Rationalizing evaluativity[1]

Dylan BUMFORD and Jessica RETT — *University of California, Los Angeles*

Abstract. Many degree constructions are *evaluative* in the sense that they require a measure of some sort to exceed a contextually-determined norm. We assume that this inference is always an implicature (Rett, 2015b), and we develop a game-theoretic pragmatic treatment to explain how and when it arises. Our analysis is couched in a Rational Speech Act (RSA) model of communication, building on the theory of vagueness-resolution proposed in Lassiter and Goodman 2014. To extend the Lassiter and Goodman model to a wider range of degree constructions, we do two things: first, we incorporate insights from Barker (2002) about the role of comparison-class uncertainty in the interpretation of gradable predicates, and second, we adapt an independent RSA model of Manner implicature (Bergen et al., 2016) to capture the effects of linguistic markedness in degree constructions (Rett, 2015b). Combining these pieces provides the first gradient model of evaluativity inferences, and we conclude by discussing some of its novel predictions.

Keywords: degrees; evaluativity; equatives; comparatives; manner implicature; vagueness; context-sensitivity; rational speech act theory; lexical uncertainty

1. Introduction

An *evaluative* expression is one that implies that the degree to which an entity exhibits a certain property exceeds some contextual standard. The distribution of these inferences across constructions has proven challenging to account for compositionally, As a result, recent analyses have cast the phenomenon as a species of implicature (Rett, 2015b), and given the quantitative nature of the inferences, we feel it is ripe for modeling in a game-theoretic pragmatics.

We start out by demonstrating that no existing game-theoretic treatment of gradable language can account for the full profile of data. We then offer a novel proposal that does. Our analysis builds on the Rational Speech Act (RSA) model in Lassiter and Goodman 2014 in two ways. First, we incorporate a treatment of markedness from the standpoint of *lexical uncertainty* (Bergen et al., 2016), and second we argue that the semantics of degree constructions should be framed in terms of two parameters of variation, one tracking the measure of an argument and the other tracking the distribution of the argument's comparison class (Barker, 2002).

Our analysis provides a strict ranking of a variety of degree constructions according to the degree to which they exhibit evaluativity. While this gradient characterization is at odds with the traditional categorical accounts of evaluativity (Bierwisch, 1989), these predictions accord with recent experimental work suggesting that evaluativity manifests differently in different contexts (Brasoveanu and Rett, 2018), and hold the promise of informing future work on the interaction of vagueness and context-sensitivity.

[1]We are extremely grateful to Jeremy Kuhn, Dan Lassiter, and Louise McNally for their comments and engagement in our online Q&A session. This project has also benefitted from early discussions with Heather Burnett, Michael Franke, and Cailin O'Connor.

 In: P. G. Grosz, L. Martí, H. Pearson, Y. Sudo, and S. Zobel (eds.) *Proceedings of Sinn und Bedeutung 25*, pp. 187–204. University College London and Queen Mary University of London.

1.1. An introduction to evaluativity

We will say that a construction is *evaluative* if it implies that by some measure (height, weight, etc.) an entity is above the norm for that entity's comparison class. We will often refer to the (flexible, contextually-determined) comparison class norm as the *standard* to which entities are compared. For instance, the expressions in (1) are evaluative, since they imply that Jane's height is unusually high/low compared to people relevantly like Jane in the context of utterance.

(1) a. Jane is tall.
b. Jane is short.

We assume that the difference between a non-gradable adjective like *prostrate* and a gradable adjective like *tall* is a matter of valence; the former denote individual predicates, while the latter denote relations between degrees and individuals. The sentences in (1) are examples of the *positive construction*, in which the degree argument of a gradable predicate (here, *tall* and *short*, respectively) is neither bound — e.g., by a degree quantifier like *-er* — nor explicitly valued — e.g., by a measure phrase like *six feet*. Instances of the positive constructions are always evaluative, regardless of whether the adjective is intuitively "positive", like *tall*, or "negative", like *short*.

A reliable test for evaluativity in declaratives is entailment to the negated antonymic counterpart (Bierwisch, 1989), as demonstrated in (2).

(2) a. Jane is tall. $\Rightarrow$ Jane is not short.
b. Jane is short. $\Rightarrow$ Jane is not tall.

Historically (Kamp, 1975; Cresswell, 1976), the semantic problem associated with evaluativity is the following: if morphologically complex adjectival forms like the comparative are compositionally derived from a positive construction, and the positive construction is evaluative, then the complex forms should all be evaluative too, but they aren't. The historic solution has been the postulation of a null operator, POS, which contributes evaluativity only in the absence of overt degree morphology binding or valuing the open degree argument in the positive construction.

(3) a. $[\![\text{tall}]\!] = \lambda d \lambda x \lambda w.\, \mathsf{ht}_w(x) \geq d$
b. $[\![\text{Jane is POS tall}]\!] = \lambda w.\, \mathsf{ht}_w(\mathsf{j}) \geq \mathbf{s}$
c. $[\![\text{Jane is taller than Keisha}]\!] = \lambda w.\, \{d \mid \mathsf{ht}_w(\mathsf{j}) \geq d\} \supset \{d \mid \mathsf{ht}_w(\mathsf{k}) \geq d\}$

The POS account thereby predicts, by design, that evaluativity (the requirement that the relevant measure exceed the contextual standard **s**) is in complementary distribution with overt degree morphology that saturates the adjective's degree argument (in the case of (3), the comparative morpheme). But this is false (Bierwisch, 1989). We will refer to this prediction as the 'Open Argument Assumption'.

As argued in Rett 2008, 2015b, adjectival constructions fall into three distinct categories with respect to evaluativity. There are those constructions that are never evaluative, regardless of which antonym they are formed with (4); there are those that are always evaluative, regardless of which antonym they are formed with (5); and there are those that are evaluative or not,

depending on the antonym they're formed with (6).[2] The comparative in (4b), for instance, is non-evaluative because it fails to entail its antonymic counterpart *Jane is not tall*. And the equative in (6b) is evaluative because it does entail its antonymic counterpart *Jane is as tall as Keisha*.

(4) **non-evaluativity**

a. Jane is 5 ft tall. *measure phrase construction*

b. Jane is taller/shorter than Keisha. *comparative*

(5) **antonym-insensitive evaluativity**

a. Jane is tall/short. *positive construction*

b. Is Jane tall/short? *polar positive question*

(6) **antonym-sensitive evaluativity**

a. How short is Jane? *degree question*

b. Jane is as short as Keisha. *equative*

c. Jane is that short too. *degree demonstrative*

Some constructions conform to the Open Argument Assumption: the sentences in (4) contain degree morphology but lack evaluativity, and the sentences in (5) lack degree morphology but are evaluative. But the constructions in (6) do not conform to the Open Argument Assumption, as they are evaluative despite containing overt degree morphology.

1.2. Extant implicature-based accounts of evaluativity

1.2.1. Rett (2015b)

Rett (2015b) argues that evaluativity, in all of its instantiations, arises as the result of a conversational implicature: Quantity in the case of antonym-insensitive constructions, and Manner in the case of the antonym-sensitive constructions. In order to associate antonym-sensitive evaluativity with Manner implicatures, Rett (2015b) reviews the large body of evidence that negative antonyms are marked relative to their positive-antonym counterparts (Lehrer, 1985).[3] Negative antonyms are often overtly morphologically marked relative to their positive counterparts (cf. *(im)possible*); when they aren't, evidence for their markedness comes from their relatively restricted cross-linguistic distribution (some languages lack negative antonyms); morphosyntactic distribution (*6 ft tall/*short*); and semantic distribution (*half as tall* vs. #*half as short*).

However, Rett (2015b) does not provide a formal account of how these implicatures arise (and how their embeddability can be modeled). It is this account that we aim to operationalize. And while there are existing treatments of antonym-insensitive evaluativity qua (something like) a Quantity implicature (Lassiter and Goodman, 2014) and of Manner implicatures writ large (Bergen et al., 2016) in formal pragmatic frameworks, we show that neither of these accounts can be generalized to extend to the full paradigm of evaluativity illustrated in (4)–(6).

[2]Crucially, this is an evaluative typology for the subclass of of gradable adjectives known as *relative*. Other subclasses of gradable adjectives — extreme, total, and partial — display distinct behavior with respect to evaluativity that is predictable from their lexical semantics (Kennedy and McNally, 2005; Rett, 2008).

[3]Rett (2015b) also extends the Manner predictions to other contrasts in markedness, like the difference between analytic (*more tall*) and synthetic (*taller*) comparatives.

1.2.2. Lassiter and Goodman (2014)

Lassiter and Goodman (2014) provide a formal account of how evaluativity comes to be associated with the positive construction; they argue that the evaluativity of positive constructions can be derived from pragmatic reasoning effects precisely because positive constructions contain an open degree argument.

The account consists of a Rational Speech Act model (Frank and Goodman, 2012). All Rational Speech Act models are built around two ideas. First, speakers choose which sentences to utter based on how costly those utterances are and how likely a listener hearing that utterance would be to deduce what situation the speaker is trying to describe. That is, all else equal, speakers avoid laborious messages, and gravitate towards messages that precisely describe the situation at hand. Of course these pressures are often in conflict, so speakers are generally forced to consider the relative *utilities* of various sentences, essentially an information-to-cost ratio. Second, listeners interpret utterances based in part on their prior expectations about the world and in part on how likely a speaker would have been to choose that utterance while trying to describe a given situation. In other words, listeners perform standard Bayesian inference to update their beliefs about the world, given a model of how the facts condition speakers' choices.

This basic communication game may be complicated by any number of additional sources of uncertainty. The speaker may not have complete knowledge of the situation they are describing. The listener may not know what question the speaker takes themself to be answering, and so may not know how to partition a set of possible worlds into the relevant hypothesis space. The message itself may contain implicit content regarding, say, the domains over which expressions quantify, or the indices that govern the referents of pronouns.

In fact, much of the work in the RSA tradition models the listener as reasoning *jointly* over any such variables of interest. Often the predicted conclusions about how different utterance choices can influence Bayesian listeners' beliefs about these discourse-contextual parameters are as interesting as the conclusions they are predicted to draw regarding the state of the world. To this end, Lassiter and Goodman assume the same type of gradable adjectives as in (3), but use the model to value the unsaturated degree argument of the adjective.

(7) a. $[\![\text{tall}]\!] = \lambda d \lambda x \lambda w.\, \mathsf{ht}_w(x) \geq d$
b. $[\![\text{Jane is } \emptyset_d \text{ tall}]\!] = \lambda w.\, \mathsf{ht}_w(\text{j}) \geq d$

Because the degree argument θ_d is left open (unexpressed and unbound) in positive constructions, listeners are forced to estimate a value for it. And because that value determines what proposition the speaker has actually expressed, which in turn determines what worlds to put increased stock in in, Lassiter and Goodman propose that listeners faced with (7b) reason simultaneously about Jane's height and what threshold the speaker might have in mind to distinguish the tall from the not tall.

The reasoning turns on the assumption that the speaker has made a rational choice to utter the sentence. This means that the speaker must think the sentence is reasonably informative, i.e., worth the cost of saying it. This increases the likelihood that the speaker's threshold is relatively high, since, for instance, learning that Jane's height exceeds 6 ft is more informative than learning that it exceeds 5 ft. But at the same time, highly informative sentences are *a priori* unlikely, since, absent any other relevant knowledge about Jane, chances are good that she is

a normally-sized person. So the listener weighs these trade-offs, and eventually concludes two things: (i) that the speaker's threshold is above the known mean for people's heights, but not much above it; and (ii) that Jane's height is above the inferred threshold, but not much above it.

This model is a compelling and imminently reasonable demonstration of how a savvy listener might interpret a vague utterance with little contextual guidance. The predictions about estimated heights and semantic cutoffs are intuitive and empirically compelling. But as it stands, it falls short of our broader explanatory goals in a couple of ways.

First, it is fundamentally an account of *vagueness*, of how interlocutors decide where to draw lines in the unspoken sand. Nonetheless, there are several ways in which we might construe the predicted inferences as also possibly *evaluative*, that is, as inferences about the relationship between a particular measurement and a contextual standard. Most obviously, we can compare the statistics of the listener's prior and posterior estimates of the subject's height and the threshold for tallness. For instance, after hearing *Jane is tall*, we might ask: (i) how much taller does the listener expect Jane to be than they did before the heard the sentence; (ii) how much did the expected value of the threshold for tallness change; or (iii) what is the posterior expected difference between the subject's height and the tallness threshold, or perhaps how much has the expected difference changed?

All of these questions provide reasonable bases on which we might say the positive construction — with either antonym — is predicted to be evaluative. But they all depend on lingering posterior uncertainty about the subject's measure and the value of the gradable predicate's implicit degree argument. In contrast, after hearing an equative like *Widget A weighs 10 lbs and Widget B is just as light*, there is no question of what Widget B's weight is. So there can be no comparison of estimates of the measure in question before and after the utterance, or to the extent that it would make sense to do so, the answers will be entirely determined by the specified measure (here, 10 lbs) and not the nature of the construction (an equative with a marked antonym). Yet the sentence is evaluative: it implies that 10 lbs is light for a widget. In other words, extending Lassiter and Goodman's (2014) account of vagueness-induced reasoning to the general phenomenon of evaluativity is tantamount to making the Open Argument Assumption.

The second concern for extending this account to evaluativity in general is the consideration of antonym-sensitive evaluativity. Given Lassiter and Goodman's semantics for gradable adjectives, the linguistic competition at the heart of the speaker's decision is between saying *Jane is tall* and saying nothing. For any non-zero setting of the tallness threshold θ_d, the former is more informative than the latter, but the latter is less costly than the former. What the speaker should do depends on how abnormal Jane's height is (that is, how likely a listener would be to guess her height correctly even if the speaker doesn't say anything). But even if the speaker also considers the possibility of saying *Jane is short*, this antonymic alternative will exert absolutely no pressure on the speaker's actions. The reason is that on every hypothesis about the value of θ_d, *Jane is tall* and *Jane is short* are contradictory; she is either above θ_d or below it. So if the speaker truthfully declares Jane to be tall, then there is no point in wondering why they didn't declare her to be short instead. The upshot of this is that there is effectively no direct competition between antonymic forms, and thus no obvious segue from this analysis to the antonym-sensitive inferences in (6).

Recall that Rett argues that antonym-sensitive evaluativity patterns emerge as a Manner implicature, driven by differences in the markedness of the two forms. To anticipate such an analysis, in the next section we turn to Bergen et al.'s (2016) recent proposal for deriving such implicatures from similar game-theoretic assumptions, and assess its prospects for augmenting the preceding account of vagueness.

1.2.3. Bergen et al. (2016)

Bergen et al. (2016) seek to demonstrate how the principle that marked messages describe marked scenarios can be derived from considerations of rationality. They set up the following computational experiment: a pragmatic listener is presented with a marked utterance that they know to be semantically synonymous with a less marked utterance. Based on this action, and the knowledge that speakers are unlikely to choose more costly messages when there are equivalent less costly messages available to them, the listener must guess what the world is like. The key to the listener's reasoning about this ostensibly irrational choice is that they allow for the possibility that *the speaker* does not consider the messages equivalent. That is, just as Lassiter and Goodman's pragmatic listener does not know exactly how the speaker means to use the word *tall* (over 66 inches?, over 67 inches? etc.), Bergen et al.'s pragmatic listener does not know exactly what proposition the speaker associates with each of the two putatively synonymous (marked and unmarked) messages. Following Potts et al. (2016), they dub this linguistic variability Lexical Uncertainty. The listener's task, then, is to perform joint inference over how the speaker interprets their own words and what situation they take themself to be describing.

Concretely, they provide the following model. Two utterances, one marked relative to the other, in principle both denote the same proposition: $\{w_1, w_2\}$. But in practice, the listener imagines that either utterance could be intended to refer to an arbitrary specific subcase of this general proposition. Perhaps they take the marked message to denote the general case $\{w_1, w_2\}$, but the unmarked message to denote just $\{w_1\}$. Or perhaps vice versa, etc. This leads to a range of hypotheses about what "lexicon" the speaker has in mind:

(8)

	$[\![\text{marked}]\!]$	$[\![\text{unmarked}]\!]$
$\mathscr{L}_0$	$\{w_1, w_2\}$	$\{w_1, w_2\}$
$\mathscr{L}_1$	$\{w_1\}$	$\{w_1, w_2\}$
$\mathscr{L}_2$	$\{w_1, w_2\}$	$\{w_1\}$
$\mathscr{L}_3$	$\{w_1\}$	$\{w_2\}$
$\vdots$	$\vdots$	$\vdots$

However, not all of these potential denotations have the same *a priori* probability. In particular, semantic considerations aside, w_1 is thought to be twice as likely as w_2, just in the ordinary sense that the events in w_1 are more typical than those of w_2. So before anything is said, the listener puts less stock in the proposition $\{w_2\}$ than in $\{w_1\}$.

Under these conditions, what should a Bayesian listener conclude after hearing the speaker choose to use the more costly message? Bergen et al. demonstrate that such a listener ought eventually to decide that the speaker takes the marked message to pick out $\{w_2\}$, the least likely proposition, and thus that the world we live in is, after all, w_2. These hypotheses provide the

best explanation for the speaker's behavior; if they intend to describe something improbable, and they associate the marked message specifically with this improbable state, then it is worth the cost of uttering it. On other hypotheses, the speaker's decision would harder to account for.

While this discussion is abstract, we ought to expect the same results from any specific competition between synonymous forms, so long as (i) one is more costly than the other, (ii) listeners entertain "lexical uncertainty" about their meanings, and (iii) there is an underlying difference in *a priori* likelihood among the worlds described by the utterances. So we ask, can the interpretive patterns of the antonym-sensitive inferences in (6) be cast in these terms? Imagine the sentences in (9) are both felicitous.

(9) You said if I looked I would find a 2-inch screw, ...
   a. and that's exactly how long this screw is
   b. and that's exactly how short this screw is

Uttering (9b) is plausibly more costly to the speaker than uttering (9a), in the sense that *short* is linguistically marked, relative to *long* (Lehrer, 1985; Rett, 2015b). And certainly it seems that at a baseline truth-conditional level, (9a) and (9b) denote the same proposition (Rett, 2015a); both are true iff the screw I found has a length of exactly 2 inches. However, there is no further uncertainty about what either of the utterances is intended to pick out, since they are both maximally informative with respect to the parameter under discussion. That is, if worlds are distinguished by the length of the screw, as in Lassiter and Goodman 2014, then the utterances in (9) both identify singleton propositions, and so simply do not have any refinements or stronger interpretations.

This leads naturally to the question of what happens when the partitioning of worlds into hypotheses is not determined solely by the length of the screw. In particular, what if there is some discernible variation among the worlds consistent with (9a)/(9b), namely, those in which the screw I find is 2 inches? Well, the Bergen et al. (2016) result is clear: people who hear (9b) will gravitate toward whichever of those worlds they consider least likely *a priori*. This could in principle be anything. The listener's worlds might include a situation where I find a 2-inch screw and an asteroid hits the earth tomorrow, and a situation where I find a 2-inch screw and an asteroid doesn't hit the earth tomorrow. In that case (9b) would lead the rational listener to start drinking. And in general this is appropriate; Manner implicatures are context-specific in many cases.

But remember we are trying to account for the very robust and systematic *evaluativity* inference that arises from (9b), namely that 2 inches is short for the relevant sort of screw. The only way this emerges consistently from the reasoning above is if the degree constructions themselves suggest a particular measure-related refinement of the space of worlds under consideration (Rett, 2015b). Imagine, for instance, that the relevant worlds consistent with (9b) were distinguished not by the presence or absence of asteroids, but by whether or not the screw is representative of its comparison class. Imagine furthermore that worlds where the screw is representative are more likely than worlds where it is uncharacteristically small. Then semantic uncertainty and markedness would lead to the empirically valid evaluative conclusion.

This is the essence of the model that we propose in the next section. We lay out a semantics for positive and equative constructions, as well as a canonical space of lexical refinements, that al-

lows listeners to reason jointly about individual measurements and the statistics of comparison classes. We then synthesize an operational notion of evaluativity from these joint considerations, and demonstrate that this model derives the target inferences.

2. The proposal: uncertainty about the comparison class

Our goal is to model how listeners reason about the relationship between individuals and their comparison classes, and how different kinds of utterances influence that reasoning. The listener's hypotheses then ought to be quotiented not just by the measurements of the subject, but also by the background distribution of the subject's competitors. Consequently, the worlds in our models will be distinguished along two dimensions: essentially, what is the gradable predicate's individual argument like, and what is everyone else like?

In this we are loosely following the analysis of vagueness in Barker 2002, who observes that there are two ways for sentences like (1a) to be informative. Obviously, hearing that Jane is tall can tell you something you didn't know about Jane's height. But in circumstances where you already know Jane's height, hearing (1a) might still tell you something about what counts as tall in the context of utterance. Barker formalizes this intuition by incorporating gradable thresholds — "what counts as *tall*" — into the worlds that make up listeners' belief states.

Rather than augmenting worlds with differentiating linguistic facts, as Barker does, we simply add facts about the measurements of other individuals. So we might, for instance, include a world where Jane is 65 inches tall while her comparison class is centered around 60 inches, and another world where Jane is 65 inches tall while her comparison class is centered around 70 inches. In our terminology, this variability models uncertainty about the standard for people like Jane. Over and above this uncertainty about what the world is like, listeners face "lexical" uncertainty about the threshold for tallness. This additional parameter governs which heights count as *tall* relative to what is standard, i.e., how atypical a person needs to be to count as *tall*.

2.1. Technical preliminaries

Like all Rational Speech Act models, we define a sequence of nested equations representing speakers' and listeners' behaviors as conditional probability distributions. The listener's interpretation is defined by a probability distribution over worlds conditioned by the utterance to be interpreted. That distribution is a function of the speaker's actions, which are in turn defined by a probability distribution over utterances conditioned by the world to be described. The particular equations we adopt are equivalent to those in Bergen et al. 2016: pg. 28.

(10)
$$\mathbb{L}_0(w \mid u, \mathscr{L}) \propto P(w) \cdot \mathscr{L}(u, w)$$
$$\mathbb{L}_1(w \mid u, \mathscr{L}) \propto P(w) \cdot \sum_{\mathscr{L}'} P(\mathscr{L}') \cdot \mathbb{S}_n(u \mid w, \mathscr{L}')$$
$$\mathbb{L}_n(w \mid u, \mathscr{L}) \propto P(w) \cdot \mathbb{S}_n(u \mid w, \mathscr{L}) \qquad \text{for } n > 1$$
$$\mathbb{S}_n(u \mid w, \mathscr{L}) \propto \exp(\alpha \cdot (\log \mathbb{L}_{n-1}(w \mid u, \mathscr{L}) - C(u))) \qquad \text{for } n \geq 1$$

The hyperparameters are α, which controls how aggressively the speaker prefers the winning utterance; C, which maps utterances to their "costs"; and the prior distributions over worlds $P(w)$ and interpretation functions $P(\mathscr{L})$. For all of the simulations we report, α is set to 4, and the cost function C is set so that the null message is free (cost 0) and the marked message is

twice as costly as the unmarked message (cost 2 and 1, respectively), again following Bergen et al. (2016).

These equations provide for the inspection of an arbitrarily "sophisticated" listener $\mathbb{L}_n$, one willing to traverse as many pragmatic round trips across the theory of mind as can be calculated. Many RSA models report only the distributions described by $\mathbb{L}_1$, but the associations between markedness and atypicality that Bergen et al. uncover only emerge robustly at higher levels of recursion, $\mathbb{L}_3$ or $\mathbb{L}_4$. In our results, we describe the $\mathbb{L}_1$ listener distribution as well as the stable distribution that ceases to change with further iterations. In every case, the stable distribution arises between $\mathbb{L}_3$ and $\mathbb{L}_6$, though the correct rank order in evaluativity among constructions is already present at $\mathbb{L}_1$.

Since our interest is in how people reason about the relationship between measures and standards, the worlds we will consider are effectively two-dimensional. Along one axis, worlds vary according to the measure of the relevant gradable predicate; along the other axis, worlds vary according to the distribution of measures among objects in the comparison class. We will use *tall* and *short* as paradigm gradable adjectives, so this means worlds are distinguished (i) by the subject's height and (ii) by the distribution of heights among relevantly similar people.

We make several simplifying assumptions in the name of tractability. First, heights are binned into 17 evenly spaced equivalence classes, numbered 1 through 17. Second, the distribution of comparison class heights is known by both interlocutors to be Gaussian with a standard deviation 2. Third, listeners only consider worlds where the subject's height is within 4 units (2 standard deviations) of the mean. Fourth, to avoid pooling effects created by the artificial floor and ceiling heights (1 and 17), the mean of the comparison class is assumed to lie between the 5th and 14th heights. Taken together, these assumptions guarantee that we have only finitely many worlds to consider, and that our probabilistic calculations are discrete.

The prior over worlds is then determined by the listener's beliefs about the comparison class mean. In the simulations we report, this distribution is uniform over the discrete interval [5, 14]. In other words, the listener is assumed to have no particular knowledge of where the mean lies within this range. This allows the most room to see how the listener responds to the gradable constructions, and quite naturally mirrors Lassiter and Goodman's (2014) assumption that the listener's prior over tallness thresholds is uniformly distributed.

In sum, our model represents speakers' simultaneous reasoning about individual measures and about comparison class means. What, then, should we say constitutes an evaluative belief in this model? We suggest that the operative statistic is the *expected deviation* of the relevant individual's height from the comparison class mean. That is, we will evaluate predicted listener responses to various utterances by comparing the expected values of the derived distribution $\mathsf{ht} - \mu$, where ht and μ are the random variables describing the axes of plots like Figure 0.

2.2. Simulations and predictions[4]

For each simulation, we identify a set of potential sentences that a speaker must choose between, and a set of possible interpretation functions mapping these sentences to propositions. In every case, the listener's prior over these "lexica" is uniform, as in Bergen et al. 2016.

[4]The code for all simulations reported here (and many others) are available online at `https://github.com/dylnb/eval-models/blob/master/src/sub25/eval-basic.ipynb`.

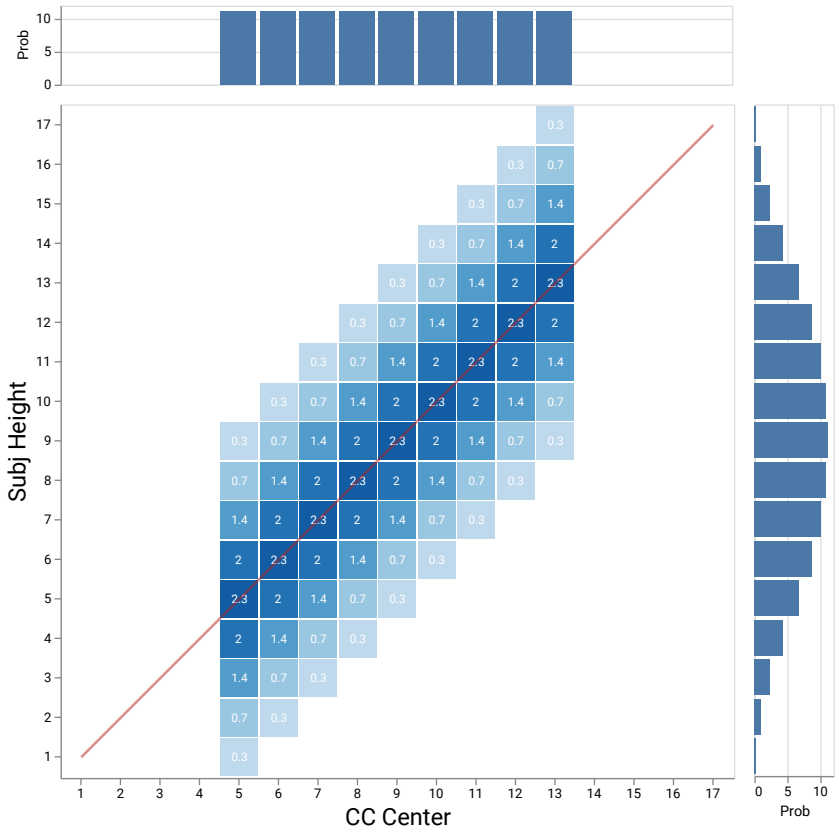


Figure 0: Listener's joint prior distribution over comparison class means (x-axis) and individual heights (y-axis). The top and right margins of the plot display the marginal distributions for the class mean and subject height, respectively. The diagonal line identifies the *Maximum A Priori* worlds, where the subject's height is in the center of the comparison class.

2.2.1. Simulation 1: The Positive Construction

In this first simulation, we simulate the reactions of a pragmatic listener to gradable antonyms in the positive form. The speaker's choices are:

(11) a. **unmarked**: *Jane is tall*
b. **marked**: *Jane is short*
c. **null**: ∅

The listener's lexical uncertainty in this case is essentially as in Lassiter and Goodman 2014, but parameterized to the uncertainty in the distribution of the comparison class. So rather than reasoning about what absolute height a person must exceed to count as *tall*, the listener reasons about how much taller than the mean μ a person must be to count as *tall*, where "the mean" is itself a random variable.

(12) Interpretations: $\left\{ \begin{bmatrix} \textbf{unmarked} & \mapsto & \lambda w.\, \mathrm{ht}_w(\mathrm{j}) \geq \mu_w + \sigma \\ \textbf{marked} & \mapsto & \lambda w.\, \mathrm{ht}_w(\mathrm{j}) \leq \mu_w + \sigma \\ \textbf{null} & \mapsto & \lambda w.\, \mathrm{true} \end{bmatrix} \middle| \; -4 \leq \sigma \leq 4 \right\}$

	Positive		$=$ Equative		$\geq$ Equative		Comparative	
	$\mathbb{L}_1$	$\mathbb{L}_*$	$\mathbb{L}_1$	$\mathbb{L}_*$	$\mathbb{L}_1$	$\mathbb{L}_*$	$\mathbb{L}_1$	$\mathbb{L}_*$
unmarked	2.08	2.56	0.84	0.71	0.11	0.32	-0.74	-0.79
marked	-3.18	-3.29	-1.06	-2.56	-1.52	-1.66	-0.44	-0.66

Table 1: Expected deviations from the comparison class mean, at the first iteration of pragmatic reasoning $\mathbb{L}_1$ and at the stable distribution $\mathbb{L}_*$. For the positive construction, this is the inferred difference between the subject's height and the mean; for the others, this is the inferred difference between the object's height and the mean.

For each interpretation function, the variable σ determines the threshold for tallness/shortness as an offset from the comparison class mean μ. For instance, with $\sigma = 1$, a person's height would have to be at least 1 unit above the mean (half a standard deviation) in order for them to count as *tall*. Since there are no worlds where the subject's height is more than 4 units above or below the mean, we do not worry about cutoffs even greater or lower than this.

What does this listener come to believe when interpreting the unmarked positive sentence in (11a)? Figure 1a shows that as the listener recurses through the pragmatic layers, they become increasingly confident that the subject's height exceeds the center of the comparison class. At $\mathbb{L}_1$, the expected deviation between the subject's height and the class mean is 2.08. At $\mathbb{L}_*$ — the level at which further iteration makes no difference — the deviation is 2.56. See also Table 1.

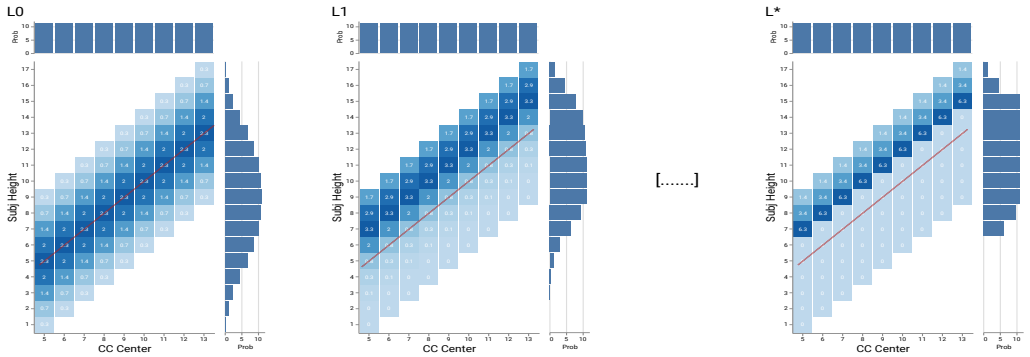
The marked positive utterance in (11b) produces even more evaluative results, shown in Figure 1b. Here the expected deviations at $\mathbb{L}_1$ and $\mathbb{L}_*$ are -3.18 and -3.29, respectively. That is, the listener becomes extremely confident that the subject's height falls well below the class mean. Evaluativity thus is not antonym-specific for the positive construction.

Note that if the subject's height is unknown, as in our simulations, nothing is learned about where *the comparison class* lies after hearing a vague positive utterance like (11a) or (11b); the marginal distributions over the class axis remain uniform throughout the reasoning. In essence then, if we fix in on any specific choice for the class mean, these results replicate Lassiter and Goodman's demonstration that positive-construction evaluativity can be seen as an emergent property of rational communication.

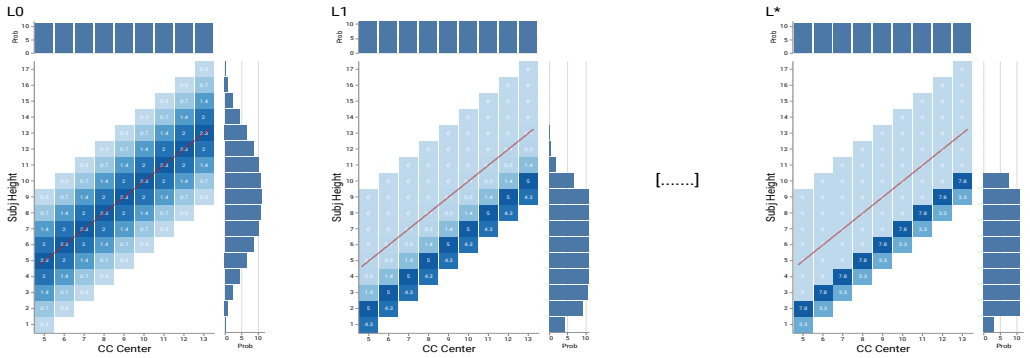
2.2.2. Simulation 2: Exact Equatives and Demonstratives

In the second simulation, we seek to demonstrate that the variability in the comparison class has a significant impact on reasoning about synonymous antonymic degree expressions, where markedness comes into play. This 'antonym-sensitive evaluativity' class of sentences includes the equative, demonstrative, and relative forms of (6b), (6c), and (9). For present purposes, we take all of these sentences to have the same (range of) meanings. Taking the equative as a paradigmatic example, the speaker's utterance choices are:

(13) a. **unmarked**: *Jane is (exactly) as tall as Keisha*
b. **marked**: *Jane is (exactly) as short of Keisha*
c. **null**: $\varnothing$



(a) Interpretations of an utterance with the **unmarked** antonym: *Jane is tall*.



(b) Interpretations of an utterance with the **marked** antonym: *Jane is short*.

Figure 1: Simulation 1 results: predicted interpretations of positive construction utterances. From left to right, the plots progress through successive pragmatic iterations, from the "literal" interpretation at $\mathbb{L}_0$ to the stable distribution at $\mathbb{L}_*$.

We assume that Keisha's height is known to both speaker and listener. This is not essential, and relaxing this assumption so that the listener must also infer Keisha's height produces the same qualitative results,[5] so we discuss only the simpler situation, corresponding to discourses like (9). Both forms in (13) entail that Jane's height is equal to Keisha's height. But crucially the space of hypotheses now includes many distinct worlds where this is the case. The listener still has to make guesses about where Jane's height (now known to be k, say) lies within the comparison class. We thus propose that the listener considers various interpretations of (13) that place k in a certain upper/lower percentile of the comparison class, just as with the positive construction above.

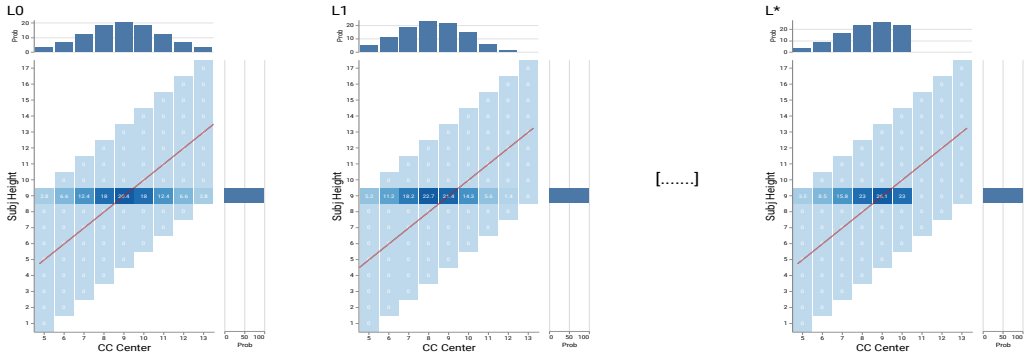
(14) Interpretations: $\left\{ \begin{bmatrix} \textbf{unmarked} & \mapsto & \lambda w.\, \mathsf{ht}_w(\mathrm{j}) = k \;\wedge\; k \geq \mu_w + \sigma \\ \textbf{marked} & \mapsto & \lambda w.\, \mathsf{ht}_w(\mathrm{j}) = k \;\wedge\; k \leq \mu_w + \sigma \\ \textbf{null} & \mapsto & \lambda w.\, \mathsf{true} \end{bmatrix} \;\middle|\; -4 \leq \sigma \leq 4 \right\}$

Again for each interpretation function, the variable σ determines the tallness/shortness cutoff as an offset from the comparison class mean μ. For concreteness, we set Keisha's height k to the median hypothetical height, which is 9. Now we ask, how will a pragmatic listener respond to the utterances in (13a) and (13b)? Of course in either case the listener will learn immediately

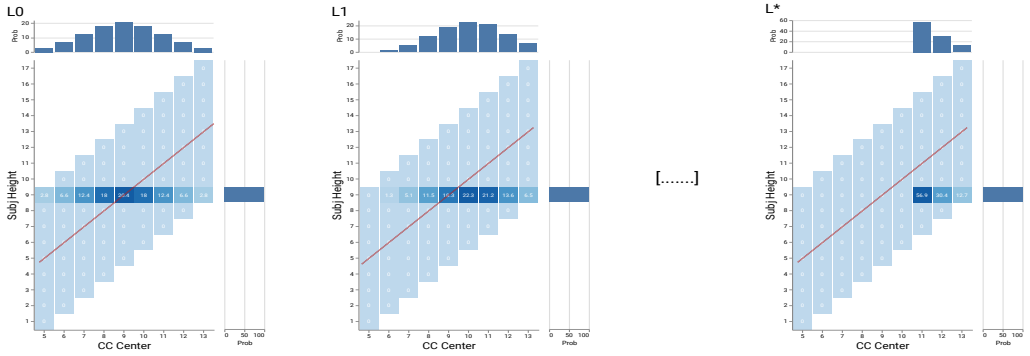
[5]These extended results are available at the link in fn. 4

that Jane's height is 9, eliminating most worlds. But they'll still have to estimate where that height lies within the comparison class. Without thinking at all about the speaker's choices, a literal listener conditioning their beliefs simply on the information that $k = 9$ would yield an (approximately) normal marginal distribution over comparison classes, centered also around 9. This is because *a priori*, what the listener knows is that most people are average. So if all they've learned is that Jane's height is 9, and they know nothing about Jane, they ought to believe that 9 is about average.

But of course this isn't all that the listener has learned. They've also learned that the speaker chose to say (13a) rather than (13b). Nevertheless, Figure 2a shows that even with unbounded iterative reasoning, this choice does not exert much influence on their belief about the comparison class. In fact the *Maximum A Posteriori* hypothesis remains the one in which the comparison class center is the same as Keisha's (and Jane's) height; in other words, that Keisha's height is typical. The $\mathbb{L}_*$ expected deviation is 0.71, as reported in Table 1.



(a) Interpretations of an utterance with the **unmarked** antonym: *Jane is exactly as tall as Keisha*.



(b) Interpretations of an utterance with the **marked** antonym: *Jane is exactly as short as Keisha*.

Figure 2: Predicted interpretations of exact equative utterances, with Keisha's height set at 9.

In contrast, as seen in Figure 2b, hearing the marked equative form — *as short as Keisha* — does produce a dramatic change in the listener's belief about the comparison class. The more the listener thinks about it, the more confident they become that Keisha's height k is in fact well below the mean; in other words, that Keisha's height is atypical. The expected deviation is now -2.56, substantially more evaluative than in the unmarked form.

This completes the main result we set out to establish. Since the antonyms in the positive construction are contradictory no matter how lexical uncertainty is resolved, the utterances do not compete. The only factors influencing the listener's interpretations are informativity, and in order to make the forms sufficiently informative, the listener is driven to evaluative conclusions. The exact equatives in (13), on the other hand, are two linguistic options that both specify precisely what the height of the subject is. Given the synonymy, listeners are compelled to find an explanation when speakers choose to use the marked form. This drives them toward evaluative worlds in which the subject's height is atypical for the comparison class.

In the next simulation we look at minimum-standard equatives, which present a competition intermediate between the previous two. The unmarked form — *at least as tall as* — and the marked form — *at least as short as* — are certainly not synonymous, but they are also not contradictory. We show that the semantic overlap between the forms is enough to lead to antonym-sensitive evaluative conclusions.

2.2.3. Simulation 3: Minimum-Standard Equatives and Demonstratives

The setup is the same as in the previous two sections. The speaker's choices are given in (15) and their range of interpretations in (16). The results of the simulation are shown in Figure 3. Again we assume for the sake of simplicity that the object's height is known to both speaker and listener, though as with the exact equatives the results are not meaningfully different if the object's height is unknown.

(15) a. **unmarked**: *Jane is (at least) as tall as Keisha*
b. **marked**: *Jane is (at least) as short of Keisha*
c. **null**: $\varnothing$

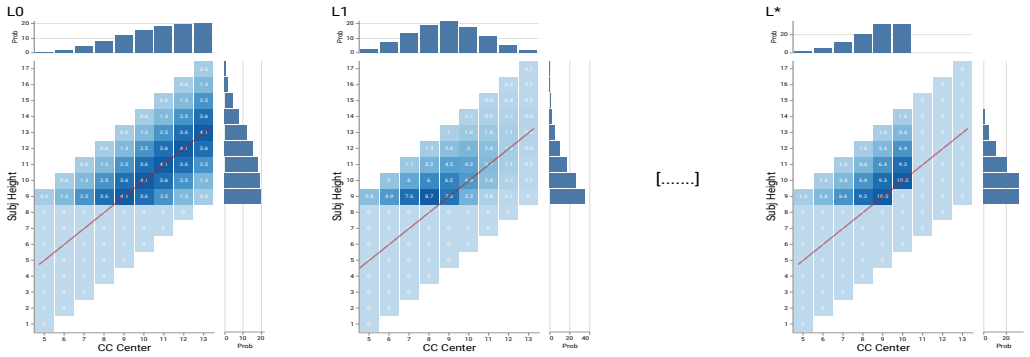
(16) Interpretations: $\left\{ \begin{bmatrix} \textbf{unmarked} & \mapsto & \lambda w.\, \mathsf{ht}_w(\mathsf{j}) \geq k \;\wedge\; k \geq \mu_w + \sigma \\ \textbf{marked} & \mapsto & \lambda w.\, \mathsf{ht}_w(\mathsf{j}) \leq k \;\wedge\; k \leq \mu_w + \sigma \\ \textbf{null} & \mapsto & \lambda w.\, \mathsf{true} \end{bmatrix} \;\middle|\; -4 \leq \sigma \leq 4 \right\}$

A listener who hears (15a) learns immediately that Jane's height is in the upper half of the prior distribution (above $k = 9$). This alone biases their estimate of the comparison class mean toward the upper end of the available spectrum, simply because these high means leave the most room for Jane's height to be above Keisha's height k. But, as seen in Figure 3a, with further pragmatic reasoning about why the speaker chose this utterance over others, the listener's belief about the class mean returns to the central values close to k. Jane's height is likewise inferred to lie above, but relatively close to, this mean.

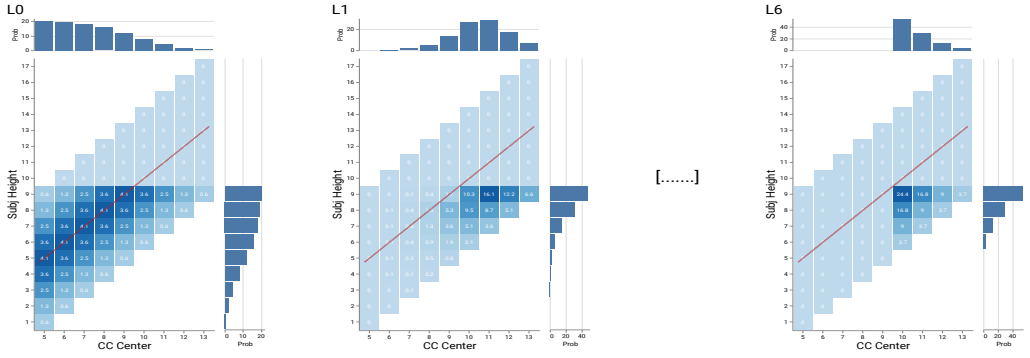
But we think that what is actually of interest in this construction is not the typicality of *Jane's* height, but rather the typicality of Keisha's. Based on judgments from standard projection tests, Rett (2015b) argues that in marked equatives (e.g., *as short as Keisha*) speakers only take for granted the evaluativity of the object (e.g., that Keisha is short). And in any case, it is rational even for a literal listener who learns that Jane's height is at or above Keisha's to come to hold evaluative beliefs about Jane, since all of their prior probability mass must be squeezed into the upper reaches of the distribution. But this has nothing to do with the construction. By this metric, *as tall as Keisha* would be evaluative only in the sense that the predicate *has a height above the mean* is evaluative; it is simply a literal (though perhaps probabilistic) consequence of the truth conditions.

So instead we look to the inferred deviation of Keisha's height k relative to the comparison class. Since in these simulations k is known to be 9, this is just the expected value of the distribution derived by subtracting 9 from the marginal over class means. For the unmarked equative *at least as tall as Keisha*, this expected deviation is 0.11 at $\mathbb{L}_1$ and 0.32 at $\mathbb{L}_*$. We note that this is even less evaluative than the unmarked exact equative *exactly as tall as Keisha* explored in the previous section. However, for the marked equative *at least as short as Keisha*, the expected deviation of k is -1.52 at $\mathbb{L}_1$ and -1.66 at $\mathbb{L}_*$.

The magnitude of the latter score is less than the analogous scores for the marked positive and exact-equative constructions, but still well above any of the scores of the unmarked equatives. This predicts, novelly as far as we know, that two degree constructions traditionally categorized as evaluative can nevertheless differ in the strength of their evaluative inference, with *Keisha is short* at the extreme end, *Jane is exactly as short as Keisha* in the middle, and *Jane is at least as short as Keisha* at the other end (see Table 1 for a summary). And it provides a response to an open question in Rett (2015b), namely, if *at least as tall* and *at least as short* aren't synonymous in the way the antonymous 'exactly' equatives are, why do 'at least' equatives still demonstrate an antonym-sensitive evaluativity pattern?



(a) Interpretations of an utterance with the **unmarked** antonym: *Jane is at least as tall as Keisha*.



(b) Interpretations of an utterance with the **marked** antonym: *Jane is at least as short as Keisha*.

Figure 3: Predicted interpretations of minimum-standard equative utterances, where Keisha's height is known to be 9.

2.2.4. Simulation 4: Comparatives

Finally, as a sort of control measure, we investigate comparative predicates like *taller/shorter than Keisha*. Unlike the minimum-standard equatives in the previous section, the antonymic forms here have no semantic overlap. And unlike the positive construction, there is very little informativity pressure coming from competition with the null message, since these predicates are not vague; they provide quite enough information to be worth the cost of saying. Thus we would expect our model to derive that comparatives are non-evaluative in both forms.

(17) a. **unmarked**: *Jane is at least as tall as Keisha*
b. **marked**: *Jane is at least as short of Keisha*
c. **null**: $\varnothing$

(18) Interpretations: $\left\{ \begin{bmatrix} \textbf{unmarked} & \mapsto & \lambda w.\, \mathsf{ht}_w(\mathsf{j}) > k \;\wedge\; k \geq \mu_w + \sigma \\ \textbf{marked} & \mapsto & \lambda w.\, \mathsf{ht}_w(\mathsf{j}) < k \;\wedge\; k \leq \mu_w + \sigma \\ \textbf{null} & \mapsto & \lambda w.\, \mathsf{true} \end{bmatrix} \;\middle|\; -4 \leq \sigma \leq 4 \right\}$

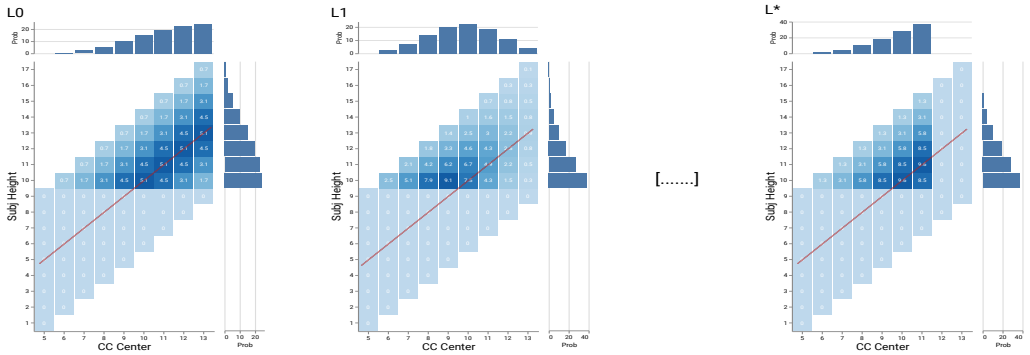
Importantly, though, as with the minimum-standard equatives, we should expect any listener (even a literal one) to develop an evaluative belief about the *subject*'s height when interpreting a comparative. This is because, without knowing anything about the comparison class, learning that Jane's height exceeds Keisha's is enough to conclude that Jane's height is probably above the mean. There are simply more ways for her to be above the mean than below it, given that her height is greater than 9 (or whatever k is inferred to be if it is unknown).

So we look to the expected deviation of the object's height k. The listener's interpretation of comparative forms is plotted in Figure 4. For the unmarked sentence in (17a), the expected deviation of k is -0.74 at $\mathbb{L}_1$ and -0.79 at $\mathbb{L}_*$. This is certainly less evaluative than the positive construction and the marked equatives. It is, in fact, slightly negative. This means that when interpreting (17a), Keisha is not only not presumed to be tall, she is actually guessed, on average, to be slightly shorter than the comparison class mean. Despite appearances, there is nothing especially mysterious about this. It is the other side of the rational coin that leads a literal listener to infer that Jane is probably above the mean; Keisha is technically probably below the mean simply because this leaves open more possibilities in which Jane's height exceeds hers.

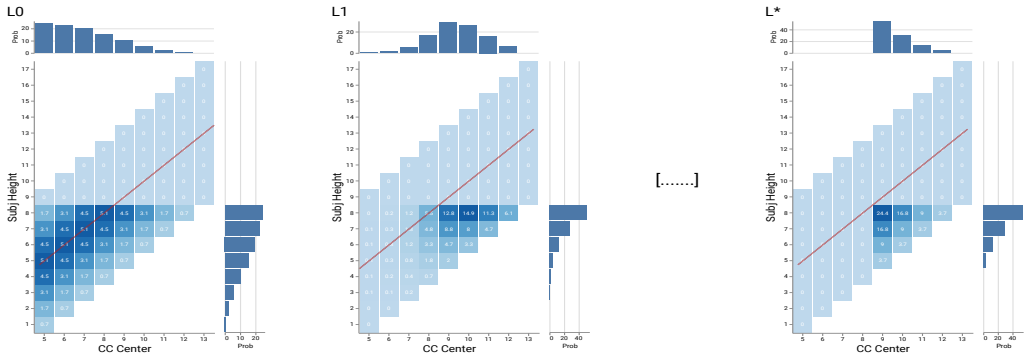
In contrast with the first three simulations, the marked comparative form in (17b) is also not very evaluative. The expected deviation of k is -0.44 at $\mathbb{L}_1$ and -0.66 at $\mathbb{L}_*$. If things were exactly symmetric to the unmarked comparative, we'd expect these numbers to have opposite signs, that is, to be slightly positive. So there is a small and potentially interesting effect here: *taller than Keisha* and *shorter than Keisha* both lead the the listener to guess, on average, that Keisha is slightly below the mean. But the large difference in evaluativity between *at least as short as Keisha* and *shorter than Keisha* is striking (and empirically correct), despite the small difference in truth conditions. This is entirely due to partial competition between antonymic forms and the way that markedness affects that competition.

3. Conclusion

We set out to develop a quantitative account of how and when evaluativity inferences arise from various degree constructions, including the following paradigmatic judgments.



(a) Interpretations of an utterance with the **unmarked** antonym: *Jane is taller than Keisha*.



(b) Interpretations of an utterance with the **marked** antonym: *Jane is shorter than Keisha*.

Figure 4: Predicted interpretations of comparative utterances, with Keisha's height fixed at 9.

Construction	Example	Evaluative?	
		unmarked	marked
Positive	*Jane is tall/short*	✓	✓
Equative	*Jane is* {*exactly, at least*} *as tall/short* as that	✗	✓
Comparative	*Jane is taller/shorter than that*	✗	✗

We argued that Lassiter and Goodman's (2014) rational account of vagueness resolution may be operationalized to derive the evaluativity of the positive construction. But since that construal relies on reasoning about the open argument of the gradable adjective, it provides no handle on the other constructions, where that argument is saturated.

To generalize the game-theoretic reasoning to these latter cases, we recast Lassiter and Goodman's hypotheses about how the heights of relevant individuals are estimated as hypotheses about how these heights relate to their comparison classes, inspired by ideas in Barker 2002. And to account for the antonym-specificity of the equative, we drew on Rett's (2015b) insight that when degree expressions are synonymous, evaluative reasoning is driven by linguistic markedness. This was incorporated into the rational model via strategic assumptions about how gradable language interacts with semantic uncertainty, in accordance with the schematic derivations of Manner implicatures in Bergen et al. 2016.

Perhaps the most interesting aspect of our analysis is that evaluativity is not an all or nothing affair. This is in part because synonymy is not an all or nothing affair either. The listeners we model are faced with the multidimensional problem of simultaneously inferring the speaker's intended meaning and the measures of individuals relative to the measures of their comparison classes. Our assumptions about semantic uncertainty mean that some utterances, like the *at least* equatives, will exhibit *intermediate* amounts of competition between antonymic alternatives, relative to the exact equative and comparative forms. As a result the evaluative conclusions that listeners draw are predicted to be less severe, but still quite robust. This is possibly in line with preliminary experimental results from Brasoveanu and Rett (2018), but more experimental work is necessary to see if such gradient predictions can be detected.

References

Barker, C. (2002). The dynamics of vagueness. *Linguistics and Philosophy 25*, 1–36.

Bergen, L., R. Levy, and N. Goodman (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics 9*, 1–83.

Bierwisch, M. (1989). The semantics of gradation. In M. Bierwisch and E. Lang (Eds.), *Dimensional Adjectives: Grammatical Structure and Conceptual Interpretation*, pp. 71–237. Springer-Verlag.

Brasoveanu, A. and J. Rett (2018). Evaluativity across adjective and construction types: An experimental study. *Journal of Linguistics 54*, 263–329.

Cresswell, M. (1976). The semantics of degree. In B. Partee (Ed.), *Montague Grammar*, pp. 261–292. Academic Press.

Frank, M. and N. Goodman (2012). Predicting pragmatic reasoning in language games. *Science 336*, 998–998.

Kamp, H. (1975). Two theories of adjectives. In E. Keenan (Ed.), *Formal Semantics of Natural Language*, pp. 123–155. Cambridge University Press.

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

Lassiter, D. and N. Goodman (2014). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Proceedings of SALT 23*. Cornell University.

Lehrer, A. (1985). Markedness and antonymy. *Journal of Linguistics 21*, 397–429.

Potts, C., D. Lassiter, R. Levy, and M. Frank (2016). Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics 33*, 755–802.

Rett, J. (2008). *Degree Modification in Natural Language*. Ph. D. thesis, Rutgers University.

Rett, J. (2015a). Modified numerals and measure phrase equatives. *Journal of Semantics 32*, 425–475.

Rett, J. (2015b). *The semantics of evaluativity*. Oxford University Press.