Adjectival vagueness in a Bayesian model of interpretation*

Daniel Lassiter
Stanford University
danlassiter@stanford.edu

Noah D. Goodman
Stanford University
ngoodman@stanford.edu

Abstract We derive a probabilistic account of the vagueness and context-sensitivity of scalar adjectives from a Bayesian approach to communication and interpretation. We describe an iterated-reasoning architecture for pragmatic interpretation and illustrate it with a simple scalar implicature example. We then show how to enrich the apparatus to handle pragmatic reasoning about the values of free variables, explore its predictions about the interpretation of scalar adjectives, and show how this model implements Edgington's (1992; 1997) account of the sorites paradox, with variations. The Bayesian approach has a number of explanatory virtues: in particular, it does not require any special-purpose machinery for handling vagueness, and it is integrated with a promising new approach to pragmatics and other areas of cognitive science.

Edgington (1992, 1997) proposes an attractive unified approach to the Sorites, Lottery, and Preface paradoxes. According to Edgington, these puzzles are all explained by a generalization of classical logic which has the formal structure of the probability calculus, with an accompanying generalized notion of valid reasoning. She gives a number of strong arguments to the effect that a degree-based theory of vagueness with the formal structure of probabilities is preferable to one with the structure of classical fuzzy logic. However, she explicitly disavows the idea that the degrees involved in her account of vagueness *are* probabilities in the usual sense of *rational degrees of credence*; instead, she labels them *verities* and leaves to the side the question of what exactly they are, or why they display the logical structure that they do.

As Douven & Decock (2014) note, the lack of a clear interpretation of Edgington's verities seems to have hindered acceptance of an otherwise very promising theory of vagueness. Douven & Decock propose an interesting, psychologically-oriented answer to these questions by deriving verities from a version of the Conceptual Spaces model of concepts (Gärdenfors 2000). As they note, a convincing derivation of this type is crucial for the overall plausibility of a probabilistic model of vagueness.

In this paper we propose an alternative explanation for the probabilistic structure of verities, focusing on the case of (relative) scalar adjectives such as *tall, heavy* and *happy*. We show that

* Thanks to Michael Franke, Chris Potts, Chris Kennedy, Adrian Brasoveanu, Paul Egré, Alexis Wellwood, Lenhart Schubert, Richard Dietz, two *Synthese* reviewers, three *SALT 23* reviewers, participants in our 2013 ESSLLI course "Probability in semantics and pragmatics", participants in Lassiter's 2014 NASSLLI course "Language understanding and Bayesian inference", and audiences at SALT 23, Stanford, Northwestern, Brown, U. Chicago, and UT-Austin. This paper is modified and extended from Lassiter & Goodman 2013, which appeared in the proceedings of the conference *Semantics & Linguistic Theory 23*. This work was supported by a James S. McDonnell Foundation Scholar Award to NDG and by ONR grant N00014-13-1-0788.

these adjectives' context-sensitive interpretation, sorites sensitivity, and the existence of borderline cases can be derived from a general Bayesian theory of pragmatics (Goodman & Stuhlmüller 2012; Frank & Goodman 2012; Goodman & Lassiter 2015). This theory is framed within a Bayesian approach to cognitive science that has flourished in recent years (Tenenbaum, Kemp, Griffiths & Goodman 2011), and has been used recently to account for a wide variety of cognitive activities (e.g., learning, reasoning, categorization, vision, motor control) as well as a number of detailed pragmatic and psycholinguistic phenomena.

The theory's application to vague terms depends on the specific lexical semantics of expressions like *heavy*, which are generally thought in linguistic semantics to rely on a free threshold variable: "heavy" is interpreted as "heavier than θ". In order to assign an interpretation to utterances containing *heavy*, listeners must use the available information — context and knowledge of the speakers' beliefs and goals — to estimate this latent variable. We propose an approach to this estimation problem and use computer simulations to demonstrate its predictions and show that the empirical phenomena under consideration can be derived. In effect, our model suggests credence functions which are able to do most of the work of Edgington's verities, where the relevant kind of uncertainty is uncertainty about a speaker's intended message — i.e., about the use to which the speaker intends to put the semantically flexible linguistic resources that their language makes available.

This way of approaching the problem has a number of conceptual advantages. In particular, it allows us to make headway on the difficult question of how communication with vague expressions is possible, in the information-theoretic sense. Second, it provides answers to two questions which are not usually considered in the literature on degree-theoretic (including probabilistic) accounts of vagueness: How are degree functions for specific expressions derived? And how and why do the degree functions of relative adjectives shift in response to a choice of reference class — in particular, in response to *statistical* properties of a reference class? Finally, the theory does not rely on any *ad hoc* semantic or pragmatic machinery for vague expressions; rather, it is a direct application of principles that have much independent motivation from recent work in formal semantics and pragmatics and in cognitive science.

Several caveats are in order here. First, our model is not in competition with an approach to vagueness based on Conceptual Spaces or other probabilistic theories of conceptual structure. While our account extends readily to scalar expressions that are not adjectives—for instance, quantifiers such as *many* and *few*, and verbs such as *love* and *fear*—it does not account for the vagueness of expressions that lack a scalar basis, such as *fruit*, *cup*, or *bird*. A probabilistic account of non-scalar vagueness would mesh well with our theory, but must be derived from additional assumptions, such as those of the Conceptual Spaces theory. Second, we simplify the empirical picture by considering only relative adjectives like *tall* and *heavy*, leaving aside the interestingly different class of "absolute" adjectives such as *full* and *empty*, which was described and connected with vagueness phenomena in an important paper by Kennedy (2007). (We believe that our theory does a good job of accounting for absolute adjectives as well, though: see Lassiter & Goodman 2013 for the details.)

Third, the present account is intended as an answer to the psychological question of how people *understand and use* scalar adjectives. We do not propose an answer to the metaphysical questions that have occupied much of the discussion of vagueness, involving when a scalar adjective *really*

is applicable to an object, and to what degree. (The model is compatible with various proposals about the latter questions, including a degree theory along Edgington's lines, as we discuss briefly.) On some plausible assumptions about the nature of meaning, the latter type of question should be illuminated — perhaps even resolved — by an answer to the former. However, we will not attempt this philosophical project here.

1 Explananda

Scalar adjectives occur in a basic "positive" form — *tall, short, happy, sad, heavy, light* — as well as various modified forms: *exactly 4 feet tall, shorter than Harry, very happy, at least as sad as Jane, extremely heavy, much lighter than a truck*, etc. In this paper we focus on the positive form, which is the usual example in discussions of adjectival vagueness. (It is certainly not the only kind of vague adjective, though: in the list above, *very happy, extremely heavy*, and *much lighter than a truck* are vague as well.) In the positive form, these adjectives display a number of interesting empirical and theoretical properties which we will attempt to explain in a unified way.[1]

First, the meanings of relative adjectives in the positive form are highly context-dependent. A cheap house is likely to be much more expensive than an expensive book. Similarly, the natural interpretation of "big" displays enormous variation among the following noun phrases: *big microbe, big finger, big baby, big football player, big tree, big building, big city, big planet, big star*. The driving force behind this variation seems to be the fact that these adjectives are interpreted in a "norm-related" way (Fara 2000): they indicate that the object that the noun phrase is predicated of has a degree of the scalar property in question (cost, size) which is somehow significantly greater than the norm for a reference class (a.k.a. comparison class). More specifically, the interpretation of these adjectives is relativized to **statistical** properties of a reference class. So, for example, a house which counts as "expensive" in Atlanta might be cheaper than a house which counts as "cheap" in San Francisco, because house prices in San Francisco are generally much higher. The comparison class is usually supplied implicitly, but can be explicit: for example, someone could be *big for a 16-year-old* but at the same time *not big for a football player*. On these properties of vague scalar adjectives see Kamp 1975; Klein 1980; Kennedy 2007; Bale 2011; Solt 2011; Lassiter 2015 among many others.

Second, vague adjectives admit of borderline cases, as illustrated in (1).

(1) [Almost all houses in this neighborhood cost $300,000-$600,000.]
   a. The Williams' $1,000,000 house is expensive.
   b. The Clarks' $75,000 house is not expensive.
   c. The Jacobsons' $475,000 house is ___?___.

In this kind of context, speakers who are asked "Is the Jacobsons' house expensive?" frequently express uncertainty, hedge, or refuse to answer the question. This suggests that the Jacobsons'

1 More precisely, these are characteristics which are robust especially (perhaps only) for the *relative* adjectives on which we focus here. Absolute adjectives such as *full, empty, wet, dry, open, closed, safe, dangerous* are somewhat different: see Kennedy & McNally 2005; Kennedy 2007; Lassiter 2015; Morzycki to appear.

house is a borderline case of *expensive* — one for which neither "expensive" not "not expensive" feels like an appropriate description.

Second, vague adjectives are susceptible to the sorites paradox.

(2) a. A house that costs \$10,000,000 is expensive (for this neighborhood).
b. A house that costs \$1 less than a house that is expensive (for this neighborhood) is also expensive (for this neighborhood).
c. ∴ A house that costs \$1 is expensive (for this neighborhood).

Finally, there is a deep puzzle brought out by the observation that much of ordinary language is vague. If the meanings of vague expressions are indeterminate, how can they be used to communicate meaningful information? The most precise and practically useful model of communication available to us — the noisy-channel model due to Shannon (1948) — does not extend in an obvious way to communication with expressions with indeterminate meanings. This *could* be construed as an argument against the application of information theory to the study of communication with natural languages. We prefer to view it as a challenge to state a model of vagueness which is compatible with the only realistic model of communication available, and which makes precise predictions about what information vague terms convey — the mapping from prior to posterior — relative to a context.

2 Bayesian Pragmatics

Throughout the paper we will assume a common variant of Montague's (1973) framework for compositional semantics, with modifications due to Gallin (1975) and the addition of a basic type d for degrees/thresholds. We also simplify by ignoring intensionality where it is not directly relevant. The details of our semantic assumptions will not play a major role, however: the main convention that should be kept in mind is the assumption that the language is interpreted by a function $[\![\cdot]\!]$ which maps English expressions to expressions in a simply typed λ-calculus.

Our pragmatic theory builds on accounts which emphasize the importance of coordination, in particular developments of Grice 1957, 1989 on game-theoretic principles (Lewis 1969; Clark 1996; Benz, Jäger & van Rooij 2005; Jäger 2007; Potts 2008; Franke 2009, 2011; Jäger & Ebert 2009). We follow closely recent work in Bayesian pragmatics (Frank & Goodman 2012; Bergen, Goodman & Levy 2012; Goodman & Stuhlmüller 2012; Smith, Goodman & Frank 2013; Goodman & Lassiter 2015), which combine Gricean and game-theoretic influences with an approach to inference and decision-making under uncertainty which has been very influential in recent cognitive science (Pearl 2000; Griffiths, Kemp & Tenenbaum 2008; Tenenbaum et al. 2011). Related ideas can be found in Golland, Liang & Klein 2010; Lassiter 2012; Franke 2012a,b; Kehler & Rohde 2013; Vogel, Bodoia, Potts & Jurafsky 2013a; Vogel, Potts & Jurafsky 2013b.

Bayesian models in cognitive science rely on two crucial formal assumptions, familiar from Bayesian epistemology. First, an agent's subjective uncertainty is represented as a probability distribution $P(\cdot)$ over a set of propositions (sets of worlds). Second, the effect on the agent of learning some proposition A is to cause her to adjust her prior probability distribution $P(\cdot)$ to a posterior $P(\cdot|A)$, i.e., to employ Bayesian inference.

2.1 Bayesian inference in brief

This section reviews some basic notions in probability and Bayesian inference that will be used repeatedly in later parts of the paper. Readers familiar with this apparatus will be able to skip to section 2.2.

Probabilistic information states are an enrichment of the sets-of-worlds picture of information states familiar from formal epistemology and formal pragmatics. The addition of a measure $P(\cdot)$ over propositions allows us to make fine-grained epistemic statements, not only about what is possible, impossible, or necessary, but also about what is more or less likely. $P(\cdot)$ is constrained as follows:

(3) a. For all $A \subseteq W$, $P(A) \in [0,1]$.
b. $P(W) = 1$.
c. For any disjoint $A, B \subseteq W$, $P(A \cup B) = P(A) + P(B)$.

Second, Bayesian models assume that, upon learning some proposition A, agents update their information state by conditioning on A, yielding a new probability distribution which continues to obey the constraints in (3).[2]

(4) $$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

In many cases, it is more straightforward to calculate $P(B|A)$ using **Bayes' rule** (but note that this rule is a straightforward consequence of the definition of conditional probability in (4)). If $\mathbf{B} = \{B_1, B_2, \ldots\}$ is a partition of W, then we have:

(5) $$P(B_i|A) = \frac{P(A|B_i) \times P(B_i)}{\sum_{j=1}^{|\mathbf{B}|} P(A|B_j) \times P(B_j)}$$

Suppose that the elements of $\mathbf{B}$ are hypotheses about the process by which observation A was generated. We can then think of the core Bayesian assumption as follows: the probability that hidden cause B_i is true, given that we have made observation A, is proportional to the product of two terms: (i) the probability that we would have observed A if hidden cause B_i were true, and (ii) the probability that we assigned to hidden cause B_i before we observed A.

(6) $$P(B_i|A) \propto P(A|B_i) \times P(B_i)$$

For convenience, we will sometimes write instances of Bayes' rule in this form, without specifying the full space of alternative hypotheses B_j which we would have to consider in order to calculate $P(B_i|A)$. The missing denominator is a constant which will be the same regardless of the choice of B_i; it is required to ensure that the resulting distribution sums to 1.

2 The ratio definition of conditional probability is convenient and simple, but not at all crucial for us. We would also be content to take conditional probability as basic, as many philosophers of probability have recommended (e.g., Hájek 2003).

2.2 Motivation and assumptions

In applying this model to linguistic communication, we assume that speakers and listeners maintain probabilistic models of each others' utterance planning and interpretation processes, and that these models drive pragmatic language use. In particular, listeners use their models of speakers' utterance choice to make more informed interpretive choices than would be possible if they simply updated their information states with the information that the utterance's semantic interpretation is true. The model thus encodes back-and-forth pragmatic reasoning as pictured in Figure 1. Later in the paper we will also propose a way for listeners to use a model of the speaker to resolve context-sensitivity on Bayesian principles, by jointly inferring the state of the world and the values of semantic variables.

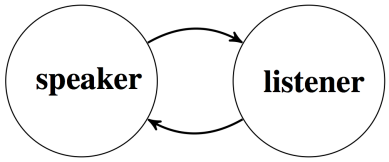

Figure 1 Recursive pragmatic reasoning.

The most straightforward way to implement recursive reasoning of this type would be along the following lines: a listener L updates her information state, given that some utterance has been made, by reasoning about how the speaker would have chosen utterances or other actions in various possible worlds, and weighting the result by the probability that those worlds are indeed actual.

$$P_L(w|u) \propto P_S(u|w) \times P_L(w) \tag{7}$$

Conversely, a speaker chooses utterances by reasoning about how the listener will interpret the utterance, together with some private utterance preferences $P_S(u)$ (representing, for example, frequency effects or a preference for brevity and ease of retrieval).

$$P_S(u|w) \propto P_L(w|u) \times P_S(u) \tag{8}$$

These equations are both instantiations of Bayes' rule. However, since they are mutually recursive the reasoning could go on forever, unless we impose some bound. In addition, it is not obvious where in (7) and (8) literal meaning, as studied in compositional semantics, intrudes (cf. Franke 2009: §1).

The solution adopted here goes back to chapter 1 of Lewis's (1969) *Convention: A Philosophical Study*. Iterated reasoning grounds out in first-order expectations about others' likely actions, and layers of reasoning about others' reasoning are built on top of this basic expectation. Recently this idea has been developed in multiple, partly overlapping lines of research in game-theoretic pragmatics and in cognitive science (Franke 2009, 2011; Jäger & Ebert 2009; Xu & Tenenbaum 2007; Goodman & Stuhlmüller 2012; Frank & Goodman 2012; Goodman & Lassiter 2015).

In this paper we focus on a particularly simple version of this model: the interpretation process of a listener who uses literal interpretation as a base case and reasons to some finite depth, as

described in detail by Franke (2009). The *pragmatic listener* L_1 reasons about the utterance choices of a simulated speaker S_1, who reasons about the interpretation of a *literal listener* L_0, who does not reason pragmatically. The literal listener provides a hook for the compositional semantics to provide conventionalized semantic information to the pragmatic reasoning process. The model is easily extended to speakers and listeners who reason to greater depths, but, as we will see, robust pragmatic effects can arise already at level 1. A graphical depiction of the model is given in Figure 2. We describe these model components in detail in the next subsection.

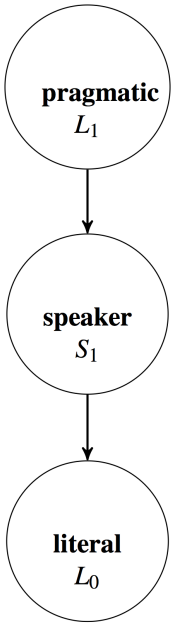

Figure 2 Bounded pragmatic reasoning grounds out in first-order expectations.

Importantly, the non-maximal speaker and listener models — S_1 and L_0 below — exist only as part of the pragmatic listener's psychology: we do not want to commit to the existence of ultra-naïve listeners, or to speakers who believe that they are speaking to such listeners.[3] This point was made by Lewis in his analysis of basic coordination behavior, where he pointed out that such iterated reasoning

> is *not* an interaction back and forth between people. It is a process in which *one* person works out the consequences of his beliefs about the world—a world he believes to include other people who are working out the consequences of their beliefs, including their belief in other people who ... By our interaction in the world we acquire various high-order expectations that can serve us as premises. In our

3 We will not address speaker modeling here, but our model does make predictions: to the extent that listeners do reason to level 1, a reflective speaker should choose utterances by reference to an L_1 model who is reasoning about an S_1 who is reasoning about an L_0. See also Qing & Franke 2014 for a model of vague interpretation which builds on the one described here, but places greater emphasis on speaker modeling.

subsequent reasoning we are windowless monads doing our best to mirror each other, mirror each other mirroring each other, and so on. (Lewis 1969: 32)

2.3 Literal listener

The literal listener L_0 is defined as an agent who responds to an utterance u in two steps: calculate $[\![u]\!]$, the literal interpretation of u in the relevant language, and condition the prior information state on the truth of $[\![u]\!]$.

$$P_{L_0}(A|u) = P_{L_0}(A|[\![u]\!] = 1) \tag{9}$$

L_0 is essentially a probabilistic version of the interpreter discussed by Stalnaker (1978) and in much work in dynamic semantics, who responds to utterances by simply assuming that they are true. There is even a close relationship between the update operations: Stalnakerian update is set intersection, and conditionalization is equivalent to intersection followed by renormalization of the measure. (We consider the case in which $[\![u]\!]$ contains free variables below. See Goodman & Lassiter 2015 for discussion of lexical and syntactic ambiguities, which we do not deal with here.)

2.4 Speaker model

We model a generic speaker S_1 as an agent who attempts to make statements which are informative relative to the current topic of conversation/Question Under Discussion (QUD: Ginzburg 1995a,b; Roberts 1996). Equivalently, the conversation determines a random variable for which the speaker knows the true value, but the listener does not, and attempts to select an utterance which will transmit as much information as possible about this variable. The random variable/QUD-relativity of the speaker model is important because we do not wish to predict that speakers will say things that are highly informative in a global sense if they are irrelevant to the current conversation. In our model, the QUD provides a set of possible answers A over which the informativity of a potential utterance is calculated. A QUD might be "Who came to the party?", or "How many people came?", or "How tall is Al?". The choice of QUD is constrained, but not fully determined, by overt questions, the information structure (e.g., prosody) of the utterance, and various other aspects of the history of the discourse. When further specification is required, we would opt for an expansion of the model presented here to include inference of the QUD, as discussed by Kao, Wu, Bergen & Goodman (2014). We will not deal with this additional complication here, though.[4]

The speaker and listener share the goal of coordinating utterance and interpretation so as to maximize the probability that the listener will infer the correct answer to the QUD. We thus define the utility of u for speaker S_1 to be proportional to its informativity to the literal listener L_0 about the

4 On the equivalence between random variables and question denotations (on the Groenendijk & Stokhof 1984 interpretation), see van Rooij 2003. Note in particular that an answer A is just a set of possible worlds, i.e., a proposition which is a cell in the partition which the question denotes.

A notable simplification in our model is the assumption that the speaker knows the true answer with certainty. This assumption could be relaxed either by allowing that the speaker samples a possible world from his personal probability distribution and then proceeds with the calculations described here, or by using expected informativity instead of simple informativity in the utility function. For relevant discussion see Goodman & Lassiter 2015.

true answer A, minus a non-negative cost $C(u)$. Following Frank & Goodman (2012), we quantify the informativity of u as the negative surprisal (positive log probability, Shannon 1948) of the true answer for L_0, once L_0 has conditioned on the literal truth of the utterance. In addition, there is a term $C(u)$ representing the intrinsic cost of producing u for the speaker. Relevant factors might include difficulty of articulation and difficulty of retrieval. Here we assume that cost increases monotonically with difficulty of articulation, which we approximate very roughly as length in words.

For the S_1 model this gives us the following utility function, where $\mathbb{U}_{S_1}(u;A)$ is the utility of utterance u for speaker S_1 on the assumption that the answer to the QUD is A.

$$\mathbb{U}_{S_1}(u;A) = \ln(P_{L_0}(A|u)) - C(u) \tag{10}$$

With this utility function in hand we turn to a specification of the speaker's choices given some possible utterances u, each with a utility $\mathbb{U}_{S_1}(u;A)$. Most work in decision theory and game theory assumes that agents deterministically choose the action with the highest utility, or choose randomly among maximal options if there is a tie. We employ a relaxed version of this model according to which agents choose stochastically, i.e., that speakers *sample* actions with the probability of making a choice increasing monotonically with its utility. **Soft-max** choice rules of this type are widely employed in psychology and machine learning (Luce 1959; Sutton & Barto 1998). Apparently sub-optimal choice rules of this type have considerable psychological motivation. They can also be rationalized in terms of optimal behavior for an agent whose computational abilities are bounded by time and resource constraints, but who can efficiently approximate optimal choices by sampling from a probability distribution (Vul, Goodman, Griffiths & Tenenbaum 2014).

$$P_{S_1}(u|A) \propto \exp(\alpha \times \mathbb{U}_{S_1}(u;A)) \tag{11}$$

This choice rule has a parameter $\alpha > 0$ which determines how closely stochastic choice approximates deterministic utility-maximization. With $\alpha = \infty$, we would recover the choice rule typically used in game theory. In simulations reported below we set α to a lowish value of 4. The qualitative results reported are not extremely sensitive to the value of this parameter, though very high and very low settings would yield (respectively) over- and under-informative interpretations for vague expressions.

We must assume a space of alternative utterances $u' \in \mathbf{ALT}$ in order to find the normalizing constant for (11).

$$P_{S_1}(u|A) = \frac{\exp(\alpha \times \mathbb{U}_{S_1}(u;A))}{\sum_{u' \in \mathbf{ALT}} \exp(\alpha \times \mathbb{U}_{S_1}(u';A))} \tag{12}$$

This choice can influence the qualitative behavior of the model, and it is not currently well-investigated from an empirical or computational perspective. For present purposes we will assume (essentially following Fox & Katzir (2011)) that the alternative utterances considered are a subset of the possible answers to the QUD. We will also assume that speakers have the option of saying nothing. In particular, we will assume that a sentence with a scalar adjective such as *Al is tall* is interpreted by consulting two alternatives: the utterance *Al is short* and the action of saying nothing

($\varnothing$). Similarly, *I ate some of your cookies* might be evaluated relative to the simplified alternative set $\{$*I ate some of your cookies, I ate all of your cookies*$, \varnothing\}$. The qualitative results reported below do not differ in major respects if we also consider other well-chosen alternatives, e.g. *Al is very tall/short* and *Al is medium height*; however, it is possible to induce counter-intuitive model behavior with certain alternative sets. A major desideratum in future work will be to get a clearer picture of how speakers and listeners choose a realistic but manageable set of alternatives for pragmatic reasoning, and more generally how people choose an action set under relatively unconstrained conditions.

2.5 Pragmatic listener

The *pragmatic listener* L_1 interprets utterances u using Bayesian inference, assigning to each A a probability proportional to the product of (a) the probability that the speaker would have chosen to employ u if A were the true answer, and (b) the prior probability that A is true.

$$P_{L_1}(A|u) \propto P_{S_1}(u|A) \times P_{L_1}(A) \tag{13}$$

$P_{L_1}(A)$ specifies L_1's background knowledge about answers to the QUD. For example, if the QUD is *How tall is Al?* and L_1 knows only that Al is an adult man, then $P_{L_1}(A)$ is an estimate of the distribution of heights among adult men. (14) is normalized by the total posterior probability of all possible answers A' given that u was in fact chosen.

$$P_{L_1}(A|u) = \frac{P_{S_1}(u|A) \times P_{L_1}(A)}{\sum_{A'} P_{S_1}(u|A') \times P_{L_1}(A')} \tag{14}$$

3 Example application: Scalar implicature

To illustrate the workings of the model, we work through a simple scalar implicature example in detail, showing how the model accounts for defeasible pragmatic enrichment in a case where free variables are not relevant.[5] The next section will show how to adapt this pragmatic reasoning to infer the value of the free threshold variable that is present in vague scalar adjectives.

Emma is saving 6 cookies for dessert — two for each member of her family. She leaves them on the kitchen counter while she goes to the bank. She calls home while she is waiting in line. Her husband Dan reports to her: "Charlie ate some of the cookies." From the literal meaning of this sentence, there is a very strong inference that (assuming Dan is reliable and well-informed) the number of cookies Charlie ate is greater than zero. Emma will typically acquire more information than this, though: she will also learn that the number he ate is less than six, i.e., that he didn't eat *all* of the cookies. Of course this is a defeasible inference — conceivably, Dan is lying or confused.

5 Plausibly, there is a domain restriction variable in the quantifier's meaning which must be inferred (Stanley & Szabó 2000). In the case at hand, the domain restriction does not vary across the alternatives under consideration, and so we can safely ignore this variable. However, there are many cases in which this variable is not clearly given, and its inference will likely interact with other aspects of the pragmatic reasoning considered here.

Nevertheless, given what Dan said, Emma can reasonably expect some dessert to be left when she gets home.

Where does this additional, defeasible information come from? The usual story in the pragmatics literature, going back to Grice (1975), goes roughly as follows. Emma reasons that, if Charlie had eaten none of the cookies, Dan — being a reliable type — would have said "Charlie ate none of the cookies". If Charlie had eaten all of the cookies, Dan could truthfully say "Charlie ate some of the cookies", as he did; but he could also have said "Charlie ate all of the cookies", and this utterance would have been strictly more informative about the family's dessert prospects. Furthermore, there is no obvious alternative explanation of why Dan would have chosen to say "some" if "all" were true: they are about equally effortful, and no obvious considerations of (e.g.) politeness seem to be relevant. So, given Dan's choice to use the sentence with "some", Emma concludes (tentatively and defeasibly) that the sentence with "all" replacing "some" would not be true. So, Charlie ate some of the cookies, but he didn't eat all of them: at least one is left.

The key to the above reasoning is that the listener enriches the interpretation beyond the literal meaning on the basis of a *rationalization* of the speaker's observed choice, based on a *background model* of the speaker's decision-making processes. Bayesian back-and-forth reasoning captures the informal Gricean explanation in a precise way. It also explains the defeasibility of this pragmatic inference, since the result is a probabilistic inference about the speaker's choices in various possible scenarios, and choice is assumed to be stochastic rather than deterministic.

Let the QUD be *How many cookies did Charlie eat?*, and let L_1 be Emma, whose interpretation process invokes a simulated version S_1 of Dan. Emma must have assumed that Charlie wouldn't eat any of her cookies (otherwise, she wouldn't have left them unprotected), but we can imagine that she had no expectations about how many he would eat if he did have some. So, let the prior probability of *Charlie ate n cookies* be high for $n = 0$ and distributed uniformly over $n \in \{1,2,3,4,5,6\}$.

$$P_{L_{0/1}}(A = \textit{Charlie ate n cookies}) = \begin{cases} .94 \text{ if } n = 0 \\ .01 \text{ if } n \in \{1,2,3,4,5,6\} \end{cases} \tag{15}$$

Fix a set of alternative utterances $\mathbf{ALT} = \{\textit{NONE}, \textit{SOME}, \textit{ALL}\}$. Referring to equation (**??**), we see that, in order to find the posterior $P_{L_1}(A|u)$, we need to know $P_{S_1}(u'|A)$ for each u' in this **ALT**. Recall (combining equations (10) and (11)) that this quantity can be found as

$$P_{S_1}(u'|A) \propto \exp(\alpha \times [\ln(P_{L_0}(A|u')) - C(u')]). \tag{16}$$

To illustrate the qualitative pattern, let's set $\alpha = 4$ and $C(u) = 2/3 \times \textit{length}(u)$, where length is measured in words. So, $C(u') = 4$ for all $u' \in \mathbf{ALT}$. Then L_1 goes through the following reasoning in attempting to rationalize Dan's choice to utter $u = \textit{SOME}$ (i.e., *Charlie ate some of the cookies*). All of the conclusions described here are summarized in graph form in Figures 3 and 4, which the reader may wish to consult as she works through the example.

Suppose $n = 0$. The truth is that $A = \textit{Charlie ate 0 cookies}$.

- Since *NONE* entails $n = 0$ and excludes $n > 0$, the literal listener's posterior probability of the true state $n = 0$ given $u = \textit{none}$ is $P_{L_0}(n = 0|u = \textit{NONE}) = 1$.

- Since *SOME* entails $n > 0$, $P_{L_0}(n = 0|u = SOME) = 0$.
- Since *ALL* entails $n > 0$, $P_{L_0}(n = 0|u = ALL) = 0$.

Plugging these results into equation (16), we find:

- $P_{S_1}(NONE|n = 0) \propto \exp(4 \times [\ln(1) - 4]) \approx 1.1 \times 10^{-7}$.
- $P_{S_1}(SOME|n = 0) \propto \exp(4 \times [\ln(0) - 4]) = 0$.
- $P_{S_1}(ALL|n = 0) \propto \exp(4 \times [\ln(0) - 4]) = 0$.

The normalized probability that S_1 will produce each utterance is computed by dividing its non-normalized probability by the sum of the non-normalized probabilities of all $u' \in \mathbf{ALT}$. This sum is approximately $1.1 \times 10^{-7} + 0 + 0 = 1.1 \times 10^{-7}$. So, $P_{S_1}(NONE|n = 0) = (1.1 \times 10^{-7})/(1.1 \times 10^{-7}) = 1$, and *SOME* and *ALL* have probability zero under this scenario.

In other words, S_1 *might* say *NONE* if $n = 0$ and *definitely wouldn't* say *SOME* or *ALL*, since these utterances would lead the literal listener to assign probability zero to the true state. So, of the alternatives that the pragmatic L_1 is considering in attempting to rationalize the observed utterance, only *NONE* would be possible if $n = 0$. Given this, the posterior probability of $n = 0$ for L_1, given the observed utterance *SOME*, will be

$$\begin{aligned} P_{L_1}(n = 0|u = SOME) &\propto P_{S_1}(u = SOME|n = 0) \times P_{L_1}(n = 0) \\ &= 0 \times .94 \\ &= 0 \end{aligned} \tag{17}$$

Suppose $n = 1$**.** Here *NONE* and *ALL* exclude the true state, so $P_{L_0}(n = 1|u = NONE) = P_{L_0}(n = 1|u = ALL) = 0$. In the case of *SOME*, we can find the literal listener's posterior probability as the prior probability normalized by the total probability of the states where this utterance is true, using the information in (15).

$$\begin{aligned} P_{L_0}(n = 1|u = SOME) &= P_{L_0}(n = 1|[\![SOME]\!] = 1) \\ &= P_{L_0}(n = 1|n \in \{1,2,3,4,5,6\}) \\ &= P_{L_0}(n = 1)/P_{L_0}(n \in \{1,2,3,4,5,6\}) \\ &= .01/(6 \times .01) \\ &= 1/6 \end{aligned} \tag{18}$$

Now we can plug this calculation into the speaker model:

- $P_{S_1}(NONE|n = 1) \propto \exp(4 \times [\ln(0) - 4]) = 0$.
- $P_{S_1}(SOME|n = 1) \propto \exp(4 \times [\ln(1/6) - 4]) \approx 8.7 \times 10^{-11}$.
- $P_{S_1}(ALL|n = 1) \propto \exp(4 \times [\ln(0) - 4]) = 0$.

And since again only one option has positive probability, the normalized probabilities are 0, 1, and 0 respectively. This information is used by the pragmatic listener to find the non-normalized posterior probability of $n = 1$ given *SOME*:

$$\begin{aligned} P_{L_1}(n = 1 | u = SOME) &\propto P_{S_1}(u = SOME | n = 1) \times P_{L_1}(n = 1) \\ &= 1 \times .01 \\ &= .01 \end{aligned} \tag{19}$$

Suppose $n \in \{2, 3, 4, 5\}$. These are precisely parallel to $n = 1$, since *NONE* and *ALL* are false in all of these scenarios as well, and their prior probabilities are equal to that of $n = 1$. Each of these scenarios has posterior probability proportional to .01.

Suppose $n = 6$. Now things get interesting: we have two true utterances in our alternative set, *SOME* and *ALL*. If $u = SOME$, then L_0's posterior probability of the true state will be $1/6$, for the same reasons spelled out above: L_0 simply conditions on the truth of *SOME*, and this is true when $n = 6$. However, suppose that $u = ALL$: then only $n = 6$ is compatible with the observed utterance, and so L_0 will assign probability 1 to $n = 6$ and 0 to all other states. As a result, there is a huge difference in informativity between *SOME* and *ALL* — the true state has probability six times higher if $u = ALL$ than it does if $u = SOME$.

This asymmetry leads to a large difference in S_1's production probabilities when $n = 6$:

- $P_{S_1}(NONE | n = 6) \propto \exp(4 \times [\ln(0) - 4]) = 0.$
- $P_{S_1}(SOME | n = 6) \propto \exp(4 \times [\ln(1/6) - 4]) \approx 8.7 \times 10^{-11}.$
- $P_{S_1}(ALL | n = 6) \propto \exp(4 \times [\ln(1) - 4]) \approx 1.1 \times 10^{-7}.$

As a function of the informativity difference, S_1 is much more likely to produce *ALL* than *SOME* when *ALL* is true. The normalized probabilities are

- $P_{S_1}(NONE | n = 6) = 0.$
- $P_{S_1}(SOME | n = 6) \approx .001.$
- $P_{S_1}(ALL | n = 6) \approx .999.$

That is, a speaker who is trying to be informative *almost certainly* won't use *SOME* if they have the option to use *ALL* and *ALL* is true.[6] Using this information, the pragmatic speaker L_1 can derive

6 Note that the precise numerical values derived here depend on the Luce choice parameter α, here set to 4. The qualitative preference for *ALL* does not depend on this parameter, though. For example, with $\alpha = 10$, the dispreference for *ALL* over some when $n = 6$ would not be on the order of $.001/.999$, but a much stronger $[1.7 \times 10^{-8}]/[1 - 1.7 \times 10^{-8}]$—almost indistinguishable from a deterministic preference for the maximum utility choice. With $\alpha = 1$, the dispreference for *SOME* has the more moderate value $.16/.84$. Ultimately we can only fit this model parameter to data, or attempt to manipulate it qualitatively in an experimental setting.

much more information about the true state than a literal listener could, using the observation that the utterance chosen was *SOME*.

$$
\begin{aligned}
P_{L_1}(n=6|u=SOME) &= \frac{P_{S_1}(u=SOME|n=6)\times P_{L_1}(n=6)}{\sum_{n'\in\{0,\ldots,6\}} P_{S_1}(u=SOME|n')\times P_{L_1}(n')} \\
&\approx \frac{.001\times.01}{0+.01+.01+.01+.01+.01+.001\times.01} \\
&\approx .015
\end{aligned} \tag{20}
$$

L_1's posterior probabilities for the other states (1-5) are much higher, around .197 each.

Upshot. The posterior probability of $n = 6$, given the observed utterance *SOME*, is much lower for the pragmatic listener L_1 than it is for a literal listener. The difference is driven by the fact that the pragmatic listener considers not only what the speaker *actually* chose to say, but also other things that the speaker *could have chosen*. The pragmatic listener reasons about the latent causes of the speaker's observed choice, using a model which predicts what choices would likely have been observed given various possible configurations of the latent causes (states of the world). This type of reasoning derives the intuitive inference that n is probably not six when *SOME* is used — i.e., that *some* strongly implicates *not all*.

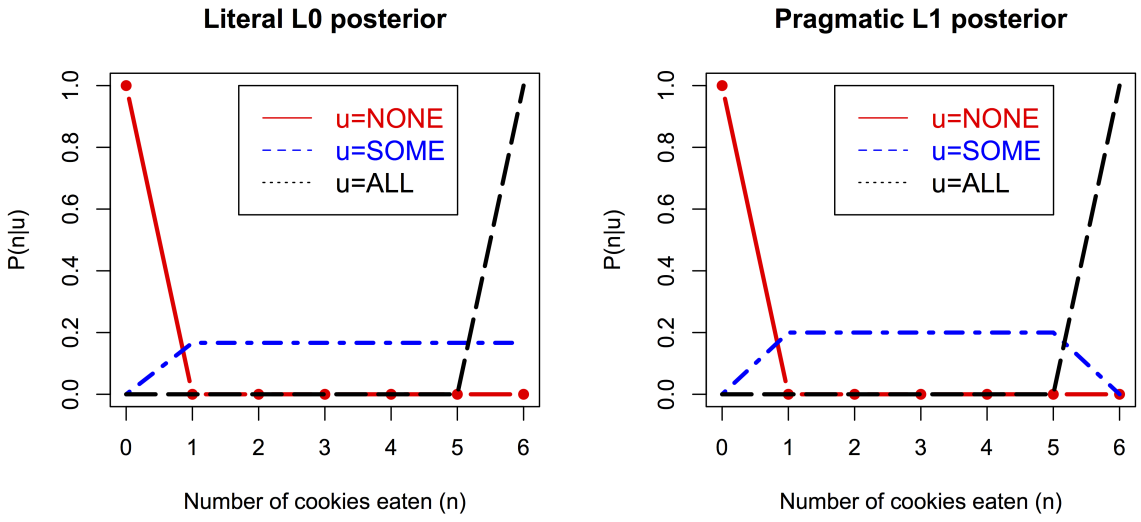

Figure 3 Posterior distributions of the literal (left) and pragmatic (right) listeners, given the various possible utterances under consideration. When $u = SOME$, the pragmatic listener uses the speaker's utterance preferences (Figure 4) to draw a strong "not all" inference.

Figures 3 and 4 summarize the results just described. Note that the difference between the L_0 and L_1 posteriors in Figure 3 lies entirely in $P(n|u = SOME)$: the probability that $n = 6$ is much lower at L_1, with the leftover probability divided evenly among the other true states 1-5. This difference is driven by the fact that L_1's reasoning about the actual state n is "fed through" the speaker model in Figure 4, in which there is a large difference in the probability of using *SOME* vs. *ALL* in this state.

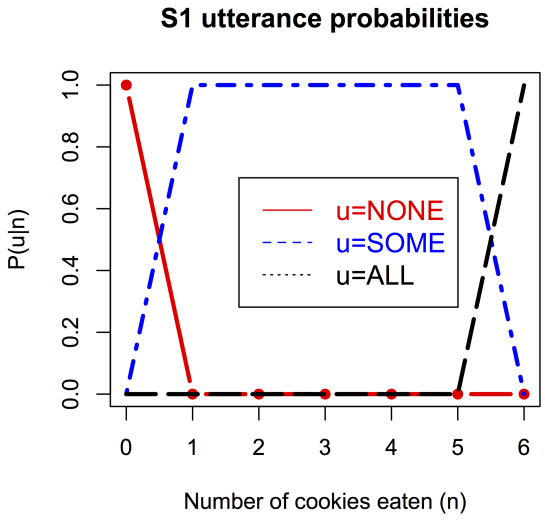

Figure 4 Speaker probability of selecting utterance u, given that the number of cookies eaten is n, for various u and n. The *SOME*-related difference between the listener models in Figure 3 is driven by the fact that $P(u = SOME|n = 6)$ is almost, but not quite, zero.

4 Application to scalar adjectives

Unlike quantity implicatures triggered by the use of quantifier phrases, the interpretation of expressions containing scalar adjectives crucially requires a listener to fill in a free threshold variable which is left unspecified in the adjective's semantic interpretation. Here we describe briefly the semantic motivation for this claim, and then describe an expanded pragmatic model in which the pragmatic listener estimates the values of free variables in much the same way that she estimated the underlying world state in the scalar implicature example just reviewed. The final interpretation reflects a balance between two countervailing pressures: the listener's preference for interpretations which are likely to be true, and the speaker's preference for interpretations that are informative. In our model, this balancing process is responsible for the extreme sensitivity of scalar adjectives' contextual meanings to statistical priors.

4.1 Semantic background

We adopt a degree semantics in which scalar adjectives relate individuals to a threshold value, schematically:

(21) $[\![A]\!] = \lambda\theta_A \lambda x[\mu_A(x) > \theta_A]$

θ_A is a degree on A's scale—the **threshold**—and $\mu_A(x)$ is the measure of x on this scale. For example, if A is *tall*, then the scale is the range of possible heights $(0, \infty)$, and μ_A is a function mapping objects to values in this range. The adjective denotes a function which receives a threshold argument (say, 3 feet) and an individual argument, returning true if the associated measure function

returns a value which exceeds the threshold when applied to the argument.[7]

We assume that the lexical entry of a gradable adjective contains two components: a specification of the relevant scale — an ordered set of degrees — and an indication of the adjective's polarity along that scale. That is, antonym pairs such as *tall/short* and *dangerous/safe* live on scales which are identical except that the ordering is reversed. Since we are reasoning only about positive-form adjectives here, we can simplify by maintaining the scale's intrinsic ordering but reversing the direction of the comparison.

(22) $[\![tall]\!] = \lambda\theta_{tall}\lambda x[\mu_{height}(x) > \theta_{tall}]$

(23) $[\![short]\!] = \lambda\theta_{short}\lambda x[\mu_{height}(x) < \theta_{short}]$

Note that we are assuming that *tall* and *short* have unrelated thresholds, and so that their meanings are independent except for their shared but inverted scales.[8]

In formal semantics the threshold is usually treated as an argument rather than a contextual parameter because this choice makes it simple to account for uses in which the threshold is semantically rather than pragmatically controlled, such as *two feet tall* or *taller than Charlie is*. However, the result of its presence in (21)-(23) is that we cannot directly compose *tall* and *Al* in order to form a sentence such as *Al is tall*. A fairly standard solution to this problem is to posit a silent morpheme *POS* which binds the value of θ_{tall} to a contextual parameter $\mathbf{s}_{tall}$ (e.g., von Stechow 1984; Kennedy 2007). For this approach we need to assume that each adjective *A* is supplied with a dedicated contextual parameter $\mathbf{s}_A$, or some variable-assigning function which determines one.

(24) a. $[\![tall]\!]^{\mathbf{s}_{tall},\mathbf{s}_{big},\mathbf{s}_{heavy},\cdots} = \lambda\theta_{tall}\lambda x[\mu_{tall}(x) > \theta_{tall}]$

b. $[\![POS]\!]^{\mathbf{s}_{tall},\mathbf{s}_{big},\mathbf{s}_{heavy},\cdots} = \lambda A\lambda x[A(\mathbf{s}_A)(x)]$

c. $[\![POS\ tall]\!]^{\mathbf{s}_{tall},\mathbf{s}_{big},\mathbf{s}_{heavy},\cdots} = \lambda x[\mu_{tall}(x) > \mathbf{s}_{tall}]$

d. $[\![Al\ is\ POS\ tall]\!]^{\mathbf{s}_{tall},\mathbf{s}_{big},\mathbf{s}_{heavy},\cdots} = \mu_{tall}(\mathbf{Al}) > \mathbf{s}_{tall}$

The result is that *Al is tall* is true iff Al's height is greater than $\mathbf{s}_{tall}$, whatever that is.

An alternative account is to treat *POS* as a type-shifter which reverses the order of the arguments, as in (25). (Additional type-shifters would be needed to pass up the unsaturated variable in non-predicative uses and non-matrix contexts; the intended account is a generalization of the treatment of free anaphors in variable-free semantics (Jacobson 1999).)

(25) a. $[\![tall]\!] = \lambda\theta_{tall}\lambda x[\mu_{tall}(x) > \theta_{tall}]$

b. $[\![POS]\!] = \lambda A\lambda x\lambda\theta_A[A(\theta_A)(x)]$

7 This is a $\langle d,\langle e,t\rangle\rangle$ semantics for adjectives in the style of von Stechow (1984). For current purposes, it does not matter whether we use this analysis or an $\langle e,d\rangle$ treatment as recommended by Bartsch & Vennemann (1973); Kennedy (1997, 2007). The only modification needed would be in the definition of the POS morpheme/type-shifter.

8 In the context of our pragmatic model, assuming two unrelated thresholds for *tall* and *short* is enough to derive reasonable interpretations, as we will see below; it even derives the fact that *tall* and *short* are contraries without stipulation. We could, if we wanted, add a lexical stipulation to this effect ($\theta_{tall} \geq \theta_{short}$), but we do not know of any compelling reason to do so. It would also be possible to assume that these expressions are contradictories ($\theta_{tall} = \theta_{short}$, so that *short* ≡ *not tall*). Whether the latter assumption is reasonable is a matter of debate: see Horn 1989; Heim 2006, 2008; Büring 2007a,b among many others for arguments pro and con.

c. $[\![POS\ tall]\!] = \lambda x \lambda\, \theta_{tall}[\mu_{tall}(x) > \theta_{tall}]$
d. $[\![Al\ is\ POS\ tall]\!] = \lambda\, \theta_{tall}[\mu_{tall}(\mathbf{Al}) > \theta_{tall}]$

These accounts are really not very different: both provide, in compositional fashion, a sentence meaning which does not determine a truth-value until the value of a certain free/unsaturated variable is determined. In either case, pragmatic inference is required to determine which proposition the sentence expresses. Our approach could also be modified for theories which countenance degrees only in the metalanguage (Lewis 1970; Barker 2002) or not at all (Klein 1980), but we will not spell out the details here. (Indeed, we believe that the depth of the differences between these approaches has been exaggerated in the literature: see Klein 1991; Lassiter 2015 for relevant discussion.)

4.2 Bayesian inference of free variables

Information from the compositional semantics—such as the semantics for adjectives just described—enters into our pragmatic model via equation (9), which instructs the literal listener to condition on the truth of the observed utterance.

(26) $$P_{L_0}(A|u) = P_{L_0}(A|[\![u]\!] = 1)$$

But conditioning is defined only for propositions, and so this model is only appropriate if $[\![u]\!]$ contains no free variables. That is, L_0 cannot condition on the truth of an utterance like *He is in Paris* without first filling in a referent for *he*. Similarly, L_0 cannot condition on the truth of *Al is tall* without first filling in the threshold value θ_{tall} which specifies how tall one must be in order to count as *tall*. Once this is done, for whatever value of θ_{tall} is supplied, L_0 can condition on the information that Al's height is greater than θ_{tall}.

More generally, if u's meaning does make reference to some set of free variables V, the listener must find some way to infer values for these variables. We propose to estimate these values by having the pragmatic listener consider as possible interpretations all possible literal meanings that could be expressed by an utterance, relative to all possible assignments of values to variables. That is, our pragmatic listener will now instantiate assignments of values V from a prior distribution $P_{L_1}(V)$ and thread them through the iterated reasoning procedure, considering how likely it is that the speaker would have produced the utterance that he did, *if the variables were resolved as they are in V*.[9]

Let V be a function which assigns values to all variables in the language, and let $[\![u]\!]^V$ be the language's interpretation function as parametrized by V. The literal listener is as in the simple model, except that he conditions on $[\![u]\!]^V$, given a value of V which is provided by the speaker

9 We do not, of course, want to claim that listeners making such inferences in real time actually consider an infinite set of $\langle A, V\rangle$ pairs as hypotheses. The problem is, however, a very general one about how humans make inferences with a very large hypothesis space—something that we are surprisingly good at. The model presented in this paper is a high-level computational theory for which a variety of techniques are available that can make inference tractable: this includes lazy computation (reason only about variables that are actually present in the utterance and need to be inferred) and Markov Chain Monte Carlo techniques such as the one that we use below to simulate Bayesian posteriors. For further discussion see, for example, Griffiths, Vul & Sanborn 2012; Vul et al. 2014; Goodman & Tenenbaum electronic.

model.

$$P_{L_0}(A|u,V) = P_{L_0}(A|[\![u]\!]^V = 1) \tag{27}$$

The variable-sensitive S_1 model also takes a given value of V and produces utterances stochastically, on the assumption that the variables are valued as they are in V.

$$P_{S_1}(u|A,V) \propto \exp(\alpha \times \ln\left[P_{L_0}(A|u,V) - C(u)\right]) \tag{28}$$

The pragmatic listener then derives a variable-sensitive interpretation by considering how likely it is that the speaker would have said u if the answer were A *and* the variables were as in V — and, as usual, multiplying this value by the prior probabilities of A and V. This gives us a function which assigns a joint posterior probability to all possible combinations of A and V.[10]

$$P_{L_1}(A,V|u) \propto P_{S_1}(u|A,V) \times P_{L_1}(A) \times P_{L_1}(V) \tag{29}$$

For example, if the utterance is "Al is tall" V will determine a value for θ_{tall} and no other relevant variable. L_1's pragmatic interpretation will thus proceed by considering, for $\theta_{tall} \in (0,\infty)$ and heights $h \in (0,\infty)$, how likely it is that the speaker would have said this if "tall" were interpreted as meaning "taller than θ_{tall}". (Note, by the way, that this is not a claim about the algorithm by which this computation is implemented: a well-designed algorithm for computing this inference will not waste effort considering heights and values of θ_{tall} which are so far beyond the normal range of heights for the class of objects in question that they cannot possibly be relevant to the issue at hand. See also footnote 9.)

The prior term $P_{L_1}(V)$ specifies L_1's background knowledge about the interpretation of free variables. We will make the assumption that the listener has no relevant background knowledge about the resolution of free variables, and so no reason to favor any choice of V. $P_{L_{0/1}}(V)$ is thus uniform for all possible combinations of values for the elements of V, and the $P_{L_n}(V)$ terms drop out of the reasoning everywhere.[11] The assumption of uniform priors on semantic variables means that, if u = *Al is tall*, all possible thresholds for *tall* are equally good candidates *a priori*; we do not, for example, build in a preference for interpretations which are statistically more frequent in uses of *tall*. This assumption seems to be justified at least in the case of scalar adjectives, where non-uniform priors would limit the flexibility of interpretation: in other words, if $P_{L_1}(\theta_{tall})$ were strongly biased toward human-like heights, this bias would influence the interpretation of *tall skyscraper* in strange ways. We eliminate this possibility by employing uniform priors. However, it is an empirical question whether this assumption holds for all types of free variables that occur in natural languages.

10 We are assuming that A and V are independent for the listener in the prior, and are related only via the interpretation process.

Note that we could continue to iterate to some higher L_n and pass the variable up. It is also possible to marginalize at non-maximal levels. We have found that, under many conditions, scalar adjectives receive implausibly weak interpretations if marginalization happens at L_0 (Goodman & Lassiter 2015). However, many other possibilities remain to be explored, e.g., a level-3 pragmatic intepreter with level-1 or -2 marginalization.

11 Note that $P_{L_{0/1}}(V)$ is an improper prior if the range of V extends to infinity in either direction.

4.3 Scalar adjective interpretation: The intuition

To get an intuition about how the various model components play into scalar adjective interpretation, we will consider in detail several sample interpretations against the following background: u is "Al is tall", we know nothing about Al except that he is an adult man, and **ALT** = $\{$*Al is tall, Al is short*, $\varnothing\}$.

- Consider very large values of θ_{tall}, for instance 7 feet. In this case the probability of the interpretation is low because of the prior on heights: it is very unlikely that Al is more than 7 feet tall, and our model encodes an assumption that speakers do not make false utterances. This fact leads to a low probability of this interpretation *even though*, if Al were in fact taller than 7 feet, the utterance would be extremely informative; in this case, the strength of the dispreference generated by u's low prior probability of truth (low $P(h > 7\,feet)$ outweighs the informativity preference, leading to a low posterior density of $\theta_{tall} = 7\,feet$.
- Consider very low values of θ_{tall}, for instance 1 foot. In this case the interpretation receives very low probability because of the speaker's preference for informative utterances. Specifically, the probability that "Al is tall" is true relative to this *tall*-threshold is effectively 1: all adult men are more than 1 foot tall. However, the probability that the speaker will produce this utterance is very low, because conditioning on the truth of a known proposition does not influence one's probability distribution. Thus the posterior distribution $P_{L_0}(h|u = \text{"Al is tall"}, \theta_{tall} = 1\,foot)$ is effectively the same as the prior, and no information is conveyed. In this case, the speaker will prefer to say nothing, since this is at least a cost-free way to convey no information. In other words, the speaker's observed choice to utter "Al is tall" can only be rationalized if it increases the probability of the true answer considerably, relative to what the prior probability of this answer is, and low values of θ_{tall} do not meet this requirement.
- Suppose θ_{tall} has some intermediate value, say, 6 feet. Then we have a reasonable compromise: saying "Al is tall" will convey a reasonable amount of information, but the prior probability that Al's height is greater than 6 feet is not so low that the pragmatic listener will discount it as probably false.

More generally, the joint posterior on interpretations and answers to the QUD that this model derives reflects a balancing process between the speaker's informativity preference and the listener's beliefs about which utterances would be true. This results in a probabilistic "sweet spot" interpretation for scalar adjectives which is highly sensitive to the statistical information encoded in the prior. Very weak interpretations, with θ_{tall} falling in the lower region of the height prior, are probably true; however, the informativity preference entails that speaker would probably not have chosen to use the utterance in such a situation. Conversely, very strong interpretations (with θ_{tall} in the extreme upper tail of the height prior) are dispreferred because they make the utterance very likely to be false — even though they would be extremely informative if true. The effect is a preference for interpretations which make Al fairly tall, but not implausibly so.

4.4 Scalar adjective interpretation: Simulations

The simulated marginal posteriors of $h = height(\textbf{Al})$ and θ_{tall}, given the utterance $u =$ "Al is tall", are plotted in Figure 5. This and all following simulations use Metropolis-Hastings, a Markov Chain Monte Carlo algorithm (Neal 1993; MacKay 2003), to approximate the posterior since it is not possible to solve the model analytically. As above, we set $\alpha = 4$ and $C(u) = 2/3 \times length(u)$, with length measured in number of words: for example, $C(Sam\ is\ tall) = 2$. The action $\varnothing$ (saying nothing) has cost 0. We use $\textbf{ALT} = \{ A_{pos}, A_{neg}, \varnothing \}$, where A_{pos} and A_{neg} are an antonym pair such as *tall/short*.

All plots below show the marginal kernel densities of A and θ_{tall} in our samples, along with the input priors $P_{L_{0/1}}(h)$ and $P_{L_{0/1}}(\theta_{tall})$.[12] These are approximations to the the marginal posterior density $P_{L_1}(\theta_{tall}|u)$ and $P_{L_1}(h|u)$, as described by equations 30 and 31 respectively.

$$P_{L_1}(\theta_{tall}|u) = \int_0^\infty P_{L_1}(\theta_{tall},\ h|u = \text{"Al is tall"})\ dh \tag{30}$$

$$P_{L_1}(h|u) = \int_0^\infty P_{L_1}(\theta_{tall},\ h|u = \text{"Al is tall"})\ d\theta_{tall} \tag{31}$$

We assume that heights (e.g., the heights of adult men) are approximately normally distributed: $P(h) = N(\mu, \sigma)$ for a specified mean μ and standard deviation σ.[13] The choice of μ and σ matter only for scale: multiplying these parameters by a constant has the effect of multiplying posteriors by the same constant (cf. Figure 7).

The inferred meaning of "tall" is essentially "significantly greater than average height", an intuition which has been expressed in the literature in various forms (e.g., Fara 2000; Kennedy 2007). Crucially, though, the resulting interpretation remains vague: the interpretation process takes us from knowing nothing about the contextual meaning of "tall" to knowing quite a lot, but there is still remnant uncertainty in $P_{L_1}(\theta_{tall}|u)$.

Since the lexical entries of *tall* and *short* in (22) and (23) differ only in the direction of the comparison with their estimated threshold, and since the prior is symmetric, we should expect the interpretation of *Al is short* to be symmetric along the prior mean. This is indeed the case: see Figure 6.

These simulations shed light on several of the explananda discussed in section 1. There we briefly described four puzzles: information transmission despite uncertainty about interpretation, context-dependence, borderline cases, and the sorites. This model accounts for the possibility of

12 The *kernel density* is a nonparametric estimate of the density of a continuous function from a finite number of samples, given certain assumptions about the smoothness of the function. See, for example, Silverman 1986.

Simulations took 5,000,000 samples from $P_{L_1}(h, \theta_{tall}|\ u)$ with a burn-in of 5000 (i.e., the first 5000 samples were discarded because to avoid dependence on the starting point of the simulation). Unlike Lassiter & Goodman (2013) we do not rescale variables to fall within $[0, 1]$, since we are not concerned with the difference between bounded and unbounded scales here.

13 "Approximately" because the normal distribution has support over the entire real line, but it does not make sense to talk about heights less than zero. However, for the priors that we will consider, only a negligible portion of the prior probability mass falls below zero.

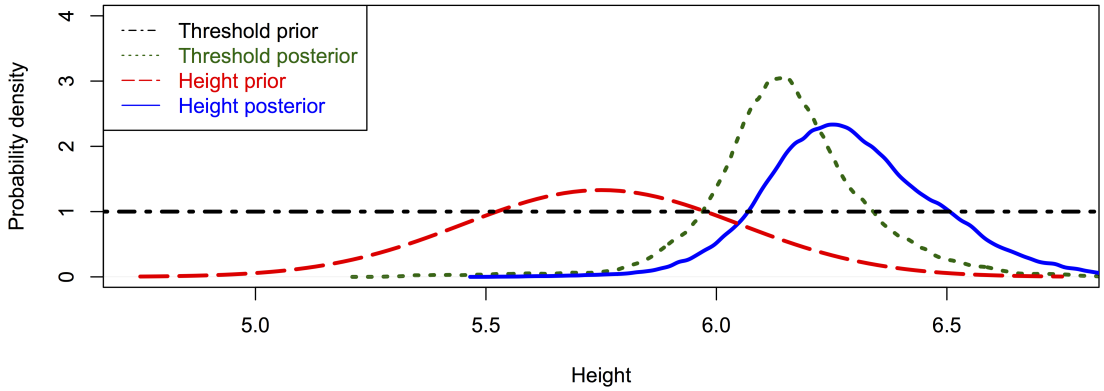

Figure 5

Simulated posterior given utterance "Al is tall", plotting the marginals of *tall*-thresholds and answers (heights) separately. The model predicts several core properties of relative adjectives: significant information conveyed despite vague meaning, sensitivity to statistical priors, the existence of borderline cases, and a plausible account of the sorites.

information transmission using vague expressions in a precise way. In information theory (Shannon 1948), transmission of information is interpreted as modification of a probability distribution. The interpretation of a vague adjective in our model is a function from priors to posteriors, and clearly significant information has been gained in the interpretation: the posteriors on Al's heights in Figures 5 and 6 are shifted substantially away from the prior, and have lower variance. In other words, even though the meaning of "tall" remains uncertain, a listener can gain significant information about the world when a speaker uses it to describe something.

Second, the context-sensitivity of vague scalar adjectives is predicted because the interpretation process is highly sensitive to the form of the input prior. In Figure 5, for example, the inferred meaning of *tall* is a distribution centered around 1.3 standard deviations above the prior mean. Since the scale in these simulations is arbitrary, the model predicts that a normal prior with a different mean and standard deviation would lead to a qualitatively similar but quantitatively different posterior, differing in mean and variance. Figure 7 illustrates how interpretation is affected when we interpret "tall" relative to two prior distributions with equal variance but different means.

This style of interpretation helps us to understand the sense in which the concept "tall" has a stable meaning, even though its interpretation can vary widely in different contexts (*tall boy*, *tall tree*, *tall building*, etc.). Assuming that heights are normally distributed in each class but differ in mean and standard deviation, the posterior on θ_{tall} and A will be shifted accordingly, while maintaining the same shape relative to the prior mean and standard deviation. Background knowledge, in the form of a statistical prior on answers to the QUD, thus interacts with lexical meaning and the pragmatic preference for informativity to yield a context-sensitive probabilistic meaning.

Third, borderline cases of "tall" are individuals whose probability of counting as "tall" is intermediate. This is itself vague, of course, but that is as it should be given the existence of

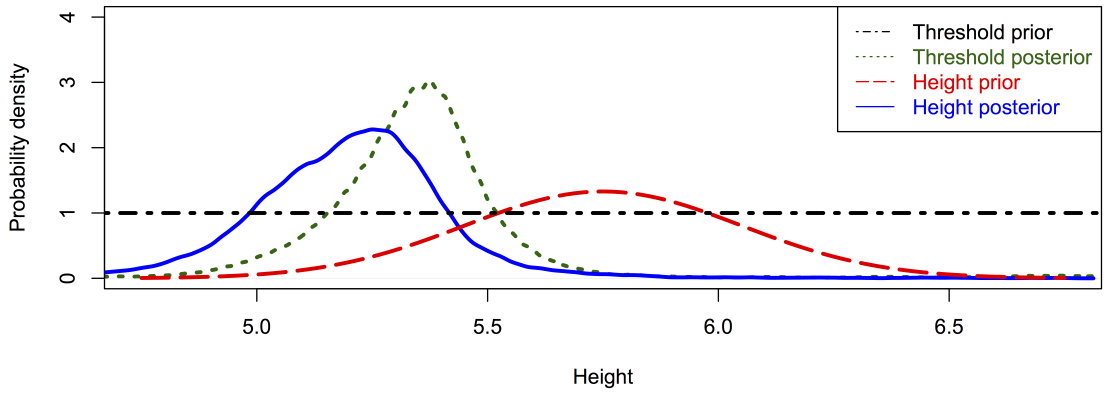

Figure 6 Threshold and degree priors and simulated posteriors given utterance "Al is short".

higher-order vagueness.[14] The key problem, though, it to fix what the probability of counting as "tall" is — or, more precisely, how to interpret "the probability that an individual with height h counts as 'tall"'. We suggest the following interpretation, where P_T is the metalinguistic probability that we, as theorists of vagueness, feel appropriate to assign here:

$$P_T(a \text{ is tall}) = \int_0^{height(a)} P_{L_1}(\theta_{tall} | u \quad = \text{"}a \text{ is tall"}) \, d\theta_{tall}. \tag{32}$$

The intuition behind equation 32 is this. If we know an individual's height and want to know how appropriate it would be to describe him as "tall", we imagine ourselves using "tall" in communicating with a listener with an appropriate prior distribution $P_{L_{0/1}}(h)$. Using the interpretation that this listener would arrive to resolve the underspecified meaning of "tall", we then compute the probability that the utterance is true relative to the context-sensitive posterior on θ_{tall} that this L_1 derives.[15]

In the case at hand (Figure 5), this means finding the area under the curve of the threshold posterior which falls to the left of a's height. If a is 5 feet 9 inches, then a probably does not count as "tall" (specifically, $P_T(a \text{ is tall}) \approx .02$). If a is 6 feet 6 inches tall, then a almost certainly counts as "tall" $P_T(a \text{ is tall}) \approx .97$). But if a is 6 feet 2 inches tall, then a is a near-perfect borderline case — $P_T(a \text{ is tall}) \approx .55$). This all depends on the choice of prior distribution and model parameters, but it seems to yield roughly reasonable results — even though the prior and model parameters were not fitted to empirical data in our simulations.

Crucially, this proposal is intended as a psychological theory of the intuitive strength of sentences involving vague scalar adjectives, and as one which could play a role in a theory of people's (including theorists') metalinguistic judgments about the degree to which a sentences is supported

14 We will not deal with higher-order vagueness in detail here, but two compatible lines of attack suggest themselves. First, we could treat *definitely* as a vague modal, following Lassiter (2011). Second, we could allow uncertainty about the relevant prior distribution, which would generate meta-uncertainty about the interpretation.

15 We call this inference "metalinguistic" because it requires us to reflect directly on linguistic interpretation and usage. Unfortunately for psycholinguists and field linguists, this is a difficult and fairly unusual task for most people: usually inferences about language serve as a means to communicate about the non-linguistic world, rather than as an end in themselves, and we have to figure out some more indirect way to get at them.

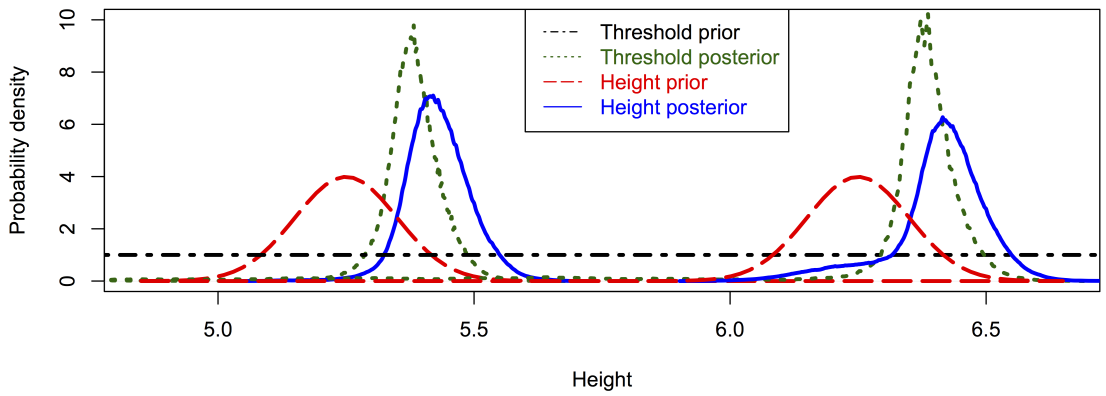

Figure 7 Simulated posteriors given utterance "Al is tall", using two input priors with different means but equal variance.

in a context. It is *not* meant as an answer to the metaphysical question of when a sentence such as "Al is tall" is *in fact* true. We suspect that many of the semantic approaches commonly explored in the literature of vagueness are compatible with this theory. For example, we could convert it into an Edgington-style degree theory by tying the degree of truth of a sentence in a context to the $P_T(\cdot)$ function that an ideal interpreter would derive in this context. This would be rather different from her original conception, since the probabilities that our model derives involve a speaker's intended meaning, rather than a description's truth. However, if the only relevant contextual factor were a speaker's intended meaning, it might be appropriate to use a rational listener's probabilistic estimate of intended meaning to fix degrees of truth. The epistemic theory of vagueness is also clearly compatible with this account: we simply assume that there is a fixed, unknown fact about the true value of θ_{tall}, and both speaker and listener are trying to estimate it probabilistically and use these estimates to inform their usage. Interpretations in terms of supervaluations or three-valued logic also seem to be available. We will not take a stand on which of these interpretations is best because our interest is primarily in the psychology and behavior of language users, and it is unclear to what extent an answer to the metaphysical question would be illuminating about these cognitive issues.[16]

Fourth, there is the sorites paradox. The next section describes several possible approaches and our model's predictions for each.

The model of pragmatic interpretation sketched here, with threshold values and other semantic variables passed up to the pragmatic listener, interacts with our simple free-variable semantics to

16 A cognitive theory of the interpretation of vague expressions does not answer the metaphysical question, but it does perhaps sharpen the subject matter of this question. If we had a satisfying probabilistic theory of how listeners interpret vague expressions and what information they extract from them, as well as an account of why and when speakers choose to use vague expressions, what phenomena would remain to be explained by an answer to the metaphysical question?

predict several of the key properties of relative adjectives in the positive form. Crucially, the model's predictions vary depending on the prior on answers to the QUD, $P_{L_{0/1}}(h)$. While a normal prior is natural for, e.g., the distribution of heights and temperatures in a natural class, there are other properties for which this is not a reasonable assumption (for example, the properties associated with *dangerous/safe* and *full/empty*). In Lassiter & Goodman 2013 we explore the predictions of this model for a broader range of priors. We propose there that the differences in posterior derived by our model, depending on prior, can account for several phenomena discussed in the literature on absolute scalar adjectives including the latter's apparent lack of borderline cases and insensitivity to the sorites paradox (Kennedy 2007).

5 The sorites

In this section we explore several possible psychological accounts of the sorites paradox that are made available by the model of interpretation that we have described. The first two follow closely ideas in Edgington 1992, 1997, and focus on the marginal posterior distribution on thresholds.[17] The third suggested interpretation is new here, and makes use of the fact that our model delivers a joint distribution on *tall*-thresholds and on the heights of individuals described as "tall". Using the Bayesian model of adjective interpretation together with some auxiliary assumptions about how argument strength tracks probability, we can extract from each a (non-paradoxical) explanation of why the sorites feels compelling, as well as specific quantitative predictions about how compelling the premises should be depending on the set-up of the sorites sequence and our semantic assumptions.

We will not come to a firm conclusion about how the sorites should be understood, since there are many options and little empirical data available on which to base a conclusion. However, the fact that we can assign quantitative strengths to the premises of the sorites suggests a reorientation toward precise, experimentally testable theories of the sorites, as we describe briefly at the end of the section.

We set up a sorites sequence as follows. The first member x_n is an individual who is clearly *Adj*, and the last, x_0, is an individual who is clearly not *Adj*. Each non-initial member of the sequence has a degree of the scalar property underlying *Adj* which is exactly ε greater than the previous member. The general form of the sorites paradox is as in (33). (34) illustrates with some familiar characters, the stock adjective *tall*, and ε = 1 mm.

(33) a. x_n is *Adj*.
b. For all m s.t. $0 < m \leq n$, if x_m is *Adj*, then x_{m-1} is *Adj*.
c. $\therefore x_0$ is *Adj*.

(34) a. André the Giant is tall.

17 Our discussion follows Edgington's seminal work most closely. For other related work connecting probability with philosophical and linguistic questions involving vagueness, see Borel 1907 (with translation and commentary in Égré & Barberousse 2014); Black 1937; Kyburg & Schubert 1993; Kyburg 2000; Lawry 2008; Frazee & Beaver 2010; Lassiter 2011; Égré 2011; Sutton 2013; Egré to appear. A detailed comparative analysis of these accounts would take us too far afield here; see however Égré & Barberousse 2014; Egré to appear for some discussion along these lines.

b. For any two people, if the first is tall and the second is 1 mm. shorter, then the second is also tall.

c. ∴ Danny DeVito is tall.

The argument is valid but its conclusion is absurd. The problem is to explain how the first premise can be true and the conclusion false, while doing justice to the intuition that the inductive premise is intuitively compelling in concrete examples such as (34b).

Edgington (1997) argues that the basic form of the sorites — the form in which people reason intuitively about it and find the premises compelling — is not the highly compressed universally-quantified version in (33) but the equivalent, more expansive version in (35). (For a related suggestion see Sorensen 2012.)

(35) a. x_n is tall.

b. If x_n is tall, then x_{n-1} is tall.

c. If x_{n-1} is tall, then x_{n-2} is tall.

d. If x_{n-2} is tall, then x_{n-3} is tall.

e. ...

f. If x_2 is tall, then x_1 is tall.

g. If x_1 is tall, then x_0 is tall.

h. ∴ x_0 is tall.

Of course, it is crucial in resolving the paradox in this form to fix our understanding of the conditionals. Edgington considers two options, a material conditional interpretation and a probabilistic interpretation, and argues that the probabilistic account is able to dissolve the paradox on either interpretation. We will give our own gloss on this reasoning, which is (we hope) faithful to the spirit of Edgington's account.

Classically, belief is thought of as an all-or-nothing matter. On this interpretation, a valid argument constrains a rational individual who fully believes the premises to fully believe the conclusion as well. But what can we do with deductive validity if we think of belief as a graded, probabilistic concept? Adams (1966) proves that validity constrains probability assignments in the following way:

> If (and only if) an argument is deductively valid, the uncertainty of the consequent (i.e., 1 minus its probability) cannot exceed the sum of the uncertainty of the premises (1 minus their individual probabilities) under any probability distribution $P(\cdot)$.

So a rational Bayesian ought to pay attention to deductively valid arguments like the sorites. However, deductive validity's constraining effects on rational belief are not as extreme when belief admits of degrees: rather, deductively valid arguments merely *tend* to lead to reasonable conclusions and can fail to do so under two kinds of conditions. First: we have to be *fairly confident about the premises*. Second: *there cannot be too many premises*. (How many are "too many" depends on how confident we are about the premises, of course.) If either of these conditions is violated, a deductively valid argument can lead us to erroneous conclusions.

As Edgington points out, a sorites argument like (35) is designed so that — if it is valid — it must violate one of these conditions: either the gap ε is large (so that we are not very confident about the premises), or the number of premises is large (so that probability shades off gradually from 1 to 0). Either way, the air of paradox dissolves once we jettison the assumption that belief is an all-or-nothing matter: we should think of the sorites primarily as a demonstration of the dangers of this assumption.

Structurally, the explanation is identical to a Bayesian account of the Lottery and Preface paradoxes, as Edgington emphasizes. I may believe, for each of the statements *S* in my book, that *S* is true; and yet I also believe that there is a false statement lurking undetected in the book. (As Edgington (1997: 295) puts it: "Who would be so rash as to claim that he has no false beliefs?") On a Bayesian interpretation, this is a consistent set of beliefs if there are many statements in the book and some of them have high but non-maximal probability. Paradox arises only if belief is all-or-nothing (or if there is a fixed context-insensitive probability threshold for belief: see Lin & Kelly 2012; Leitgeb 2014).

Implicit in this explanation is a strong claim about the psychological aspect of the sorites (cf. Fara 2000). Edgington assumes that the reason we find the premises compelling is that they have high (conditional) probability. That is, the intuitive strength of a premise is taken to be directly correlated with its probability, rather than, for example, corresponding to a step function which falls off sharply with the transition from probability 1 to probability less than 1. In full detail, the implicit assumption seems to be:

> Consider a sentence *A* (which may be a conditional sentence). Our intuitions about the strength of *A* do not track the binary question of whether we have "full belief" in *A* — that is, there is not a sharp difference in the felt strength of *A* between the situation in which we are completely certain and a situation in which we are almost, but not quite, certain. Rather, intuitions about the strength of a premise decline gradually with the degree of belief that we have in *A* — that is, with its probability.

Bracketing the issue of how to assign probabilities to conditional sentences (more below), it seems clear that this assumption is very important for the overall plausibility of Edgington's theory. Without it, her explanation would fail to account for the psychological aspect of the sorites — even if it is logically impeccable — since it would not yet give us an explanation of why people find premises (35b)-(35g) compelling. Fortunately, this interpretation of intuitive strength has received a great deal of support in the recent psychological literature on reasoning, which has recently moved away from a binary concept of strength connected with the binary concept of belief. The dominant approach in this literature, as well, is a graded notion of strength which is closely connected to (conditional) probability (Oaksford & Chater 2007; Over 2009).

Taking for granted that our intuitions about premise strength track the premises' probability, then, what remains to be shown is that there is a semantically and pragmatically well-motivated way to assign high probability to each of the premises of (33) while also assigning low probability to the conclusion. Take first the material conditional interpretation of conditionals (MC). On this interpretation the paradox is logically equivalent to the expanded version of the negative existential sorites, viz.:

(36) a. x_n is tall.
b. It's not that case that x_n is tall and x_{n-1} is not tall.
c. ...
d. It's not that case that x_1 is tall and x_0 is not tall.
e. $\therefore x_0$ is tall.

We can use equation (32) to find the probability that an individual x_m is tall while x_{m-1} is not: it is just the probability that θ_{tall} falls between $height(x_m)$ and $height(x_{m-1})$.

$$(37) \qquad P_T(x_m \text{ is tall and } x_{m-1} \text{ is not tall}) = \int_{height(x_{m-1})}^{height(x_m)} P_{L_1}(\theta_{tall} | u = \text{“}x_m \text{ is tall”})\, d\theta_{tall}.$$

The probability of the negation of this statement is 1 minus this value. Since they are logically equivalent, this is also the probability of the material conditional interpretations of premises (36b)-(36d).[18] In general, if the gap ε is small then this probability will also be small. This is clear if we imagine dividing the posterior on θ_{tall} in Figure 5 into k rectangles of equal width along the x-axis. When k is large, very little of the posterior mass of θ_{tall} will reside in any single rectangle. Figure 8 shows the result with ε set to the generous value of .5 inch.

Figure 9 plots the probabilities of the MC-sorites premises in (35) using the probabilistic interpretation that was plotted in Figure 5, with $\varepsilon = .5$ inch. We start with an individual x_n who is clearly tall (6 feet 9 inches) and end with an individual x_0 who is clearly not tall (4 feet 9 inches). In our simulation, the probability that x_n is tall is indistinguishable from 1 (solid red line on the right edge of the graph in Figure 9), and the probability of the other premises never drops below .97 (dashed green line along top). So, the premises are felt as compelling. Nevertheless, the conclusion of the argument has probability indistinguishable from 0 (solid red line on the left edge of the graph in Figure 9). If this is the indeed the way that people understand the sorites, this account — which follows Edgington's closely — appears to account for the logical and psychological aspects of the paradox.

What about the probability conditional interpretation (PC)? On the version of this account that Edgington presumably has in mind, conditionals do not have truth-conditions, but they are assigned probabilities according to Adams' Thesis (a.k.a. "The Equation": see Adams 1975; Edgington 1995 for explication).

$$(38) \qquad P(\text{If } A \text{ then } B) = P(B|A).$$

On this theory, the probability of *If A then B* can be interpreted roughly as the degree of belief that an individual whose degrees of belief are as in $P(\cdot)$ would assign to B on the supposition that A is

18 We simplify here by assuming that the value of θ_{tall} is fixed for all instances of "tall" in a single sentence whose arguments are drawn from a common prior. This entails that "y is tall and z is not" has probability 0 when y and z have a common prior and z is taller than y. The account of the sorites is not much different qualitatively if we relax this assumption, allowing that θ_{tall} is drawn independently for the two instances of θ_{tall}.

Note also that, given the assumption of common prior and θ_{tall}, it doesn't matter whether the utterance u referred to in equation 32 is "x_m is tall" or "x_{m-1} is tall". On the Edgington-inspired account of the sorites that we are considering, the prior is what drives the interpretation, not the individuals' actual height.

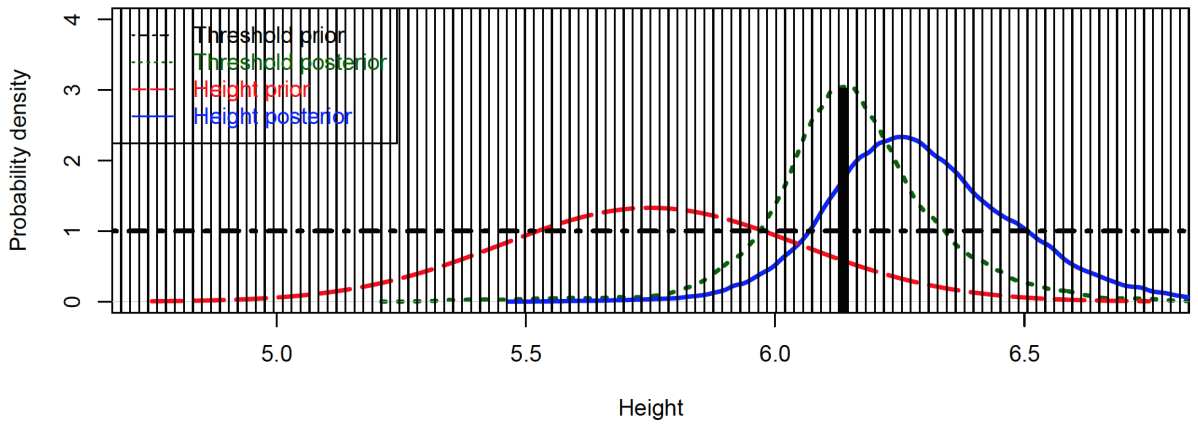

Figure 8

Interpretation of *tall* from Figure 5, with vertical lines overlaid representing the heights of individuals in the sorites sequence. Each premise of the MC-sorites is represented by a rectangle formed from two adjacent lines, and its probability is equal to the probability that the *tall*-threshold does not fall inside that rectangle. The filled-in area, with about 3% of the area under the curve of the threshold posterior, gives the failure probability for the least likely premise. This premise still has probability greater than .97.

true. So, in particular, the premises of the PC-sorites will be compelling as long as the following conditions hold:

(39) a. $P(x_n \text{ is tall})$ is high.
b. $P(x_{n-1} \text{ is tall} | x_n \text{ is tall})$ is high.
c. $P(x_{n-2} \text{ is tall} | x_{n-1} \text{ is tall})$ is high.
d. $P(x_{n-3} \text{ is tall} | x_{n-2} \text{ is tall})$ is high.
e. ...
f. $P(x_1 \text{ is tall} | x_2 \text{ is tall})$ is high.
g. $P(x_0 \text{ is tall} | x_1 \text{ is tall})$ is high.

As before, we have rigged things up so that "$P(x_n \text{ is tall})$" is indistinguishable from 1, and $P(x_0 \text{ is tall})$ is indistinguishable from 0.

We can inspect our simulation results to discover the probabilities of the intermediate premises on current assumptions. Building on the proposal above that equation 32 is an appropriate way to understand the metalinguistic judgments that (39) requires, we suggest that premises (39b)-(39g) can be unpacked as in equation 40.

(40) $$P(x_{m-1} \text{ is tall} | x_m \text{ is tall}) = \frac{\int_0^{h(x_{m-1})} P_{L_1}(\theta_{tall} | u = \text{“}x_m \text{ is tall”})\, d\theta_{tall}}{\int_0^{h(x_m)} P_{L_1}(\theta_{tall} | u = \text{“}x_m \text{ is tall”})\, d\theta_{tall}}$$

In the case of the MC-sorites, the strength of a premise involving x_m and x_{m-1} was given by the proportion of the total area under the curve of $P(\theta_{tall}|u)$ which fell either below $h(x_{m-1})$ or above

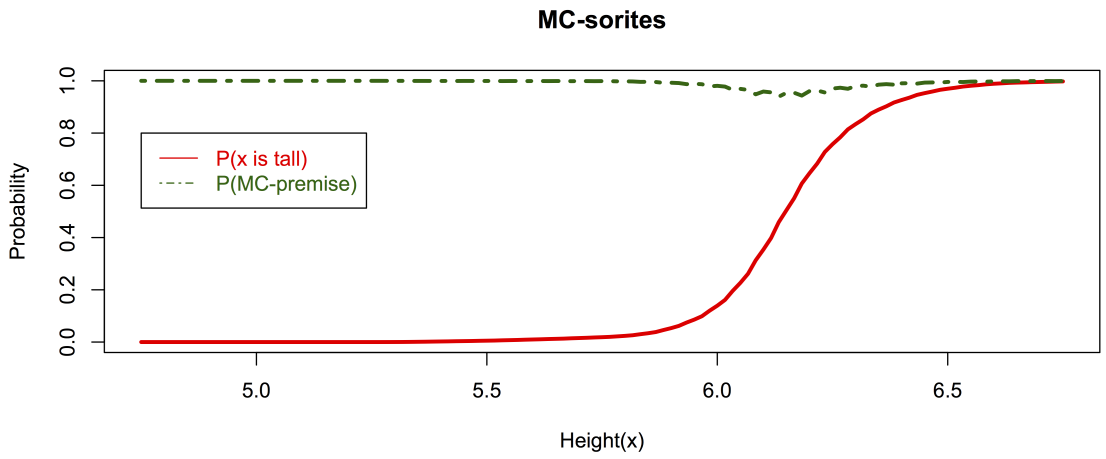

Figure 9 Probability of the premises in the expanded sorites sequence when the conditionals in (35) are interpreted as material conditionals.

$h(x_m)$. In the PC-sorites, the strength of a premise is the proportion of the mass which falls below $h(x_m)$ which does not also fall between the heights of these two individuals. In fact, a given PC-premise is guaranteed to have probability at most as great as the corresponding MC-premise: the PC-sorites is a more stringent test of the probabilistic theory's explanatory power.

Our model's predictions about this version of the sorites are plotted in Figure 10, again assuming a rather large gap size ε = .5 inch.

Initially this version looks less promising: the probability of the inductive premises begins to drop off sharply when the taller of the two individuals has height less than 5.5 feet. However, something subtle is going on here: the premises maintain high probability (> .9) over the entire range in which the antecedent "x_m is tall" has even a tiny chance of being true. The probability of the premise first drops below .9 when x_m is about 5 feet 7 inches tall, and x_{m-1} is 5 feet 6.5 inches. By this point, the probability that x_m is tall is vanishingly small, less than .01. Here, and at lower values, we are reasoning about conditionals whose antecedents are almost certainly false. (Indeed, the curve ends abruptly soon after because we were unable to estimate the probability: out of 5 million samples, there were none in our simulations in which individuals this short counted as "tall".) Perhaps this property can be used to explain away the low probability of the PC-premises at very low heights, on the grounds that these cases are too remote to play a role in our intuitive reasoning about the sorites.

Of course, these are just two of the many possible interpretations of the conditional available. There are modal, non-probabilistic, non-material-conditional interpretations, and there are probabilistic interpretations that do not satisfy Adams' Thesis. While we cannot survey the space of possible theories exhaustively, we are optimistic that the success of the probabilistic account of the sorites in these two, very different theories of the conditional is indicative of its robustness to modifications in the details of conditional semantics.

We assumed above that the function of interest for the interpretation of the sorites was the

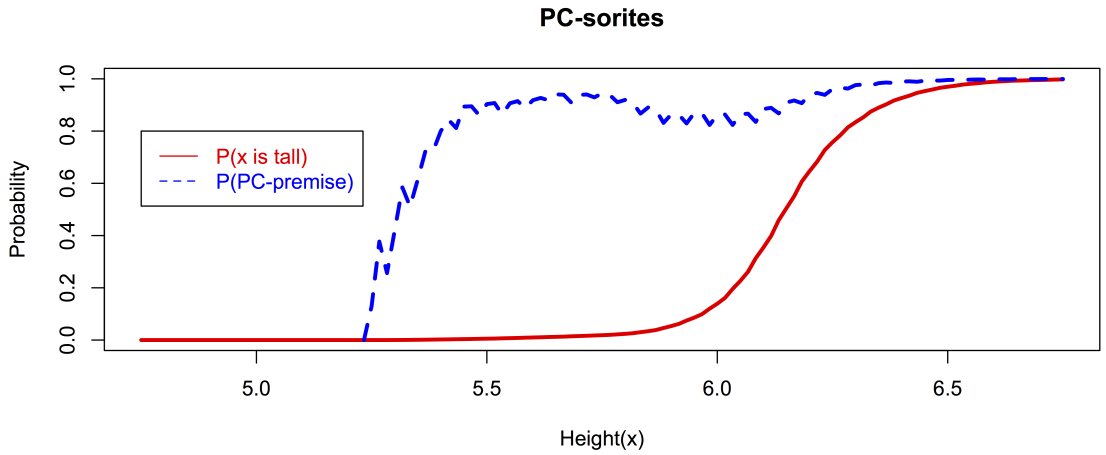

Figure 10 Probability of the premises in the expanded sorites sequence when the conditionals in (35) are assigned probabilities by Adams' Thesis (equation 38).

metalinguistic probability $P_T(x_i \text{ is tall})$, computed using the marginal probability that the threshold falls below x_i's height. However, the model has access to the joint posterior $P_{L_1}(h, \theta_{tall}|u)$, which includes information not only about what "tall" means but also about the likely heights of people who have been described as "tall", relative to a certain prior. Our model makes room for an additional possible interpretation of the sorites which draws on this joint posterior, which we describe briefly in closing.

A natural alternative framing of the sorites is to use neither explicit quantification nor a long list of fully resolved premises, but rather a free variable in the statement.

(41) a. x_n is tall.
b. If y is tall and z is ε shorter than y, then z is tall.
c. $\therefore x_0$ is tall.

Standardly, a statement like (41b) would be understood as implicitly universally quantified. However, our probabilistic account suggests a different interpretation, related to the widely held (but not uncontroversial) claim that indefinites are interpreted as variables of some sort whose values must be inferred using global pragmatic mechanisms (Kamp 1981; Heim 1982; Kratzer 1998, etc.). If this is right, then (41b) is interpreted along the lines of (42):

(42) If someone is tall and someone else is ε shorter, then the latter person is also tall.

On this account, we can keep the variables in the inductive premise. The conditional is understood as an instruction to suppose not just that the antecedent is *true*, but also that it has just been *asserted*: a cooperative speaker S_1 has uttered "y is tall", and the listener is to use this supposition to inform inferences about y's height, the meaning of *tall*, and the probability of the consequent. (41b) should then be intepreted as: what is the probability that a person z, who is just ε shorter than someone who has been described as "tall", also counts as "tall"?

It is worth pausing here to consider the plausibility of this psychological hypothesis in a slightly broader context not involving vagueness. Could language users' instinctive reactions to the conditional premise (41b) by attributed to a general strategy for conditional interpretation? Consider example (43) (inspired by discussion of tolerance principles in Egré to appear).

(43) If x can drink n glasses of wine and still be sober, then x can drink n glasses of wine plus one ounce without becoming ill.

(43) seems like a reasonable claim. Yet, if we universally close over the variables, it is simply false: we only have to set $n = 0$, and consider the small but significant number of individuals who lack the ability to digest alcohol, and become ill with any amount of wine. Nevertheless, (43) strikes us as an intuitively compelling premise—at least as much as (41b). We can understand why (43) feels compelling if we suppose that listeners generate interpretations for conditionals by imagining situations in which the antecedent has been asserted, and considering the truth-value of the consequent in the simulated situation. In this case, our pragmatic model predicts that the listener will tend to consider values for the variables x and n for which a speaker would be likely to comment "x can drink n glasses of wine and still be sober". Since this sentence is uninformative with $n = 0$ — anyone can stay sober by drinking no wine — a speaker would be unlikely to make such an assertion in this case, and so the listener will tend not to consider such scenarios when interpreting the consequent. Instead, listeners will tend to consider values of n which are significant enough to be worth comment. Furthermore, since the base rate of alcohol intolerance is extremely low, these individuals will tend not to be considered when interpreting the consequent. A metalinguistic, simulation-based account of conditional interpretation along these lines would thus be able to predict the intuitive strength of (43).

Continuing with this hypothesis about the interpretation of conditionals containing free variables, we can derive a prediction about the intuitive strength of the version of the sorites in (41). The fact that an individual y has been described as "tall" places constraints on y's height: y is probably not 5 foot 10, even though this is a perfectly common height, because someone of that height probably wouldn't be called "tall". When the listener samples thresholds and heights from the *joint distribution* $P_{L_1}(height(y), \theta_{tall} | u = \text{“}y \text{ is tall”})$ that our model provides, she will find that the height of an individual just ε shorter than y nearly always falls above θ_{tall} as well.

$$
\begin{aligned}
P_J(\text{Premise (41b)}) &= P_{L_1}(height(z) > \theta_{tall} \mid u = \text{“}y \text{ is tall”}) \\
&= P_{L_1}(height(y) - \varepsilon > \theta_{tall} \mid u = \text{“}y \text{ is tall”}) \\
&= P_{L_1}(\theta_{tall} \notin [height(y) - \varepsilon, height(y)) \mid u = \text{“}y \text{ is tall”}),
\end{aligned} \tag{44}
$$

where the last line follows on the assumption that the speaker is being truthful, so that $height(y) > \theta_{tall}$. This interpretation is rather different from the "metalinguistic" probabilities that were computed above by sampling from the distribution $P_{L_1}(\theta_{tall} | u)$ which results from marginalizing out h. Assuming that the relevant prior is as in Figure 5, we can compute the probability described in equation 44 from our sampled joint posterior by asking in what proportion of samples the difference between the sampled height and θ_{tall} exceeds ε. The probability that (41b) is true on this interpretation, setting $\varepsilon = .5$ inch, is approximately .91.

On this interpretation, then, the inductive premise receives quite high probability even when the gap size ε is fairly large. We can also explore how the probability of the inductive premise varies

as a function of ε. Figure 11 illustrates the results for a grid of values between .01 inches and 4 inches. The probability assigned to the inductive premise is very or fairly high for a wide range of values. In fact, for this choice of prior and model parameters the inductive premise has probability .5 for all gap sizes less than 3.1 inches.

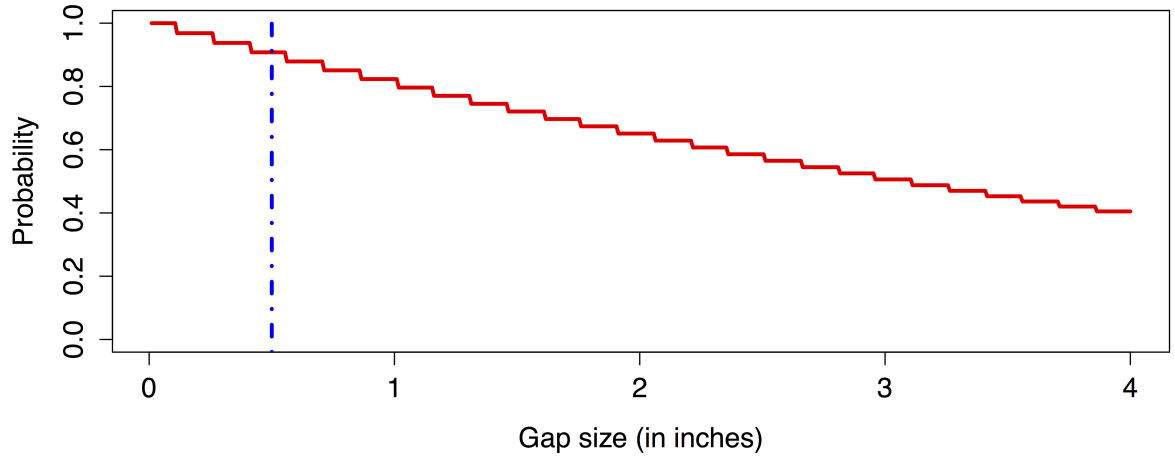

Figure 11

Probability of the inductive premise of the sorites as interpreted in equation 44, as a function of gap size (.01 to 4 inches). The vertical line represents $\varepsilon = .5$, the primary example in the main text, which yields probability .91.

We certainly do not wish to argue that this interpretation is the only way to understand the sorites. However, it does seem to be a reasonable candidate for how people approach the paradox on an intuitive level, and it is not (to our knowledge) one that has been discussed in the previous literature. One particular advantage of this interpretation over the previous two that we considered is that it predicts, quite plausibly, that people will primarily consider relatively tall individuals when considering the inductive premise of a *tall*-sorites. In contrast, for the other two interpretations considered above all instantiations of the inductive premise are on a par, including those for which the antecedent is extremely implausible — for example, when the taller individual is 5 feet tall.

We are not able to decide conclusively among these three interpretations of the sorites, and there are no doubt additional plausible candidates.[19] Especially worth noting here is that our model derives specific quantitative predictions about the strength of the sorites premises which depend on the choice of assumptions about the interpretation of the sorites. These predictions could be tested in behavioral experiments using the methods of experimental psychology. This opens up the interesting possibility that a philosophical theory of the interpretation of the sorites paradox could be subject to experimental verification or falsification — at least, inasmuch as a theory of the sorites is a theory of how language users understand the paradox and why they find it paradoxical.

19 Considerations specific to the theory of conditionals are of great relevance here: for example, the inadequacies of the material conditional interpretation of English *if* are well-known, and we do not know whether the metalinguistic, suppositional account just sketched will ultimately be viable.

6 Conclusion

Our model of adjective interpretation is an application of a general pragmatic model which emphasizes the continuity of language understanding and uncertain reasoning and decision-making in other areas of cognition. Adopting a simple free-variable semantics for adjectives, we have used concepts from game-theoretic pragmatics and tools from Bayesian modeling to predict context-sensitive vague interpretations. The model we proposed derives novel quantitative and qualitative predictions about adjective interpretation and suggests new approaches to a number of crucial issues around the vagueness and context-sensitivity of relative adjectives and especially the sorites paradox.

Some of the choices that we have made in setting up the model are overly simplified, such as the assumption that speakers care *only* about informativity and cost. The interpretive effects of rich speaker goals remain to be explored, as does varying the Question Under Discussion and numerous other possible enrichments and modifications of the model. Likewise, our approach relies crucially on intuitions about reasonable priors; it will be necessary in future work to use empirical measures to validate the choice of priors and check the predictions about their mapping to contextual interpretations. Other natural extensions of the model include inferring the QUD (Kao et al. 2014) and choosing the relevant scale with adjectives for which there are several options (Kennedy 1997; Sassoon 2013).

We hope that this work will be seen as a demonstration of the potential for fruitful interaction between philosophy of language and logic, formal semantics and pragmatics, and computational cognitive science (see also Goodman & Lassiter 2015). Bayesian modeling makes it possible to combine logical and probabilistic reasoning seamlessly, and the recent growth of this style of modeling in cognitive science has opened up new directions which we hope will prove to be useful for linguistics and philosophy as well. Most relevantly for this paper, we believe that this style of modeling has great potential to illuminate the ways in which speakers and listeners use context and background knowledge to communicate rich context-sensitive meanings despite ever-present uncertainty.

References

Adams, Ernest W. 1966. Probability and the logic of conditionals. *Studies in Logic and the Foundations of Mathematics* 43. 265–316.

Adams, Ernest W. 1975. *The logic of conditionals: An application of probability to deductive logic*. Springer.

Bale, Alan C. 2011. Scales and comparison classes. *Natural Language Semantics* 19. 169–190.

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

Bartsch, Renate & Theo Vennemann. 1973. *Semantic Structures: A Study In the Relation Between Semantics and Syntax*. Athenäum.

Benz, Anton, Gerhard Jäger & Robert van Rooij. 2005. *Game Theory and Pragmatics*. Palgrave Macmillan.

Bergen, Leon, Noah D. Goodman & Roger Levy. 2012. That's what she (could have) said: How alternative utterances affect language use. In Naomi Miyake, David Peebles & Richard P.

Cooper (eds.), *Thirty-fourth annual meeting of the cognitive science society*, 120–125. Cognitive Science Society.

Black, Max. 1937. Vagueness. An exercise in logical analysis. *Philosophy of Science* 4(4). 427–455.

Borel, Émile. 1907. Sur un paradoxe économique: Le sophisme du tas de blé et les vérités statistiques. *Revue du Mois* 4. 688–699.

Büring, Daniel. 2007a. Cross-polar nomalies. In Tova Friedman & Masayuki Gibson (eds.), *Semantics and Linguistic Theory (SALT) 17*, CLC Publications.

Büring, Daniel. 2007b. *More* or *less*. In *Proceedings from the annual meeting of the chicago linguistic society*, vol. 43 2, 3–17. Chicago Linguistic Society.

Clark, Herbert H. 1996. *Using language*. Cambridge University Press.

Douven, Igor & Lieven Decock. 2014. What verities may be. Ms., University of Groningen and Vrije Universiteit Amsterdam.

Edgington, Dorothy. 1992. Validity, uncertainty and vagueness. *Analysis* 193–204.

Edgington, Dorothy. 1995. On conditionals. *Mind* 104(414). 235. doi:10.1093/mind/104.414.235.

Edgington, Dorothy. 1997. Vagueness by degrees. In Rosanna Keefe & Peter Smith (eds.), *Vagueness: A reader*, 294–316. MIT Press.

Égré, Paul. 2011. Perceptual ambiguity and the sorites. In Rick Nouwen, Robert van Rooij, Uli Sauerland & Hans-Christian Schmitz (eds.), *Vagueness in communication*, 64–90. Springer.

Egré, Paul. to appear. Vagueness: Why do we believe in tolerance? *Journal of Philosophical Logic* .

Égré, Paul & Anouk Barberousse. 2014. Borel on the heap. *Erkenntnis* 79(5). 1043–1079.

Fara, Delia Graff. 2000. Shifting sands: An interest-relative theory of vagueness. *Philosophical Topics* 20. 45–81.

Fox, Danny & Roni Katzir. 2011. On the characterization of alternatives. *Natural Language Semantics* 19(1). 87–107.

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

Franke, M. 2009. *Signal to act: Game theory in pragmatics*: University of Amsterdam dissertation.

Franke, Michael. 2011. Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics* 4. 1–82.

Franke, Michael. 2012a. On Scales, Salience and Referential Language Use. In Maria Aloni, Floris Roelofsen & Katrin Schulz (eds.), *Amsterdam colloquium 2011*, Springer.

Franke, Michael. 2012b. Scales, Salience and Referential Safety: The Benefit of Communicating the Extreme. In Thomas C. Scott-Phillips, Mónica Tamariz, Erica A. Cartmill & James R Hurford (eds.), *The evolution of language: Proceedings of the 9th international conference (evolang 9)*, 118–125.

Frazee, Joey & David Beaver. 2010. Vagueness is rational under uncertainty. In Maria Aloni, Harald Bastiaanse, Tikitu de Jager & Katrin Schulz (eds.), *Logic, language and meaning: 17th amsterdam colloquium, amsterdam, the netherlands, december 16-18, 2009, revised selected papers*, 153–162. Springer.

Gallin, Daniel. 1975. *Intensional and higher-order modal logic: With applications to Montague semantics*. Elsevier.

Gärdenfors, Peter. 2000. *Conceptual spaces: The geometry of thought*. MIT press.

Ginzburg, Jonathan. 1995a. Resolving questions, I. *Linguistics and Philosophy* 18(5). 459–527.
Ginzburg, Jonathan. 1995b. Resolving questions, II. *Linguistics and Philosophy* 18(6). 567–609.
Golland, Dave, Percy Liang & Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Empirical methods in natural language processing (emnlp)*, 410–419.
Goodman, Noah D. & Daniel Lassiter. 2015. Probabilistic semantics and pragmatics: Uncertainty in language and thought. In Shalom Lappin & Chris Fox (eds.), *Handbook of Contemporary Semantic Theory*, Wiley-Blackwell 2nd edn.
Goodman, Noah D. & Andreas Stuhlmüller. 2012. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science* 5(1). 173–184.
Goodman, Noah D. & Joshua B. Tenenbaum. electronic. Probabilistic models of cognition. Retrieved February 5, 2015 from http://probmods.org.
Grice, H. Paul. 1957. Meaning. *The Philosophical Review* 66(3). 377–388.
Grice, H. Paul. 1975. Logic and conversation. In P. Cole & J. Morgan (eds.), *Syntax and semantics 9: Pragmatics*, 41–58. Academic Press.
Grice, H. Paul. 1989. *Studies in the Way of Words*. Harvard University Press.
Griffiths, Thomas L., Charles Kemp & Joshua B. Tenenbaum. 2008. Bayesian models of cognition. In Ron Sun (ed.), *Cambridge handbook of computational psychology*, 59–100. Cambridge University Press.
Griffiths, Thomas L, Edward Vul & Adam N Sanborn. 2012. Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science* 21(4). 263–268.
Groenendijk, Jeroen & Martin Stokhof. 1984. Studies in the Semantics of Questions and the Pragmatics of Answers. Ph.D. thesis, University of Amsterdam.
Hájek, Alan. 2003. What conditional probability could not be. *Synthese* 137(3). 273–323.
Heim, Irene. 1982. *The semantics of definite and indefinite noun phrases*: dissertation.
Heim, Irene. 2006. Remarks on comparative clauses as generalized quantifiers. Ms., MIT .
Heim, Irene. 2008. Decomposing antonyms. In *Sinn und bedeutung*, vol. 12, 212–225.
Horn, Laurence. 1989. *A Natural History of Negation*. University of Chicago Press.
Jacobson, Pauline. 1999. Towards a variable-free semantics. *Linguistics and Philosophy* 22(2). 117–185.
Jäger, Gerhard. 2007. Game dynamics connects semantics and pragmatics. In A.-V. Pietarinen (ed.), *Game theory and linguistic meaning*, vol. 18, 103–118. Elsevier.
Jäger, Gerhard & Christian Ebert. 2009. Pragmatic rationalizability. In *Proceedings of sinn und bedeutung 13*, 1–15.
Kamp, Hans. 1975. Two theories about adjectives. In E. Keenan (ed.), *Formal semantics of natural language*, 123–155. Cambridge University Press.
Kamp, Hans. 1981. The paradox of the heap. In U. Mönnich (ed.), *Aspects of philosophical logic*, 225–277. Springer.
Kao, Justine T, Jean Y Wu, Leon Bergen & Noah D Goodman. 2014. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences* 111(33). 12002–12007.
Kehler, Andrew & Hannah Rohde. 2013. A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics* 39. 1–37.
Kennedy, Chris. 1997. *Projecting the adjective: The syntax and semantics of gradability and*

comparison: U.C., Santa Cruz dissertation.
Kennedy, Chris. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy* 30(1). 1–45.
Kennedy, Chris & Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81(2). 345–381.
Klein, Ewan. 1980. A semantics for positive and comparative adjectives. *Linguistics and Philosophy* 4(1). 1–45.
Klein, Ewan. 1991. Comparatives. In A. von Stechow & D. Wunderlich (eds.), *Semantik: Ein internationales Handbuch der zeitgenössischen Forschung*, Walter de Gruyter.
Kratzer, Angelika. 1998. Scope or pseudo-scope? Are there wide-scope indefinites? In Susan Rothstein (ed.), *Events and grammar*, 163–196. Kluwer.
Kyburg, Alice. 2000. When vague sentences inform: A model of assertability. *Synthese* 124(2). 175–191.
Kyburg, Alice & Lenhart Schubert. 1993. Reconciling sharp true/false boundaries with scalar vagueness. In *First Conference of the Pacific Association for Computational Linguistics (PACLING 1993)*, 53–62. Simon Fraser University.
Lassiter, Daniel. 2011. Measurement and Modality: The Scalar Basis of Modal Semantics. Ph.D. thesis, New York University.
Lassiter, Daniel. 2012. Presuppositions, provisos, and probability. *Semantics & Pragmatics* 5. 1–37.
Lassiter, Daniel. 2015. Adjectival modification and gradation. In Shalom Lappin & Chris Fox (eds.), *Handbook of Contemporary Semantic Theory*, Wiley-Blackwell 2nd edn.
Lassiter, Daniel & Noah D. Goodman. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In Todd Snider (ed.), *Semantics & Linguistic Theory (SALT) 23*, 587–610. CLC Publications.
Lawry, Jonathan. 2008. Appropriateness measures: an uncertainty model for vague concepts. *Synthese* 161(2). 255–269.
Leitgeb, Hannes. 2014. The stability theory of belief. To appear in *The Philosophical Review*.
Lewis, David. 1969. *Convention: A Philosophical Study*. Harvard University Press.
Lewis, David. 1970. General semantics. *Synthese* 22(1). 18–67.
Lin, Hanti & Kevin T Kelly. 2012. Propositional reasoning that tracks probabilistic reasoning. *Journal of philosophical logic* 41(6). 957–981.
Luce, R. Duncan. 1959. *Individual choice behavior: A theoretical analysis*. John Wiley.
MacKay, David J.C. 2003. *Information theory, inference, and learning algorithms*. Cambridge University Press.
Montague, Richard. 1973. The proper treatment of quantification in ordinary English. In J. Hintikka, J. Moravcsik & P. Suppes (eds.), *Approaches to natural language*, vol. 49, 221–242. Reidel.
Morzycki, Marcin. to appear. *Modification*. Cambridge University Press.
Neal, Radford M. 1993. Probabilistic inference using markov chain monte carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.
Oaksford, Mike & Nick Chater. 2007. *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Over, David E. 2009. New paradigm psychology of reasoning. *Thinking and Reasoning* 15(4). 431–438.

Pearl, Judea. 2000. *Causality: Models, reasoning and inference*. Cambridge University Press.

Potts, Christopher. 2008. Interpretive Economy, Schelling points, and evolutionary stability. Ms., University of Massachusetts at Amherst.

Qing, Ciyang & Michael Franke. 2014. Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In Todd Snider (ed.), *Semantics & Linguistic Theory (SALT) 24*, 23–41. CLC Publications.

Roberts, Craige. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics* 91–136.

van Rooij, Robert. 2003. Questioning to resolve decision problems. *Linguistics and Philosophy* 26(6). 727–763.

Sassoon, Galit W. 2013. A typology of multidimensional adjectives. *Journal of Semantics* 30(3). 335–380.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27. 379–423.

Silverman, Bernard W. 1986. *Density estimation for statistics and data analysis*. Chapman & Hall.

Smith, Nathaniel J, Noah Goodman & Michael Frank. 2013. Learning and using language via recursive pragmatic reasoning about other agents. In Christopher J.C. Burges, Léon Bottou, Max Welling, Zoubin Ghahramani & Kilian Q. Weinberger (eds.), *Advances in neural information processing systems 26*, 3039–3047. Curran Associates, Inc.

Solt, Stephanie. 2011. Notes on the comparison class. In Rick Nouwen, Robert van Rooij, Uli Sauerland & Hans-Christian Schmitz (eds.), *Vagueness in communication*, 127–150. Springer.

Sorensen, Roy. 2012. The sorites and the generic overgeneralization effect. *Analysis* 72(3). 444–449.

Stalnaker, Robert. 1978. Assertion. In Peter Cole (ed.), *Syntax and semantics 9: Pragmatics*, Academic Press.

Stanley, Jason & Zoltan Szabó. 2000. On quantifier domain restriction. *Mind & Language* 15(2-3). 219–261.

von Stechow, Arnim. 1984. Comparing semantic theories of comparison. *Journal of Semantics* 3(1). 1–77.

Sutton, Peter. 2013. *Vagueness, Communication and Semantic Information*: King's College London dissertation.

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

Tenenbaum, Joshua B., Charles Kemp, Tom L. Griffiths & Noah D. Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science* 331(6022). 1279–1285.

Vogel, Adam, Max Bodoia, Christopher Potts & Dan Jurafsky. 2013a. Emergence of Gricean maxims from multi-agent decision theory. In *Human language technologies: The 2013 annual conference of the North American chapter of the Association for Computational Linguistics*, 1072–1081. Stroudsburg, PA: Association for Computational Linguistics.

Vogel, Adam, Christopher Potts & Dan Jurafsky. 2013b. Implicatures and nested beliefs in approximate Decentralized-POMDPs. In *Proceedings of the 2013 annual conference of the Association*

for Computational Linguistics, 74–80. Stroudsburg, PA: Association for Computational Linguistics.

Vul, E., N.D. Goodman, T.L. Griffiths & J.B. Tenenbaum. 2014. One and done? Optimal decisions from very few samples. *Cognitive Science* 38(4). 599–637.

Xu, Fei & Joshua B. Tenenbaum. 2007. Word learning as Bayesian inference. *Psychological Review* 114(2). 245–272.