter Fig. 6 shows the predictions of the RSA model for Gaussian and uniform distributions. We can see that when the cost $c = 2$, the RSA



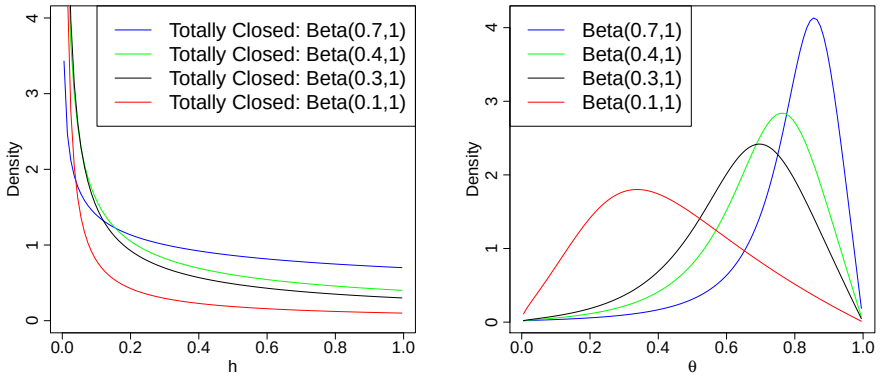
(a) "Tall": Normal prior N(0,1)

(b) "Full": Uniform prior Beta(1,1)

Figure 6 RSA predictions with $\lambda = 4$ and various costs.

model gives reasonable threshold posteriors (blue lines); Lassiter & Goodman (2013) argue that this captures the relative-absolute distinction. When c is small, however, threshold posteriors both shift to the left. In particular the threshold posterior of the absolute adjective (Fig. 6(b)) soon moves away from the endpoint. This means that positive forms have to require enough production effort in order for the RSA model to predict correctly, especially for its account of the absolute adjectives. Whether this assumption is warranted is, of course, an empirical question and depends on a precise theory about production effort. Nonetheless, the above feature also implies that if it took only little effort to utter positive forms, even absolute adjectives would receive weak meanings. This seems rather counter-intuitive. In contrast, the SOM does not crucially rely on the cost parameter for its predictions. In particular, the absolute-relative distinction does not rely on specific values of production effort.

Sensitivity to the degree prior Fig. 7 shows the RSA model's predictions for various closed scale priors. We can see that the maximum of the threshold posterior always shifts as the degree prior changes. This calls for an explanation of the contextual invariance of absolute adjectives. Specifically, given that the maximum of its threshold posterior is sensitive to the prior and is never actually the maximal or minimal degree, what in the RSA model's prediction corresponds to the stable maximal/minimal reading of absolute adjectives?



(a) Beta(α, 1) (b) RSA model's prediction

Figure 7 Predictions by the RSA for Beta(α, 1), with $\lambda = 4, c = 2$.

Here we want to emphasize the difference between vagueness and loose talk. Indeed, people often use absolute adjectives in ways that do not conform to maximal or minimal readings, e.g., people would use *full* for a glass of water even when there is still some room left. It may seem that the absolute-relative distinction is blurry after all. While this may be true for actual language use, there remain rather distinct semantic intuitions about absolute and relative adjectives, i.e., for absolute adjectives we can give a precise threshold with confidence if forced to, which is impossible for relative adjectives. This semantic aspect of vagueness also needs explanation, even if the probabilistic use of positive forms has been accounted for. It is unclear how the RSA model will address this issue.

In contrast, the SOM predicts stable optimal endpoint thresholds for absolute adjectives under reasonably small changes in priors. Only when the prior significantly deviates from typical cases will the optimal threshold change. Essentially, the SOM predicts two sources of the uncertainty about positive forms. First, the threshold is usually not optimized perfectly due to bounded rationality. This predicts the probabilistic use of both types of adjectives in reality. Second, vagueness is also the result of uncertainty about the prior degree distribution from imperfect knowledge about the comparison class. In this respect, absolute and relative adjectives exhibit different properties in terms of the stability of the optimal threshold. This accounts for their difference in context-variability.

Counter-intuitive metalinguistic effects Finally, the RSA model predicts rather counter-intuitive metalinguistic effects in a scenario that only involves purely descriptive language use. Suppose people are talking about John and somebody says *John is tall*. You have never met John, but you have some world knowledge about the distribution of adult male height, $\phi(h)$. According to the RSA model, prior to the utterance you know nothing about what *tall* means for an adult male (even with the world knowledge you have), but afterwards you both know what *tall* means and gain some information about John's height.

However, it seems more plausible to say that you are able to gain some information about John's height because you already know what *tall* roughly means for men even *before* the utterance, and you apply this knowledge afterwards to infer John's height.[13] Also, if this utterance is all you get and you receive no extra information about John's height, then intuitively it seems that after the utterance you know nothing more about what *tall* means beyond what you have anticipated from the prior world knowledge. The SOM adopts this perspective in its predictions.

Semantic vs pragmatic optimality So far we have seen how the SOM's predictions differ from the RSA model in several respects. Here we will further discuss the conceptual relation between the two models. We will argue that the SOM, which stems from previous evolutionary approaches, emphasizes considerations of *semantic optimality*, while the RSA model focuses on *pragmatic optimality*.

To be precise, we distinguish between three senses of pragmatics: (1) language use in general, (2) how contexts affect meaning, and (3) how language can be used beyond the conventional/literal meaning. The following discussion is in the sense of (3), as both models deal with (1) and (2).

The SOM is a semantic approach with respect to (3) in that it treats the probability distribution of the threshold as conventional semantic knowledge within a linguistic community. The probability of each threshold depends on the utility of the corresponding semantic system. Thus for each threshold, in order to evaluate its communicative success, we introduce speakers and listeners who use the positive form according to the semantics and measure the expected success in conveying the true degree. The optimization problem is for the whole community to select thresholds that form good semantic systems.

The RSA model, on the other hand, is pragmatic in that its speaker faces a choice between several alternatives. As a result, the use of one expression depends on not only its semantics, but also its relation to the alternatives. Such a model has been highly successful for typical pragmatic phenomena such as scalar implicature.

13 If such knowledge does not exist beforehand, then if you ask the question *Is John tall?* before anyone else uses *tall*, it will have an implausibly weak meaning.

We believe that the two approaches should be complementary rather than contradictory, and ideally they should be integrated to fully capture the complexity of meaning and use. The challenge is that it is often unclear which aspect weighs heaviest for any particular phenomenon we are interested in. For instance, we have only considered two possible utterances in the current setting for simplicity, and we have seen that the semantic approach of the SOM provides reasonable predictions. However, when we are to extend the model to allow for more alternatives, such as antonyms and compositional expressions, we can either hold on to a fully semantic approach, i.e., optimizing over all possible semantic conventions, or use the semantic approach to derive the base semantics of each alternative and put them into a pragmatic model to compete with each other. We leave more elaborate discussions for future work.

5 Conclusion

In this paper, we combined previous evolutionary and probabilistic approaches to the meaning of gradable adjectives and proposed a new speaker-oriented model. The contributions are the following. First, we addressed how the probability distribution of conventional thresholds is contextually determined. Second, we added a fully predictive speaker model and illustrated that it makes plausible predictions under various parameters. Third, we distinguished between vagueness and loose talk to address the difference between relative and absolute adjectives' interpretation. In particular, we argued that an important source of vagueness is the uncertainty about the exact degree distribution within the comparison class.

Although we have been critical to the RSA model, it is clear that our approach owes a lot to it. We believe that SOM and RSA focus on different aspects of positive forms of gradable adjectives and that more work is necessary to further integrate the two approaches. This becomes especially relevant when trying to test the predictions of probabilistic models of this kind experimentally (for a first attempt, see Qing & Franke 2014).

References

Frank, Michael C. & Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336(6084). 998. doi:10.1126/science.1218633.

Franke, Michael. 2012. On scales, salience & referential language use. In Maria Aloni, Floris Roelofsen & Katrin Schulz (eds.), *Amsterdam Colloquium 2011* Lecture Notes in Computer Science, 311–320. Springer.

Goodman, Noah D. & Andreas Stuhlmüller. 2013. Knowledge and implicature:

Modeling lanuage understanding as social cognition. *Topics in Cognitive Science* 5. 173–184.

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

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

Lassiter, Daniel. 2011. Vagueness as probabilistic linguistic knowledge. In Rick Nouwen, Robert van Rooij, Uli Sauerland & Hans-Christian Schmitz (eds.), *Vagueness in Communication*, 127–150. Springer.

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

Luce, Duncan R. 1959. *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

Potts, Christopher. 2008. Interpretive Economy, Schelling points, and evolutionary stability. Manuscript, UMass Amherst.

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

Schmidt, Lauren A., Noah D. Goodman, David Barner & Joshua B. Tenenbaum. 2009. How tall is tall? compositionality, statistics, and gradable adjectives. In Niels Taatgen, Hedderik van Rijn, Lambert Schomaker & John Nerbonne (eds.), *Cognitive Science Society (CogSci 2009)*, vol. 31, 3151–3156. Curran Associates, Inc.

Solt, Stephanie & Nicole Gotzner. 2012. Experimenting with degree. In Anca Chereches (ed.), *Semantics and Linguistic Theory (SALT) 22*, 166–187. CLC Publications.

Sutton, Richard S. & Andrew G. Barto. 1998. *Reinforcement learning: An introduction*. Cambridge: MIT Press.

Ciyang Qing
Institute for Logic, Language and Computation
Universiteit van Amsterdam
P.O. Box 94242
1090 GE Amsterdam
The Netherlands
qciyang@gmail.com

Michael Franke
Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen
Wilhelmstraße 19
72072 Tübingen
Germany
mchfranke@gmail.com