PROBABILISTIC SEMANTICS FOR VAGUENESS

BY

STEVEN FONG-YI LEE

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Philosophy in the Graduate College of the University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Professor Peter N. Lasersohn, Chair
Professor Timothy G. McCarthy
Associate Professor Jonathan M. Livengood
Assistant Professor Benjamin A. Levinstein
Abstract

In this dissertation I argue that truth-conditional semantics for vague predicates, combined with a Bayesian account of statistical inference incorporating knowledge of truth-conditions of utterances, generates false predictions regarding negations and metalinguistic inference. I thus propose a fundamentally probabilistic semantics for vagueness on which the meaning of a vague predicate is a likelihood function on the states it encodes, with these likelihoods being generated via reinforcement learning in a signaling game.
for Mom and Dad, and Lillian, Stella, and Nathan
Acknowledgements

This dissertation would not have been possible without the help of many people, most of whom will go unnamed. First, I would like to thank my advisor, Peter Lasersohn. Thank you for your years of support, encouragement, and patience as I’ve worked on this dissertation. I would also like to thank the other members of my committee, Timothy McCarthy, Jon Livengood, and Ben Levinstein. Thank you also to Rob Koons and Dan Bonevac, from UT Austin, where I started my doctoral studies. And, to the members of my preliminary committee who could not be part of my final committee, Andy Arana and Dan Korman. I would also like to thank my good friend Eric Mack, for years of life together as philosophy graduate students. Thank you also to all of my teachers at the Talbot School of Theology, including JP Moreland, Garry DeWeese, Doug Geivett, and David Horner. And, to Noel Saënz, for your encouragement as I have finished my doctoral studies. Thank you also to Christy Ann Foster and Erin Tarr, for taking care of the administrative duties without which graduate studies would be impossible. Thank you also to our departmental chair, Kirk Sanders. I would also like to thank everyone at Caffe Paradiso, Flying Machine Avionics, Brewlab, and Café Kopi, where so much of this dissertation has been worked on. Thank you also to my good friends Josue and Mariane Bastos, and Sam Chen and Monica Huang. And thank you to everyone at the Montessori School of Champaign-Urbana, for taking such good
care of Nathan and Stella as I have written this dissertation.
Contents

Chapter 1: Introduction ............................................. 1

1.1 Introduction ..................................................... 1

1.2 Degree-Theoretic Semantics .................................. 12
  1.2.1 Motivating Degrees as a Semantic Type ................. 12
  1.2.2 Relative vs Absolute Gradable Adjectives ............... 15
  1.2.3 A Semantics Proper .................................... 18

Chapter 2: Against Threshold Semantics ......................... 26

2.1 Introduction ..................................................... 26

2.2 Bayes-Grice Models of Vagueness ............................ 28

2.3 Negation and Relative Gradable Adjectives .................. 38
  2.3.1 Unexpected Inferences .................................. 38
  2.3.2 Equal and Opposite Interpretations ..................... 41

2.4 Solutions Within the Model .................................. 44
  2.4.1 Varying the Parameters .................................. 44
  2.4.2 Alternative Utterances .................................. 47

2.5 Extending the Model by Pragmatic Ascent, Part I .......... 53
  2.5.1 Abstract Scale Structures and Non-Uniform Threshold Priors . . . . 53
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5.2 Privileging Non-negated Posteriors</td>
<td>65</td>
</tr>
<tr>
<td>2.6 Modifying the Model: Assuming $\theta_1 &gt; \theta_2$</td>
<td>73</td>
</tr>
<tr>
<td>2.7 Extending the Model by Pragmatic Ascent, Part II</td>
<td>87</td>
</tr>
<tr>
<td>2.8 An Alternative Bayes-Grice Model</td>
<td>99</td>
</tr>
<tr>
<td>2.8.1 Motivation: Unrealistic Sender Behavior</td>
<td>99</td>
</tr>
<tr>
<td>2.8.2 A Distance-Weighted Informativity Model</td>
<td>103</td>
</tr>
<tr>
<td>2.9 Non-Uniform $\rho_0 \theta$ Priors</td>
<td>108</td>
</tr>
<tr>
<td>Chapter 3: Game Theory and Probabilistic Signaling</td>
<td>119</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>119</td>
</tr>
<tr>
<td>3.2 The Puzzle of Game Theory and Vagueness</td>
<td>120</td>
</tr>
<tr>
<td>3.3 Evolutionary Game Theory to the Rescue?</td>
<td>124</td>
</tr>
<tr>
<td>3.3.1 Replicator Dynamics</td>
<td>124</td>
</tr>
<tr>
<td>3.3.2 Best Response Dynamics</td>
<td>129</td>
</tr>
<tr>
<td>3.3.3 Two Solutions</td>
<td>133</td>
</tr>
<tr>
<td>3.4 Symmetrization and Normal Distributions</td>
<td>142</td>
</tr>
<tr>
<td>3.4.1 Symmetrization</td>
<td>145</td>
</tr>
<tr>
<td>3.4.2 Normal State Distributions</td>
<td>150</td>
</tr>
<tr>
<td>Chapter 4: A Dynamic Probabilistic View</td>
<td>158</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>158</td>
</tr>
<tr>
<td>4.2 Showing Equivalence</td>
<td>160</td>
</tr>
<tr>
<td>4.2.1 From Barker to Stalnaker/Kaplan</td>
<td>160</td>
</tr>
<tr>
<td>4.2.2 From Stalnaker to Montague</td>
<td>171</td>
</tr>
</tbody>
</table>
Chapter 5: Conclusion .............................................. 240

5.1 How we got here ................................................. 240

5.2 Areas for Further Exploration ................................. 244

5.2.1 Various Extensions ........................................... 244

5.2.2 Scale structure ................................................. 247

5.2.3 The Sorites .................................................... 252

Chapter 6: Bibliography ............................................ 256
Chapter 1

Introduction

1.1 Introduction

This dissertation constitutes an extended argument against truth-conditional semantics for relative gradable adjectives—that is, words like ‘tall’, which uniformly give rise to a group of phenomena associated with what is known in philosophical circles as vagueness—and in favor of an alternative, probabilistic conception of a semantics for relative gradable adjectives. The phenomena around which my arguments are centered are the statistical inferences that we typically make upon hearing sentences such as ‘Feynman is tall.’ Consider if I say to you, ‘Describe Feynman to me’, and you say, ‘Well, Feynman is tall’: I typically decrease the probability that I assign to him being 5′9″, increase the probability that I assign to him being 6′1″, and also increase the probability that I assign to him being 6′6″, (but not by very much, since being 6′6″ is so rare.) It seems obvious that these statistical inferences are a result at least in part of my beliefs concerning the meaning of ‘tall’, since if I believed ‘tall’ meant what we ordinarily mean by ‘short’, or I did not know that ‘tall’ meant anything at all, I would have engaged in an entirely different pattern of statistical inference. Thus,
if we identify the meaning of ‘Feynman is tall’ with its truth-conditions, then my beliefs concerning the truth-conditions of that sentence must enter into an explanation of how I engage in these statistical inferences.

It is initially puzzling, however, how to do this: one of the distinguishing features of relative gradable adjectives, and vague words is general, is that they admit of borderline cases: under what conditions is it true that Feynman is tall? When he is 5′11″? Or 6′0″, or 6′1″? It seems clearly true that Feynman is tall if he is, say, 6′6″, and clearly not true if he is, say, 5′6″. But between such clear cases are borderline cases, even after we account for some relevant comparison class—in this case, perhaps, the class of American adult males. Thus, the most prominent semantic theories for relative gradable adjectives in the literature today handle the difficult question of the truth-conditions for such adjectives, and the determination of the truth or falsity of their borderline cases, by appealing to a context-sensitive threshold: in the case of ‘tall’, a threshold such that above it, ‘Feynman is tall’ comes out true, and below it, it comes out false. (In §1.2 of this introduction I present in more detail one particularly prominent example of such a theory, due to Kennedy (2007).) This has the advantage of allowing such semantic theories to avoid committing to the claim that, say, one counts as tall for an American adult male when one is 6′2″, but no shorter. However, such theories are still committed to a sharp boundary between the conditions under which ‘Feynman is tall’ is true, and the conditions under which it is false; in contrast, the statistical inferences we typically engage in when we hear that sentence asserted seem smooth and continuous: my own sense is that upon hearing Feynman called ‘tall’ I would smoothly increase from 0 the probability of his being a given height, starting from about 5′10″, up to about 6′2″, and then
smoothly decrease it for all heights from there on up, with little probability being assigned to him being, say, 6′6″. How then we can generate smooth probability distributions from sharp cutoffs?

In Chapter 2, I examine a recent proposal by Lassiter and Goodman (2017) that attempts to fill in this gap between a hearer’s knowledge of the truth-conditions of ‘Feynman is tall’, and their probability distribution for Feynman’s height posterior to hearing that sentence. Their model is a development of Gricean pragmatics along Bayesian lines, in which a pragmatic receiver $\rho_2$ assumes that he is listening to a pragmatic speaker $\sigma_1$ of finite rationality who is, while choosing between alternative utterances—whether to call him ‘short’, ‘tall’, or saying nothing at all about his height—attempting to be maximally informative as to the height they observe Feynman to be, while balancing for the cost of the message. This informativity is defined relative to $\rho_2$’s conception of a hypothetical literal receiver $\rho_0$ who, given that he hears ‘Feynman is tall’ used with a particular threshold value, assigns 0 probability to all heights less than that threshold, and proportionally redistributes the missing probability mass over all heights greater than or equal to that threshold. The truth-conditions of the utterance thus show up in the literal receiver $\rho_0$, who conditionalizes on the literal truth, relative to the given threshold, of the utterance: if the threshold was 6′1″ and I learned only that Feynman was no shorter than that, I would assign no probability to him being any shorter than that, and increase the probability of him being 6′1″, increase but a little less this time the probability of him being 6′2″, and so on for greater heights. The pragmatic receiver $\rho_2$ thus has an utterance production model $\sigma_1(‘Feynman is tall'|h, \theta)$ that tells him, for a given height for Feynman and threshold for tallness, the probability that $\sigma_1$ will say ‘Feynman is tall’. By
conditionalizing on a prior distribution for heights and a prior distribution for thresholds, $\rho_2$ infers a posterior probability distribution for Feynman’s height after marginalizing out his posterior threshold distribution.

The gist of my argument in Chapter 2 is then that that same model, given the same truth-conditional semantics for ‘tall’, along with the common understanding of the semantics of negation, makes the wrong predictions about the statistical inferences we make regarding Feynman’s height, upon hearing an interlocutor’s utterance of ‘Feynman is not tall.’ That is, upon hearing you say ‘Feynman is not tall’, intuitively, it seems that I decrease the probability I assign to him being 6'6", or even 6'3", and correspondingly increase the probability that I assign to him being 5'9", or even 5'6". However, the model in question, combined with the assumption that ‘Feynman is not tall’ is true just in case ‘Feynman is tall’ is false, predicts that I will engage in a pattern of statistical inference regarding Feynman’s height that is quite different from the intuitively correct one: it predicts that I will engage in a pattern of statistical inference concerning Feynman’s height much closer to that which, intuitively, I engage in when I hear you say ‘Feynman is short.’ In fact, under any plausible parameter settings, their model predicts that upon hearing ‘Feynman is not tall’, I will engage in a pattern of statistical inference regarding Feynman’s height that is in fact stronger than that which I engage in upon hearing ‘Feynman is short’, and this is implausible: it is one thing to say someone is not tall, it is entirely more to say they are short. Furthermore, their model predicts that upon hearing ‘Feynman is not tall’ I will engage in a pattern of statistical inference on which that utterance is interpreted (as far is its implications regarding the probabilities assigned to various heights is concerned), in a manner opposite from, but far
stronger than, ‘Feynman is tall.’ Again, this is intuitively incorrect: we want ‘Feynman is not tall’ to imply something weaker than ‘Feynman is tall’, as far as the probabilities assigned to Feynman’s height is concerned.

After presenting the details of Lassiter & Goodman’s model, and then presenting the problems described in the foregoing paragraph, I spend the remainder of Chapter 2 attempting to resolve these problems: First, I search for solutions within their model, by either varying its parameters, or examining effects of the conversational context. Second, I attempt to extend their model in a manner consistent with its initial assumptions, by ascending to higher levels of the sender-receiver hierarchy: what, for example, a $\rho_4$ receiver would infer from a $\sigma_3$ sender thinking about our $\rho_2$ receiver. Third, I attempt to modify their model. After finding that none of these strategies satisfactorily resolves either of the foregoing problems, I suggest that we might attempt to fix, external to their model, a plausible distribution for the threshold above which ‘Feynman is tall’ is true, and below which it is false. I then show that if we do so, then we can solve, at least at lower levels of the sender-receiver pragmatic hierarchy, both of the problems with negations of vague predicates.

I then argue that fixing this threshold distribution independently of any model of interpretation is liable to the objection that there is a simpler model, on which the conditional encoding probability over heights $h$ and thresholds $\theta$ for tallness—that is, $\sigma_1(\text{‘Feynman is tall’}|h,\theta)$—is replaced by a conditional encoding probability over just heights: $\sigma_1(\text{‘Feynman is tall’}|h)$. That is, if I think that the threshold for tallness among American adult males is definitely at 6’1”, I will think that the probability that the speaker will say ‘Feynman is tall’ if Feynman is less than 6’1” is 0, and 1 if he is as tall as or taller than 6’1”. Similarly, if I think there is
a uniform probability that the threshold distribution is between 6'0" and 6'2", I will think that the probability that the speaker will say ‘Feynman is tall’ if Feynman is less than 6'0" is 0, and 1 if he is as tall as or taller than 6'2", and will linearly increase between 6'0" and 6'2".

On this simpler model, with the thresholds absent, the truth-conditions have dropped out too; listeners consider only the probability that the speaker would say ‘Feynman is tall’ given that Feynman is a given height. If this model is viable, then we might well wonder how it is that we come use words in this probabilistic manner, and thus how listeners come to know these conditional encoding probabilities. In the background is a question as to the source of our knowledge of the threshold distribution: do we infer a threshold distribution for ‘tall’ from our observations of how the encoding behavior of our co-linguals varies with the heights of the things they are describing? Or do we infer a threshold distribution on the basis of what we take to be our pre-existing common knowledge of the truth-conditions of our words, along with a characterization of the speaker’s relevant psychological features, such as their degree of rationality, and their subjective cost of talking? As one might anticipate, I will plump for the former option, but that requires that we be able to give a plausible explanation for how such probabilistic encoding behavior might arise.

Thus, in Chapter 3, I discuss some recent attempts to provide such an explanation. I start by introducing the notion of a signaling game from D. Lewis (1969): here, a sender observes a state, determined by nature and unobservable to a receiver, and must choose a signal to encode that state. The receiver must decode that signal by taking appropriate actions, and both players receive a common reward if the sender and receiver appropriately coordinate.
their encoding and decoding strategies. I also introduce the associated notion of a convention, which is a strict Nash equilibrium satisfying certain other constraints. I then present the proof from Lipman (2009) that vagueness, understood in the context of a signaling game as probabilistic encoding over different states of the world, is never a strict Nash equilibrium, and thus cannot be explained purely in terms of Lewis’ notion of convention. I then consider the prospects for an alternative explanation, in terms of evolutionary game theory, for how probabilistic signaling behavior might arise over multiple iterations of play of a signaling game: First, I consider the replicator dynamics, under which (roughly) the probability at a given time that a given strategy will be played—in the case of a signaling game where the states of nature are the heights of an object, the probability of a sender encoding a given height with a given signal at a given iteration of the game, and the probability of a receiver decoding a given signal as a given height at a given iteration of the game—is determined by the success of that strategy in past iterations of the game. I then show that probabilistic encoding behavior is not predicted to arise under the replicator dynamics. Second, I consider the best response dynamics, according to which at a given iteration of the game, a player chooses the best response to other player’s choices in previous iterations of the game. Again, I show that probabilistic encoding behavior is not predicted to arise under the best response dynamics of evolutionary game theory.

Then, I turn to two recent attempts to explain probabilistic encoding behavior in signaling games: First, I discuss work from Franke, Jäger, and Van Rooij (2010), on which probabilistic encoding (and decoding) behavior is derived from an assumption of limited rationality on the part of the sender and receiver: the sender and receiver do not deterministically choose
payoff maximizing encoding and decoding strategies; instead, the probability that they will choose a given strategy monotonically increases with its payoff. However, I find that this model fails to provide an adequate explanation of why it is that senders and receivers fail to choose payoff maximizing strategies, and thus it fails to predict to what degree, and under what conditions, probabilistic encoding and decoding behavior will emerge.

Next, I present work from O’Connor (2014), which draws on reinforcement learning to explain how probabilistic encoding behavior arises in signaling games. On this view, as in the replicator dynamics, the probability of encoding a given height with a given signal at a given iteration of the game is determined by the success of that strategy in past iterations of the game. Crucially, however, the success of that strategy in a given iteration of the game also increases the probability that heights nearby that height will be encoded by that signal in the next iteration of the game. Similarly, the success of a given decoding strategy in a given iteration of the game will increase the probability that in the next iteration of the game, the state that that signal was decoded as, and also nearby states, will be used to decode that same signal. This approach then yields stable probabilistic encoding behavior over multiple iterations of game play.

Finally, I extend O’Connor’s work in two ways: First, I argue that an adequate game theoretic model of the origins of vagueness should allow players to switch roles as sender and receiver over various iterations of the game. Now, the sender’s probabilities of encoding a given height are represented in a state-to-signal weight matrix; for any given observed state, the probability that a sender chooses a given signal to encode that state is the ratio of the weight for that signal for that height, to the sum of the weights of all signals, for that
height. Likewise, the receiver’s probabilities of decoding a given signal as a given state is represented in a signal-to-state matrix; for any given received signal, the probability that a receiver chooses a given height to decode that signal is the ratio of the weight for that height for that signal, to the sum of the weights of all heights, for that signal. But, if we allow a sender’s state-to-signal weight matrix to be transposed and become a signal-to-state weight matrix, and vice-versa, then we can have a single representation of both the encoding and decoding probabilities, and a player’s encoding behavior can be affected by his past encoding and decoding behavior; similarly for their decoding behavior. Thus, an encoder can become a decoder, and vice-versa. I show that, as one might hope, vagueness persists under these conditions.

My second extension of O’Connor’s framework concerns the distribution of the states that the sender observes: O’Connor only examines signaling games where the heights that the receiver observes follow a uniform distribution over a bounded range. However, often the features that we encode using vague predicates follow a Gaussian, and not uniform, distribution; thus, the heights of American adult males is roughly a Gaussian distribution, with an average of 5’9” and a standard deviation of 3”. I demonstrate that the kind of probabilistic signaling behavior characteristic of vagueness can emerge when the distribution of heights is not uniform, but instead Gaussian.

Now, I argued in Chapter 2 that one specific model of how, from knowledge of the truth-conditions of sentences such as ‘Feynman is tall’, listeners engage in the statistical inference patterns that we pre-theoretically expect them to engage in upon hearing that sentence assertively uttered, is subject to problems with negation that motivate a move to an alternative,
simpler theory of statistical inference: one in which the truth-conditions of the utterance no longer appear, having been replaced by the receiver’s conception of how the probability of the sender uttering what they did, varies with the height of the object described. In Chapter 3 I then attempted to justify this theory by pointing to a plausible explanation, pre-existing in the literature, for how such probabilistic encoding behavior might arise independent of any sender or receiver considering the truth-conditions of their utterances. In Chapter 4, I argue that the problems with negation discussed in Chapter 2 generalize to a class of models, all of which attempt to explain, based on our knowledge of the truth-conditions of sentences containing vague predicates, how we engage in the expected statistical inference patterns upon hearing such sentences. I also present in Chapter 4 two additional arguments against such a class of models. First, however, I demonstrate that a number of operations that have featured prominently in recent semantics and pragmatics are in fact equivalent, and are all special cases of an equation identical in form if not substance to one seen in Chapter 2, that generates statistical inference patterns from utterances of sentences containing vague predicates. I then argue that conformation to this equation is in fact a requirement for any plausible Bayesian model of how, given utterances containing vague predicates, listeners draw the pre-theoretically expected statistical inference patterns from beliefs concerning the truth-conditions of such utterances. This equation then characterizes the general class of models that I am concerned to reject.

Of course, it must be admitted that just because a model does not belong to this class of models, does not entail that it does not exist; thus, perhaps there is a non-Bayesian model for how we can go from knowledge of the truth-conditions of ‘Feynman is tall’ to the
kinds of statistical inferences we typically make upon hearing that utterance. Nonetheless, without first producing such a model we seem unable to determine whether one might exist. Accordingly, my strategy here is not to exhaustively categorize all of the possible models of how the expected statistical inference patterns can be generated from knowledge of the truth-conditions of utterances containing vague predicates; instead, I take the initial plausibility of this class of models as evidence that it is a good starting point, and make what changes are necessary to the semantics, in order to generate the expected statistical inferences.

I thus present an alternative model of the manner in which utterances of sentences with vague predicates give rise to the expected statistical inference patterns, of both metalinguistic and non-metalinguistic kinds. On this model, in non-metalinguistic uses of vague predicates the truth-conditions of utterances of sentences containing vague predicates have been, at bottom, replaced by the kind of probabilistic encoding and decoding behavior the origins of which were discussed in Chapter 3. Bayesian calculations based on anything recognizable as the truth-conditions of such sentences, are used only to generate metalinguistic inference patterns, with a probability distribution for thresholds defined by taking the derivative of the sender’s encoding probabilities. Knowledge of the truth-conditions of an utterance containing a vague predicate is only reflected as knowledge of the encoding probabilities, conditional on first, the subject of the utterance’s degree of deviation, with respect to the property expressed by that vague predicate, from some measure of central tendency, and second, the threshold for application of that vague predicate, again expressed as some requisite degree of deviation, with respect to the property expressed by that vague predicate. The upshot is that knowledge of the truth-conditions of an utterance of a sentence containing a vague predicate
does not play any role in generating the non-metalinguistic statistical inference patterns that we expect from utterances of sentences with vague predicates. Finally, on the assumption that it is knowledge of the meaning of vague predicates that allows us to engage in the the kind of statistical inferences we typically engage in when we hear utterances of sentences containing vague predicates, I propose a probabilistic semantics for vague predicates, and demonstrate how we can integrate such a view with an existing degree-theoretic, truth-conditional semantics.

1.2 Degree-Theoretic Semantics

1.2.1 Motivating Degrees as a Semantic Type

In order to isolate the kinds of words that give rise to vagueness, we can observe that some words are gradable, and some are not. Furthermore, following Kennedy (2007) and earlier work in Kennedy and McNally (2005), Rusiecki (1985), and Unger (1975), we can distinguish between absolute and relative gradable adjectives, only the latter of which give rise to the phenomenon of vagueness. As Kennedy points out, gradable adjectives—such as ‘tall’, ‘thick’, ‘pure’, ‘flat’, and ‘opaque’—can be distinguished from non-gradable adjectives—such as ‘biological’, ‘postal’, and ‘tertiary’—in that the former, but not the latter, can feature in constructions with various kinds of degree morphology, including comparative morphemes (more, less, as), intensifiers (very, quite, extremely), and sufficiency morphemes (enough, too, so). For example,
(1) Non-gradable adjectives
   a. * That book is more biological than this one
   b. * This office is very postal
   c. * This derivative is too tertiary for the computer

all sound unacceptable. In contrast,

(2) Gradable adjectives
   a. This piece of glass is as thick as that one
   b. This table is very flat
   c. This window is so opaque that I cannot see through it

all sound acceptable. Another difference not yet noted to my knowledge is that many (but not all) gradable, but not non-gradable, adjectives occur with measure phrases:\footnote{Ben Levinstein points out examples such as:}

(5) a. * This derivative is 1.8 degrees tertiary
   b. This piece of glass is 2 inches thick

\footnote{Ben Levinstein points out examples such as:}

(3) a. Alice is 5 months pregnant.
   b. Bob is 5 years gone.

but these do not seem to indicate the degree of pregnancy or of deadness. Instead I suggest they are elliptical for:

(4) a. Alice has been pregnant for 5 years.
   b. Bob has been dead for 5 years.
The common feature of degree morphology and measure phrases is that they appear to modify the degree or gradation to which the modified adjective holds of the object in question; to be told a piece of glass is thick is one thing, to be told a piece of glass is very thick is another, and to be told a piece of glass is 2 inches thick is yet another. The difference between the first and second message appears to be that the latter message tells us that the glass has a greater degree of thickness than the former message; the difference between the first and third message appears to be that the latter message tells us exactly what degree of thickness the glass possesses, while the former message merely conveys a range of possible thicknesses. On account of their ability to combine with degree morphology and measure phrases, gradable adjectives have been semantically analyzed as involving functions on degrees, where totally ordered sets of degrees constitute a scale. Thus, the meaning of the positive form ‘tall’ has been taken to be, roughly, having a degree greater on the height scale than a contextually determined degree standard for some relevant comparison class (which is, depending on the version of the view, either pragmatically or semantically determined). Likewise the meaning of the comparative form ‘taller than y’ has been taken to be, roughly, having a degree on the height scale greater than the maximal degree of the height of y. And the meaning of the measure phrase form ‘1.8 meters tall’ has been taken to be, roughly, having a degree on the height scale equal to 1.8 meters. Thus, in addition to entities, truth-values, possible worlds, and times, such theories of gradable adjectives countenance degrees among the basic semantic types.
1.2.2 Relative vs Absolute Gradable Adjectives

Next, I want to explain the distinction between absolute and relative gradable adjectives, for three reasons: First, in order to more precisely characterize the kinds of adjectives that give rise to vagueness, and why they do so. Second, in order to show that there is a real distinction here, so that differences in the structures of the degree scales have real explanatory work to do. And third, in order to provide context for how I will later diverge from the usual characterization of the degree scales for both absolute and relative gradable adjectives.

As for the distinction: Kennedy (2007) points out two ways (among others) of observing the difference between absolute and relative gradable adjectives. First, if we control for imprecision, we can see that absolute gradable adjectives allow for natural precisifications, but relative gradable adjectives do not. An utterance of (5b) can be varyingly imprecise, depending on the context of use; if we are talking about armored cars, (5b) might be appropriately used to describe a piece of glass 1.95 inches thick, but if we are talking about satellite telescope lenses, (5b) might be inappropriate to talk about a piece of glass 1.999 inches thick. Similarly, in certain contexts

(6) The building is empty

might be appropriately used to describe a building with a janitor in it late at night; in other contexts it might require that the building be absolutely unpopulated, such as if we were about to demolish it. The fact that ‘empty’ has imprecise uses should not obscure the fact that it has a natural precisification of there being absolutely no persons present inside; it is thus an absolute adjective, but still, in view of ‘emptier’, ‘as empty as’, and so on, a gradable
adjective. In contrast, ‘tall’ admits of no natural precisification, even after we account for a comparison class; to draw the line for ‘tall’ among male collegiate basketball players at exactly 80.000 inches seems as arbitrary as at exactly 80.001 inches.

Second, relative gradable adjectives lack entailments had by absolute gradable adjectives; contrast tall with pure. Something completely pure has absolutely no impurities; to say that something is impure is then to say that it has a degree of impurity greater than the minimum of having absolutely no impurities. Then the denial that something is impure entails that it has no impurities, and we predict that

(7) The assay is not impure, but there are some undesired surfactants in it is contradictory. In contrast, ‘tall’ has no such entailments; consider:

(8) John is not tall, but he is slightly taller than average.

The latter half of (8) requires that John have some degree of height, but John not being tall of course does not rule that out.

Third, these and other facts suggests that absolute gradable adjectives and relative gradable adjectives can be distinguished by the presence or absence of absolute minimum or maximum values on the scale of degrees that these adjectives stand for; this in turn suggests that degree modifiers that pick out maximal or minimal degrees should be available or unavailable depending on the presence or absence of the minimum or maximum values. Maximal degree modifiers are words such as completely, perfectly, and totally; minimal degree modifiers are words such as slightly, partly, and a little. And indeed, we find that
are all unacceptable, but the maximal and minimal modifiers of antonym pairs that encode degree scales with absolute maximums are acceptable:

(10)  a. perfectly/slightly: certain, safe, pure, accurate  

     b. perfectly/slightly: uncertain, dangerous, impure, inaccurate

as are the minimal and maximal modifiers of antonym pairs that encode degree scales with absolute minimums:

(11)  a. perfectly/slightly: bent, bumpy, dirty, worried  

     b. perfectly/slightly: straight, flat, clean, unworried

Finally, we see that some words allow both maximal and minimal degree modifiers:

(12)  a. perfectly/slightly: full, open, opaque  

     b. perfectly/slightly: empty, closed, transparent

Perhaps most importantly for a theory of vagueness, the lack of natural precisifications seems to give rise to the classic symptoms of vague predicates: susceptibility to Sorites arguments, the existence of borderline cases, and context-sensitivity. With a natural precisification we seem able to resist the first step in the Sorites argument: ‘Here, and here alone, is where the theater is empty, all else is loose talk’; without a natural precisification, we cannot stop and say ‘Here, here is where the tall are, and no less.’ With a natural precisification, we
can eliminate borderline cases: ‘Empty means empty, nada, zilch, no one’; without, we must admit: ‘Is 6’0” tall? Kind of.’ And with a natural precisification we can see that what really counts as empty is not context sensitive: if I say

\[(13)\text{ Please give me the empty one,}\]

and ask for one of two cups, one of which is emptier than the other but neither of which are really empty, you will not know which one to give me, as neither is empty. If I ask you

\[(14)\text{ Please give me the tall one,}\]

and ask for one of two cups, one of which is significantly taller than the other, but neither of which are actually tall, you will know which one to give me.

### 1.2.3 A Semantics Proper

From the foregoing discussion it is clear that there is some semantic element in common between the positive, comparative, superlative, and other variants of a gradable adjectives, both relative and absolute: ‘tall’, ‘taller’, ‘tallest’, and other such variants all share something in common, and a good semantic analysis of ‘tall’ should show just what that is. ‘taller’, for example, has been analyzed as involving both a comparative degree morpheme, (the \([\text{Deg-er}]\) ending), and an adjectival phrase \([\text{AP tall}]\) that the comparative degree morpheme takes as input. \([\text{AP tall}]\) is thus taken to be a particular measure function; that is, a function from various entities \(e\) to various degrees of height \(d\):
\[
\llbracket \text{[AP tall]} \rrbracket = \lambda x.\text{tall}(x) \tag{1.1}
\]

'\text{er}' is then taken to be a function on measure functions: given any measure function, it delivers a function which, given an entity (the thing to which the subject is compared), delivers a function which given another entity (the subject), delivers a truth value:

\[
\llbracket \text{[Deg \text{-er}]} \rrbracket = \lambda g(\langle e,d \rangle \lambda y.e \lambda x.e.g(x) > g(y) \tag{1.2}
\]

Thus, since semantic composition is functional application, we have

\[
\llbracket \text{[P \text{DegP} [AP tall] -er]} \rrbracket = \llbracket \text{[DegP -er]} \rrbracket(\llbracket \text{[AP tall]} \rrbracket)
= [\lambda g(\langle e,d \rangle \lambda y.e \lambda x.e.g(x) > g(y))(\lambda x.e.\text{tall}(x))]
= \lambda y.e \lambda x.e.\text{tall}(x) > \text{tall}(y) \tag{1.3}
\]

Likewise, a common analysis of the unmarked, positive form tall is that it involves the same adjectival phrase \text{[AP tall]} and an unarticulated, but still present, morpheme \text{[Deg pos-]} that takes a measure function and delivers a function from entities to truth values. It does so by deriving from the measure function a degree of possession of the property measured by the measure function that an entity must have in order for it to satisfy the vague predicate in question; in short, it derives a standard of application from the measure function. This is usually taken to involve a function \textbf{s} that takes the measure function as input, and outputs the degree required for satisfaction of the vague predicate. In this case, we have:

19
\[
\llbracket [\text{DegP pos-}] \rrbracket = \lambda g_{(e,d)} \lambda x_e. g(x) \geq s(g)
\] (1.4)

If we compose \([\text{Deg pos-}]\) and \([\text{AP tall}]\) to get the bare positive form ‘tall’, we have:

\[
\llbracket [\text{DegP pos-} [\text{AP tall}]] \rrbracket = \llbracket [\text{DegP pos-}] \rrbracket (\llbracket [\text{AP tall}] \rrbracket)
\]
\[
= [\lambda g_{(e,d)} \lambda x_e. g(x) \geq s(g)] (\lambda x_e. \text{tall}(x))
\] (1.5)
\[
= \lambda x_e. \text{tall}(x) \geq s(\text{tall})
\]

Now, one clear problem with an analysis such as in (1.4) is that it fails to account for the context-sensitivity of ‘tall’: in a context where we are watching Feynman standing down on the basketball court next to some NBA players, ‘Feynman is tall’ seems false; in a context where Feynman is standing around with his shorter colleagues, ‘Feynman is tall’ seems true. And this is of course related to the problem of the comparison class: one way to account for the apparent context-sensitivity of vague predicates is to posit, in the absence of a prepositional modifier such as ‘for a physicist’, an implicit comparison class that is contextually determined. In the first case above, the comparison class is perhaps the set of basketball players, or the set of people on the court; in the second case, the comparison class seems to be his shorter colleagues standing around him. But context and ‘for’-phrases do not seem to be the only ways to introduce a comparison class; another way to do so is to use a vague predicate to modify a noun phrase: for example, ‘Feynman is a tall physicist.’ If we think that ‘Feynman is a tall physicist’ and ‘Feynman is tall for a physicist’ and ‘Feynman is tall’ (when uttered in a context in which the comparison class is physicists) all have the
same meaning, then we might be tempted to give an analysis such as:

\[ \llbracket \text{DegP pos-} \rrbracket = \lambda g_{(e,d)} \lambda k_{(e,t)} \lambda x \cdot g(x) \geq s(k)(g) \]  

where our comparison class \( k \) is introduced by either a nominal such as ‘physicist’ or the prepositional phrase ‘for a physicist’, and \( s \) computes the standard of comparison from the restriction of the function \( g \) to the set of physicists. There are problems with such an account, however.

First, we can make good sense of

(15) Feynman is a tall physicist, but he is not tall for a physicist.

If, however, modified nominal constructions and prepositionally modified constructions introduce the same comparison class in the same way, as per (1.6), then (15) should be a flat-out contradiction. In fact, (15) indicates that in modified nominal constructions such as ‘Feynman is a tall physicist’, the nominal does not determine the comparison class, in contrast to prepositionally modified constructions, in which it does.

Second, we have been assuming that \( s \) computes either the average, or the median, or some precise standard of comparison from the measure function. As Graff (2000) has noted, however, adding the comparison class of physicists to ‘tall’ should then eliminate any vagueness in ‘tall for a physicist’, when in fact it does not: if there is a group of famous, extremely tall physicists standing near Feynman, I might know that the mean height for physicists is 6’1” and still not think ‘Feynman is tall for a physicist’ is true—and likewise for the median,
mode, or other statistical measure of the heights of physicists. Thus, even for prepositionally modified constructions, there must be some further element of context-sensitivity that is not captured by (1.6).

Thus, (1.6) is at best an analysis of a bare positive construction such as ‘Feynman is tall’, where context supplies the comparison class, and vagueness results from uncertainty surrounding the identity of the comparison class. Even this, however, seems wrong: once we make explicit the comparison class via a prepositional modifier such as ‘for a physicist’, we should eliminate all vagueness, if (1.6) is correct. As we have seen, this is not so.

If (1.6) is insufficient, how can we account for the manner in which comparison classes can affect the semantics of vague predicates? One option, as in Kennedy (2007), is to claim that the denotation of a vague predicate is such that \( s \) never takes as input a comparison class variable as it does in (1.6), whether that predicate occurs in constructions such as ‘Feynman is tall’, ‘Feynman is a tall physicist’, or ‘Feynman is tall for a physicist’. Instead, on this view, we treat ‘Feynman is a tall physicist’ as equivalent to ‘Feynman is tall and a physicist’, and account for the apparent manner in which ‘physicist’ affects the degree required to be tall by claiming that \( s \) is a context-sensitive function, and the occurrence of ‘physicist’ is one factor of context that affects just which degree is required. On Kennedy’s view, \( s \) is a function that chooses a standard of comparison such that ‘Feynman is tall’ or ‘Feynman is tall for a physicist’ is true of Feynman just in case in the context of utterance the degree to which Feynman is tall stands out, where ‘stands out’ is to be understood in purely distributional terms, (and not agent-relative or speaker-relative terms, in order to avoid the objections that Stanley (2003) raises against Graff (2000) and Boguslawski (1975)).
How then, if we keep the denotation of the positive null morpheme \([\text{Deg pos-}]\) as in (1.4), are we to account for the meaning of prepositionally modified constructions such as ‘tall for a physicist’? Kennedy notes that vague predicates seem to exhibit a surprising difference between when they are used to modify nominals, and when they are themselves modified by prepositional phrases: constructions of the form \([x \text{ is } [\text{DegP pos- } [\text{AP A for an NP }]]]]\) seem to presuppose that the referent of \(x\) is a member of the class expressed by the noun phrase. Thus, he claims that sentences such as:

(16)  
\begin{enumerate}
\item Kyle’s car is big for a Buick.
\item Kyle’s car is not big for a Buick.
\item Is Kyle’s car big for a Buick?
\end{enumerate}

all require Kyle’s car to be a Buick. He points to the following pairs of sentences as evidence, where the apparent dissonance of the first member seems due to presupposition failure.

(17)  
\begin{enumerate}
\item ? Kyle’s BMW is expensive for a Honda.
\item Kyle’s BMW is (really) an expensive Honda.
\end{enumerate}

(18)  
\begin{enumerate}
\item ? Kyle’s BMW is not expensive for a Honda.
\item Kyle’s BMW is (obviously) not an expensive Honda.
\end{enumerate}

(19)  
\begin{enumerate}
\item ? Is Kyle’s BMW expensive for a Honda?
\item Is Kyle’s BMW (actually) an expensive Honda?
\end{enumerate}

He predicts presupposition failure by positing that the adjectival phrase ‘tall’ has a variant ‘tall*’ such that:
which is just the same function as ‘tall’, except with its domain restricted to those things satisfying \( \lambda x, e. k(x) \). \([tall^*]\) thus combines with the denotation of the noun phrase contributed by a prepositional modifier. Keeping (11) the same, we get

\[
[[A \text{ tall}^*]] = \lambda k_{(e, t)} \lambda x_e : k(x). \text{tall}(x) 
\]  
(1.7)

Thus, (1.9) is a function defined only on physicists, and a sentence such as ‘Frege is tall for a physicist’ will result in a failure to compute a denotation, as expected in cases of presupposition failure. Rather than act as an input to \( S \), the comparison class is an input to a type-shifted adjectival phrase.

As far my project in this dissertation is concerned, the important feature of the denotations for ‘tall’ and ‘tall for a physicist’ in (1.5) and (1.9), respectively, are that \( s \) preserves the context-sensitivity of ‘tall’; we might think of it as a function from a context and a measure function to a threshold for application of the adjective. Or, we might think there are many different functions, one for each adjective: \( s_{\text{tall}}, s_{\text{short}}, \) and so on, each a function from a
context $c$ to a threshold $\theta$. Or, we might relativize the truth-value of ‘Feynman is tall’ to a context $c$, and let the threshold for tallness $\theta_{\text{tall}}$ be whichever one was the output of $s(c)$. Or, since it is the context that determines the thresholds, we might simply let the thresholds be the context, and have:

\[
[F\text{eynman is tall}]^{\theta_{\text{tall}},\theta_{\text{short}},\ldots} = \text{tall}(F\text{eynman}) \geq \theta_{\text{tall}} \tag{1.10}
\]

With the context sensitivity of the truth-conditions front and center, we can now ask, how is it that from a knowledge of these truth-conditions we can engage in the statistical inference patterns characteristic of vague predicates? It is to this question that I turn in Chapter 2.
Chapter 2

Against Threshold Semantics

2.1 Introduction

We saw in Chapter 1 a common semantics from Kennedy (2007) for both absolute and relative gradable adjectives, the latter of which I will also call vague predicates, as the absolute gradable adjectives arguably do not give rise to vagueness, or at least do so with much lower regularity. On this semantics, ‘Feynman is tall’ is true just in case Feynman’s degree of height is greater than or equal to some contextually determined threshold of application $\theta_{\text{tall}}$. For our purposes, this comes to:

$$[\text{Feynman is tall}]^{\theta_{\text{tall}},\theta_{\text{short}}, \ldots} = \text{tall}(\text{Feynman}) \geq \theta_{\text{tall}}, \quad (1.10)$$

where $\text{tall}$ is a function from objects to degrees of tallness. This does not address, however, a important question concerning relative gradable adjectives: how we as listeners are able to interpret usages of relative gradable adjectives to engage in the statistical inferences that we all clearly do engage in. Thus, from hearing someone called ‘tall’, I lower the probability that
they are 5’7”, and increase the probability that they are 6’1”. With no natural precisifications
at which we can place the contextually variant degree standard for tallness, we seem unable
to determine where the threshold might be, and thus how tall someone might be, when they
are called ‘tall’.

Lassiter and Goodman (2017) address this question by claiming that a receiver can engage
in a process of Gricean-inspired Bayesian reasoning to learn about both what Feynman’s
height might be and where the threshold for tallness might be. They posit a receiver who
estimates the probability that a sender would utter ‘Feynman is tall’, given that Feynman is
tall to degree $h$, and given that the threshold is $\theta$. This receiver also has a prior probability
distribution for Feynman’s degree of tallness—a normal distribution, as seems realistic—and
an uninformative, uniform prior probability distribution for the threshold for tallness. Then,
they engage in a process of Bayesian reasoning to calculate posterior probabilities for both
Feynman’s degree of tallness, which we may for now think of as his height, and the threshold
for tallness. Thus, on their model, receivers do not have to have prior information about
the distribution of thresholds in order to learn how tall Feynman might be upon hearing
‘Feynman is tall’; instead, pragmatic reasoning processes allows them to gain information
about both Feynman’s height and the thresholds for tallness.

After introducing their model of interpretation, I argue that it makes the wrong predictions
about the statistical inferences we engage in when we hear ‘Feynman is not tall’, given
the standard semantics for negation, on which ‘Feynman is not tall’ is true just in case
‘Feynman is tall’ is not true. After examining several different ways of addressing this
problem and finding all of them unsatisfactory, I show that we can predict the intuitively
correct patterns of statistical inference from their model, at least at lower levels of the sender-receiver pragmatic hierarchy, if instead of uninformative, uniform prior threshold distributions, we think of the receivers as having informative, non-uniform prior threshold distributions. However, having fixed prior threshold distributions suggests a simpler model of interpretation on which there are no thresholds; listeners consider only the probability that the speaker would say ‘Feynman is tall’ given that Feynman is a given height, and leverage a prior distribution for Feynman’s height to calculate a posterior distribution for his height. This then raises the question, what is the source of our knowledge of how these utterance probabilities vary with Feynman’s height? If we do not infer a threshold distribution on the basis of what we take to be our pre-existing common knowledge of the truth-conditions of our words, along with a characterization of the relevant features of the speaker’s psychology, then how do we know what the threshold distribution is like? I claim that we infer this from our observations of the encoding behavior of members of our linguistic community, but this first requires that the encoding behavior characteristic of how we use vague predicates can arise in the first place. I address this question in Chapter 3.

2.2 Bayes-Grice Models of Vagueness

Lassiter and Goodman (2017) attempt to address this question, (and offer some somewhat competing explanations of the origins of the features characteristic of vagueness), in terms of an iterated sender-receiver hierarchy. They posit a pragmatic receiver (or listener) $\rho_2$ who thinks they are receiving a message from a pragmatic sender $\sigma_1$ assumed (by $\rho_2$) to know the
correct answer \( h \) to the topic of conversation, or Question Under Discussion (Ginzburg 1995a, b)—for example, perhaps that topic is Feynman’s height \( h \). Drawing on Fox and Katzir (2011), they further specify that \( \rho_2 \) assumes that \( \sigma_1 \) chooses from some set of alternative utterances \( \text{ALT} \) which represent a subset of possible answers to the QUD; thus, \( \rho_2 \) thinks of \( \sigma_1 \) as choosing between uttering, say, *Feynman is tall, Feynman is short*, or nothing at all. \( \rho_2 \) also thinks of \( \sigma_1 \) as in turn assuming that \( \sigma_1 \) is speaking to a literal receiver \( \rho_0 \). Finally, at bottom \( \rho_0 \) updates their probability estimate on Feynman’s height in accordance with the truth-conditions for gradable adjectives that we see in Equation (1.10).

Let us say that \( u_1 = \text{Feynman is tall}, u_2 = \text{Feynman is short}, \) and the null utterance = \( u_0 \). For the sake of brevity, we will also use \( \theta_1 \) for \( \theta_{\text{tall}} \) and \( \theta_2 \) for \( \theta_{\text{short}} \). The probability density that \( \rho_0 \) assigns to Feynman’s height \( h \) upon hearing \( u_1 \) used with some given threshold value \( \theta_1 \) is derived by assigning 0 to \( h \) where \( h \) is less than \( \theta_1 \), and where \( h \) is greater than or equal to \( \theta_1 \), assigning to \( h \) his prior probability density for \( h \)—that is, \( \phi(h) \)—normalized by the cumulative probability distribution \( \phi(h) \) for \( h \geq \theta_1 \). That is, whatever the literal receiver thinks the threshold for *tall* is, upon hearing *Feynman is tall*, the literal receiver assumes there is no probability that Feynman is shorter than that threshold, and for every height equal to or greater than that threshold, assigns a new probability density to that height by dividing the old probability density by the total old probability mass above the threshold.

In the discrete case, suppose there were 5 possible heights \( h \), and \( \phi(h) \) was uniform for each height and thus \( \phi(h) = \frac{1}{5} \) for each \( h \). Then, if the literal receiver thought the threshold for *tall* was the middle height, then upon hearing *Feynman is tall*, the literal receiver would assign 0 probability to the two lowest heights, and then assign probability \( \frac{1}{3} = \frac{1}{5} \div \frac{3}{5} \) to the
other heights. Thus, we can say:

\[
\rho_0(h|u_1, \theta_1) = \phi'(h|h \geq \theta_1) = \begin{cases} 
\frac{\phi(h)}{\int_{\theta_1}^{\infty} \phi(h) dh} & \text{if } h \geq \theta_1 \\
0 & \text{otherwise}
\end{cases}
\] (2.1)

Similarly for \( u_2 = \text{Feynman is short} \), which is interpreted by \( \rho_0 \) as the mirror-image of \( u_1 \):\(^1\)

\[
\rho_0(h|u_2, \theta_2) = \phi'(h|h \leq \theta_2) = \begin{cases} 
\frac{\phi(h)}{\int_{-\infty}^{\theta_2} \phi(h) dh} & \text{if } h \leq \theta_2 \\
0 & \text{otherwise}
\end{cases}
\] (2.2)

The probability density that \( \rho_0 \) assigns to John’s height \( h \) upon hearing no utterance at all is the same as \( \rho_0 \)'s prior probability density \( \phi(h) \) for John’s height. Thus \( \theta_1 \) and \( \theta_2 \) drop out since \( u_0 \) is interpreted the same way regardless of \( \theta_1 \) or \( \theta_2 \):

\[
\rho_0(h|u_0) = \phi(h)
\] (2.3)

Next, given these definitions of how \( \rho_0 \) interprets, Lassiter and Goodman (2017) posit that \( \sigma_1 \) probabilistically chooses an utterance as a function of the utility of the utterance; in the case of \textit{tall}, assuming that \( \sigma_1 \) desires to be as informative to \( \rho_0 \) about Feynman’s height \( h \) as possible while balancing for the production cost of the message, the utility increases with increasing informativity about \( h \), and decreases with the production cost of the message. The informativities \textbf{INFO} of the utterances \( u_0, u_1, u_2 \), relative to a given height \( h \), (and

\(^1\)For reasons given below, the non-finite limits in (2.1) and (2.2) will be replaced by finite limits.
given threshold $\theta_1$ and $\theta_2$ in the case of $u_1$ and $u_2$, respectively), are thus defined as the natural log of the probability density that $\rho_0$ assigns to the height $h$ given the utterance and the corresponding threshold, if any:

\[
\text{INFO}(u_0, h) = \ln(\rho_0(h|u_0)) \quad (2.4a)
\]

\[
\text{INFO}(u_1, h, \theta_1) = \ln(\rho_0(h|u_1, \theta_1)) \quad (2.4b)
\]

\[
\text{INFO}(u_2, h, \theta_2) = \ln(\rho_0(h|u_2, \theta_2)) \quad (2.4c)
\]

To see the intuition behind these definitions, consider the informativity of $u_1$ in the case of a discrete probability distribution over heights: suppose that Feynman’s height is some particular value $h$. If $\rho_0(h|u_1, \theta_1)$ is close to 1, that means $\rho_0$ has received almost maximal information about Feynman’s height from hearing *Feynman is tall* and assuming that the threshold for *tall* is $\theta_1$: for $\rho_0$, Feynman’s height is almost certain to be $h$. Then $\ln(\rho_0(h|u_1, \theta_1))$ will be close to 0. Where $\rho_0(h|u_1, \theta_1)$ is close to 0, that means $\rho_0$ has received minimal information about Feynman’s height from hearing *Feynman is tall* and assuming that the threshold for *tall* is $\theta_1$: for $\rho_0$, Feynman’s height is almost certain to not be $h$. Then $\ln(\rho_0(h|u, \theta_1))$ will be close to $-\infty$. Thus for discrete probability distributions over heights $h$, for any given $h$, $\text{INFO}(u_1, h, \theta_1)$ is between $-\infty$ and 0, and $\text{INFO}(u_1, h, \theta_1)$ increases as
\(\rho_0\) assigns higher probability to \(h\) given that the threshold for \textit{tall} in \textit{Feynman is tall} is \(\theta_1\).\(^2\)\(^,\)\(^3\)

As for the production cost \(C(u)\) of the message, Lassiter and Goodman (2017) assume that it monotonically increases with the difficulty of articulation, and approximate the difficulty of articulation as the number of words in the message. They report plausible results with \(C(u)\) set to \(2/3\) the number of words, and thus fix it to that.

Given these definitions of informativity and cost we can define the utility \(U\) to \(\sigma_1\) of message \(u_1\) (\(u_2\)) used with threshold \(\theta_1\) (\(\theta_2\)) to encode height \(h\) as:

\[\text{INFO}(u_0, h), \text{INFO}(u_1, h, \theta_1), \text{INFO}(u_2, h, \theta_2), U_{\sigma_1}(u_0, h), U_{\sigma_1}(u_1, h, \theta_1), \text{and } U_{\sigma_1}(u_2, h, \theta_2)\text{ range between } -\infty \text{ and } \infty, \]

and so \(0 \leq \sigma_1^*(u_0|h), \sigma_1^*(u_1|h, \theta_1), \sigma_1^*(u_2|h, \theta_2) \leq \infty.\)

\(^3\)This notion of informativity might seem to confuse the amount of information with the quality of information; put another way, it might seem to confuse precision for accuracy. Or to put it yet another way, if Feynman’s height is \(h\) then if \(\rho_0(h|u_1, \theta_1)\) is close to 0, and thus \(\rho_0\) upon hearing \textit{Feynman is tall} and assuming that the threshold for \textit{tall} is \(\theta_1\) thinks that Feynman’s height is almost certain to \textit{not} be \(h\), then it might seem like \(u_1\) used with threshold \(\theta_1\) provides not near-minimal information, but instead lots and lots of information—just \textit{bad} information. However, consider that if \(\text{ALT} = \{u_0, u_1\}\) then if \(\rho_0(h|u_0) < \rho_0(h|u_1, \theta_1)\) then \(u_1\) used with \(\theta_1\) provides more information about \(h\) than \(u_0\), in the sense that \(\rho_0\) assigns higher probability to \(h\) upon hearing \textit{Feynman is tall} and assuming that the threshold for \textit{tall} is \(\theta_1\), than upon hearing nothing at all. Thus, it seems that we only assumed \(u_1\) used with \(\theta_1\) provides bad information, because we assumed that \(\rho_0\) already has better background information in the sense that \(\phi(h) > \rho_0(h|u_1, \theta_1)\); if this is not so, we do not think that \(u_1\) used with \(\theta_1\) is so mis-informative. Similar comments apply when \(\text{ALT} = \{u_0, u_1, u_2\}\).
\[ U_{\sigma_1}(u_1, h, \theta_1) = \text{INFO}(u_1, h, \theta_1) - C(u_1) \] (2.5a)

\[ U_{\sigma_1}(u_2, h, \theta_2) = \text{INFO}(u_2, h, \theta_2) - C(u_2) \] (2.5b)

Similarly, for \( u_0 \) we have

\[ U_{\sigma_1}(u_0, h) = \text{INFO}(u_0, h) - C(u_0) \] (2.6)

Thus, we see that the utility to \( \sigma_1 \) of a given message increases with informativity of the message, and decreases with the production cost of the message. Now, since \( \sigma_1 \) probabilistically chooses an utterance as a function of the utility of the utterance, with the probability of usage increasing as the utility increases, we can use a soft-max function

\[ \sigma^*_1(u_1, h, \theta_1) = e^{\lambda \times U_{\sigma_1}(u_1, h, \theta_1)} \] (2.7a)

\[ \sigma^*_1(u_2, h, \theta_2) = e^{\lambda \times U_{\sigma_1}(u_2, h, \theta_2)} \] (2.7b)

to define the non-normalized probability \( \sigma^*_1 \) of \( \sigma_1 \) using \( u_1 \) (\( u_2 \)) for any given height \( h \) and threshold \( \theta_1 \) (\( \theta_2 \)), for \( 0 \leq \lambda \leq \infty \). Here, \( \lambda \) is a Luce choice parameter, the effect of which is described below; in the meantime, to again see the intuition in the discrete case for \( u_1 \): since \( -\infty \leq U_{\sigma_1}(u, h, \theta_1) \leq 0 \), it follows that \( 0 \leq e^{\lambda \times U_{\sigma_1}(u, h, \theta)} \leq 1 \), since \( \lambda \) is non-negative.
and the utility is non-positive. In particular, assuming $\lambda > 0$, as $U_{\sigma_1}(u, h, \theta_1)$ decreases to $-\infty$, $e^{\lambda U_{\sigma_1}(u, h, \theta_1)}$ decreases to 0; as $U_{\sigma_1}(u, h, \theta_1)$ increases to 0, $e^{\lambda U_{\sigma_1}(u, h, \theta_1)}$ increases to 1. We similarly use a soft-max function to define the non-normalized probability of $\sigma_1$ using $u_0$ for any given height:

$$
\sigma_1^*(u_0, h) = e^{\lambda U_{\sigma_1}(u_0, h)}
$$

(2.8)

Then, the normalized probabilities that $\sigma_1$ will choose to encode state $h$ with message $u_0$, $u_1$, or $u_2$ given thresholds $\theta_1$ for $u_1$ and $\theta_2$ for $u_2$ are defined according to:

$$
\sigma_1(u_0| h, \theta_1, \theta_2) = \frac{\sigma_1^*(u_0, h)}{\sigma_1^*(u_0, h) + \sigma_1^*(u_1, h, \theta_1) + \sigma_1^*(u_2, h, \theta_2)}
$$

(2.9a)

$$
\sigma_1(u_1| h, \theta_1, \theta_2) = \frac{\sigma_1^*(u_1, h, \theta_1)}{\sigma_1^*(u_0, h) + \sigma_1^*(u_1, h, \theta_1) + \sigma_1^*(u_2, h, \theta_2)}
$$

(2.9b)

$$
\sigma_1(u_2| h, \theta_1, \theta_2) = \frac{\sigma_1^*(u_2, h, \theta_2)}{\sigma_1^*(u_0, h) + \sigma_1^*(u_1, h, \theta_1) + \sigma_1^*(u_2, h, \theta_2)}
$$

(2.9c)

Now we can see the effect of the Luce choice parameter $\lambda$: if $\lambda = 0$, $\sigma_1^* = 1$ for any message $u$, and hence $\sigma_1$ has equal chances of selecting any message no matter its utility. As $\lambda$ goes to $\infty$, $\sigma_1$ has probability 1 of choosing the utility maximizing message, no matter how small the difference in utility may be.

Next, upon hearing $u_0$, $u_1$, or $u_2$, receiver $\rho_2$ can infer the joint probability density of $h$, $\theta_1$, and $\theta_2$ by using Bayes’ rule, and assuming the independence of $h$ from $(\theta_1, \theta_2)$.
\[
\rho_2(h, \theta_1, \theta_2 | u_0) = \frac{\sigma_1(u_0 | h, \theta_1, \theta_2) \times \phi(h) \times p(\theta_1, \theta_2)}{\int_{a}^{b} \int_{a}^{b} \int_{a}^{b} \sigma_1(u_0 | h, \theta_1, \theta_2) \times \phi(h) \times p(\theta_1, \theta_2) \, dh \, d\theta_1 \, d\theta_2} \tag{2.10a}
\]

\[
\rho_2(h, \theta_1, \theta_2 | u_1) = \frac{\sigma_1(u_1 | h, \theta_1, \theta_2) \times \phi(h) \times p(\theta_1, \theta_2)}{\int_{a}^{b} \int_{a}^{b} \int_{a}^{b} \sigma_1(u_1 | h, \theta_1, \theta_2) \times \phi(h) \times p(\theta_1, \theta_2) \, dh \, d\theta_2 \, d\theta_1} \tag{2.10b}
\]

\[
\rho_2(h, \theta_1, \theta_2 | u_2) = \frac{\sigma_1(u_2 | h, \theta_1, \theta_2) \times \phi(h) \times p(\theta_1, \theta_2)}{\int_{h}^{b} \int_{a}^{b} \int_{a}^{b} \sigma_1(u_2 | h, \theta_1, \theta_2) \times \phi(h) \times p(\theta_1, \theta_2) \, dh \, d\theta_1 \, d\theta_2} \tag{2.10c}
\]

Note that the upper and lower limits for the outer integrals in the denominators of (2.10b) and (2.10c), respectively, are \(h\) because when \(h \leq \theta_1\), \(\sigma_1^*(u_1, h, \theta_1) = 0\) and when \(h \geq \theta_2\), \(\sigma_1^*(u_2, h, \theta_2) = 0\). Regarding \(p(\theta_1, \theta_2)\), Lassiter and Goodman think of \(\rho_2\) as assuming the uniformity and independence of \(\theta_1\) and \(\theta_2\); but in what follows we will explore diverging from this assumption. To allow for this we express \(p(\theta_1, \theta_2)\) as \(p(\theta_1 | \theta_2) \times p(\theta_2)\) or \(p(\theta_2 | \theta_1) \times p(\theta_1)\) in (2.11) - (2.13). As for the integral limits \(a, b\) specified in (2.10), and in (2.11)-(2.13): since we assume a uniform distribution for \(\theta_1\) and \(\theta_2\) we must restrict the ranges of integration for \(\theta_1\) and \(\theta_2\), and hence also for \(h\), to the finite range indicated in figure (2.1). This introduces no complications when \(\phi(h)\) is a Beta or uniform distribution since these are defined only over finite ranges anyway, but when \(\phi(h)\) is a normal distribution, it must be truncated to \((a, b)\).

Finally, we can then calculate \(\rho_2\)’s posterior probability distribution for \(h\) \((\theta_1)\) upon hearing \(u_1\) by marginalizing over \(\theta_1\) and \(\theta_2\) \((h\) and \(\theta_2)\):

\[\text{Note that in (2.11) we leave out the normalization factor from (2.10b) for brevity. Similarly for (2.12) and (2.10c), and (2.13) and (2.10a).}\]
\[
\begin{align*}
\rho_2(h|u_1) &\propto \int_a^b \int_a^b \sigma_1(u_1|h,\theta_1,\theta_2) \times \phi(h) \times p(\theta_2|\theta_1) \times p(\theta_1) \ d\theta_2 d\theta_1 \quad (2.11a) \\
\rho_2(\theta_1|u_1) &\propto \int_{\theta_1}^{b\theta_1} \int_a^b \sigma_1(u_1|h,\theta_1,\theta_2) \times \phi(h) \times p(\theta_2|\theta_1) \times p(\theta_1) \ d\theta_2 d\theta_1 \\
\rho_2(\theta_2|u_1) &\propto \int_a^{\theta_2} \int_a^b \sigma_1(u_1|h,\theta_1,\theta_2) \times \phi(h) \times p(\theta_2|\theta_1) \times p(\theta_1) \ d\theta_2 d\theta_1 \\
\rho_2(h|u_2) &\propto \int_a^{b\theta_2} \int_a^b \sigma_1(u_2|h,\theta_1,\theta_2) \times \phi(h) \times p(\theta_2|\theta_1) \times p(\theta_1) \ d\theta_2 d\theta_1 \\
\rho_2(\theta_2|u_2) &\propto \int_a^{\theta_2} \int_a^b \sigma_1(u_1|h,\theta_1,\theta_2) \times \phi(h) \times p(\theta_2|\theta_1) \times p(\theta_1) \ d\theta_2 d\theta_1 \\
\rho_2(h|u_0) &\propto \int_a^b \int_a^b \sigma_1(u_0|h,\theta_1,\theta_2) \times \phi(h) \times p(\theta_2|\theta_1) \times p(\theta_1) \ d\theta_2 d\theta_1 \quad (2.12) \\
\end{align*}
\]

Likewise, we can calculate \(\rho_2\)’s posterior probability distribution for \(h\) (\(\theta_2\)) upon hearing \(u_2\) by marginalizing over \(\theta_1\) and \(\theta_2\) (\(h\) and \(\theta_1\)):

Finally, we can calculate \(\rho_2\)’s posterior probability distribution for \(h\) upon hearing \(u_0\) by marginalizing over \(\theta_1\) and \(\theta_2\):

\[
\begin{align*}
\rho_2(h|u_0) &\propto \int_a^b \int_a^b \sigma_1(u_0|h,\theta_1,\theta_2) \times \phi(h) \times p(\theta_2|\theta_1) \times p(\theta_1) \ d\theta_2 d\theta_1 \\
\end{align*}
\]

With this model in mind we can return to Lassiter and Goodman (2017)’s claim that the features of relative gradable adjectives tend to arise when listeners (that is, \(\rho_2\)) have prior knowledge that the probability distribution of the relevant comparison class over the degree scale encoded by the adjective (that is, \(\phi(h)\)) has little to no probability mass on any endpoints—as when \(\phi(h)\) is a normal distribution with \(\mu, \sigma = 0, 1\), truncated to the range \(\pm 4\).
We see in Figure (2.1) the predictions of their model for such a distribution when $\lambda = 4$ and $C(u_1) = C(u_2) = 2$. Here, and in all other figures in this chapter where $\phi(h)$ is a normal distribution, the $x$-axis represents the standard deviation, and the $y$-axis represents the probability density. Since the lines all represent probability density functions, the areas under the curves sum to 1. To get a more specific intuition about these scales, consider that the average height of an adult American male is 5’9”, and the standard deviation is approximately 3”; thus roughly 95.45% of adult American males lie between $-2$ and 2 on the $x$-axis in Figure (2.1).

If we think of $u_1$ and $u_2$ as *Feynman is tall* and *Feynman is short*, respectively, and the relevant comparison class as adult men, the model seems to make sensible predictions about actual receivers' interpretations of relative gradable adjectives: Upon hearing $u_1$, $\rho_2$ does not have a precise value for Feynman’s height, nor a sharp upper- or lower-bound for its probability distribution; instead, $\rho_2$’s probability density function for Feynman’s height shows gradual monotonic increase and decrease to and from some global maximum. Furthermore, varying the $\mu$ and $\sigma$ of the relevant comparison class will afford $\rho_2$ varying information about the height of the object; they thus seem to capture the context-sensitivity of relative gradable adjectives. Finally, the existence of borderline cases seems to be a straightforward consequence of $\rho_2$’s posterior distribution for $\theta_1$: an object whose height lies close to the maximum of $\rho_2(\theta_1 | u_1)$ is borderline ‘tall’.
2.3 Negation and Relative Gradable Adjectives

However, I think there are two problems with their explanation of the features of relative gradable adjectives, both of which are related to negation. First, their explanation predicts that listeners will make certain kinds of inferences involving negation of adjectives and the antonyms of those same adjectives, when in fact listeners do not make those inferences. That is, their model fails to predict certain kinds of inference failures. Second, their explanation wrongly predicts that the negation of a relative gradable adjective will be interpreted equally strongly as, but opposite to, how the original adjective will be interpreted.

2.3.1 Unexpected Inferences

While Lassiter and Goodman do not discuss how to apply their model to an explanation of the entailment patterns of relative gradable adjectives, I think we can see how at least some
of it would have to go. Consider the failure of the inference from ‘Feynman is not tall’ to
‘Feynman is short.’ Capturing this pattern requires that we modify our model to account for
negation; to this end we can extend Lassiter and Goodman’s work by assuming that negation
forces \( \rho_0 \) to re-normalize the prior probability mass below (above) the threshold. That is,
we propose that \( \rho_0 \) interprets ‘Feynman is not tall’ and ‘Feynman is not short’, respectively,
as in (2.14) and (2.15):

\[
\rho_0(h|\neg u_1, \theta_1) = \phi(h|h < \theta_1) = \begin{cases} \frac{\phi(h)}{\int_{-\infty}^{\theta_1} \phi(h)dh} & \text{if } h < \theta_1 \\ 0 & \text{otherwise} \end{cases}
\] (2.14)

\[
\rho_0(h|\neg u_2, \theta_2) = \phi(h|h > \theta_2) = \begin{cases} \frac{\phi(h)}{\int_{\theta_2}^{\infty} \phi(h)dh} & \text{if } h > \theta_2 \\ 0 & \text{otherwise} \end{cases}
\] (2.15)

The informativities are then defined as:

\[
\text{INFO}(\neg u_1, h, \theta_1) = \ln(\rho_0(h|\neg u_1, \theta_1))
\] (2.16a)

\[
\text{INFO}(\neg u_2, h, \theta_2) = \ln(\rho_0(h|\neg u_2, \theta_2))
\] (2.16b)

and the cost, utility, and non-normalized sending probability for \( \neg u_1 \) and \( \neg u_2 \) are defined
similarly as above.
\[ \sigma_1^*(-u_1|h, \theta_1) = e^{\lambda \times (\ln(\rho_0(h|\neg u_1, \theta_1)) - C(-u_1))} \]  

(2.17a)

\[ \sigma_1^*(-u_2|h, \theta_2) = e^{\lambda \times (\ln(\rho_0(h|\neg u_2, \theta_2)) - C(-u_2))} \]  

(2.17b)

The normalized sending probabilities \( \sigma_1(-u_1, h, \theta_1, \theta_2) \) and \( \sigma_1(-u_2, h, \theta_1, \theta_2) \) are then:

\[
\sigma_1(-u_1|h, \theta_1, \theta_2) = \frac{\sigma_1^*(-u_1, h, \theta_1)}{\sigma_1^*(u_0, h) + \sigma_1^*(u_1, h, \theta_1) + \sigma_1^*(u_2, h, \theta_2) + \sigma_1^*(-u_1, h, \theta_1) + \sigma_1^*(-u_2, h, \theta_2)} \]  

(2.18a)

\[
\sigma_1(-u_2|h, \theta_1, \theta_2) = \frac{\sigma_1^*(-u_2, h, \theta_2)}{\sigma_1^*(u_0, h) + \sigma_1^*(u_1, h, \theta_1) + \sigma_1^*(u_2, h, \theta_2) + \sigma_1^*(-u_1, h, \theta_1) + \sigma_1^*(-u_2, h, \theta_2)} \]  

(2.18b)

The normalized \( \sigma_1 \) sending probabilities for \( u_0, u_1, \) and \( u_2 \) are similarly defined. Then, the posterior probability density functions \( \rho_2(h, \theta_1, \theta_2|\neg u_1) \) and \( \rho_2(h, \theta_1, \theta_2|\neg u_2) \) are defined as in (2.19a) and (2.19b):

\[
\rho_2(h, \theta_1, \theta_2|\neg u_1) = \frac{\sigma_1(-u_1, h, \theta_1, \theta_2) \times \phi(h) \times p(\theta_1, \theta_2)}{\int_a^b \int_a^b \int_a^b \sigma_1(u_1|h, \theta_1, \theta_2) \times \phi(h) \times p(\theta_1, \theta_2) \, d\theta_2 dh \, d\theta_1} \]  

(2.19a)

\[
\rho_2(h, \theta_1, \theta_2|\neg u_2) = \frac{\sigma_1(-u_2, h, \theta_1, \theta_2) \times \phi(h) \times p(\theta_1, \theta_2)}{\int_a^b \int_a^b \int_a^b \sigma_1(u_2|h, \theta_1, \theta_2) \times \phi(h) \times p(\theta_1, \theta_2) \, d\theta_1 dh \, d\theta_2} \]  

(2.19b)

The posterior probability density functions \( \rho_2(h, \theta_1, \theta_2|u_0), \rho_2(h, \theta_1, \theta_2|u_1), \) and \( \rho_2(h, \theta_1, \theta_2|u_2) \) remain defined as in (2.10a), (2.10b), and (2.10c), respectively. Note that the model does
not require additional thresholds $\theta_3, \theta_4$ for $\neg u_1, \neg u_2$, so we can continue to normalize and marginalize over only $h, \theta_1, \theta_2$.

Consider then relative antonym pairs such as short and tall; let us assume a context in which Feynman’s height is the QUD and $\rho_2$ assumes that $\sigma_1$ is choosing between Feynman is short, Feynman is not short, Feynman is tall, Feynman is not tall, or saying nothing at all (that is, $\text{ALT} = \{u_0, u_1, \neg u_1, u_2, \neg u_2\}$). When we hear Feynman is not tall we do not infer that he is as short as when we hear Feynman is short, and similarly for Feynman is not short and Feynman is tall. However, this does not seem corroborated by the model, insofar in Figure (2.2), where again $\lambda = 4$, $\rho_2$’s posterior probability distributions for $\neg u_1$ and $\neg u_2$ are actually more extreme than for $u_2$ and $u_1$, respectively, regardless of whether $C(u) = \frac{1}{3} \times \text{length}(u)$ or $C(u) = \frac{4}{3} \times \text{length}(u)$, or whether $\lambda = 4$ or $\lambda = 8$. Thus, the model predicts that from not tall, we can infer short, and from not short we can infer tall. In fact, Figure (2.2) indicates that the model predicts that upon hearing someone called not short, we would infer that they are taller than we would upon hearing them called tall, and likewise for not tall and short, (given that the cost of uttering the first is higher than the second).

### 2.3.2 Equal and Opposite Interpretations

To see the second problem, contrast what we might infer about Feynman’s height from hearing Feynman is not tall versus Feynman is tall: given my background knowledge of the height distribution of American males, in particular that the average is about 5’9”, upon hearing the latter I might infer he is probably around 6’0”, give or take a couple of inches. Upon hearing the latter, I might infer that he is probably around 5’8”, again give or take a
couple of inches. Notice that the peak probability in the latter case is closer to the average than in the former case; this is the sense in which tall and not tall are interpreted in opposite and unequal manners: hearing someone called not tall seems to merely sharply decrease the proportion of probability mass at the upper end of the distribution, leaving a long tail as we move down the scale of heights, whereas hearing someone called tall seems to sharply decrease the proportion of probability mass at the lower and middle parts of the distribution, effectively moving the distribution up the scale of heights. In terms of the model, this means that $\rho_2(h|u_1)$ should peak closer to the average of the prior distribution than $\rho_2(h|\neg u_1)$, with a long tail towards the bottom of the height scale. Again assume that Feynman’s height is the QUD and $\text{ALT} = \{u_0, u_1, \neg u_1\}$; perhaps you know that I am looking to round out my intramural basketball team with someone able to rebound and block shots and am would thus much prefer, all things being equal, a tall player. You might recommend someone who is tall, but not particularly athletic, or you might recommend someone who is not tall, but explosively athletic and intensely competitive. Again, $\text{ALT}$ does not have to be realistic,
only realistic for a listener to assume a speaker chooses from; nonetheless, it seems the simplest way to get such an ALT is for ALT to be realistic, lest we have to posit other mechanisms to account for why we do not systematically miscommunicate.

To define the posterior $\rho_2(h|\neg u_1)$ given this ALT we first define the normalized sending probability as

$$\sigma_1(\neg u_1|h,\theta_1) = \frac{\sigma_1^*(\neg u_1, h, \theta_1)}{\sigma_1^*(u_0, h) + \sigma_1^*(u_1, h, \theta_1) + \sigma_1^*(\neg u_1, h, \theta_1)} \quad (2.20)$$

and then marginalize over $\theta_1$ in the posterior probability density function:

$$\rho_2(h|\neg u_1) = \frac{\int_a^b \sigma_1(\neg u_1, h, \theta_1) \times \phi(h) \times p(\theta_1) d\theta_1}{\int_a^b \int_a^b \int_a^b \sigma_1(\neg u_1, h, \theta_1) \times \phi(h) \times p(\theta_1) dhd\theta_1} \quad (2.21)$$

(Note that for ALT = \{u_0, u_1, \neg u_1\}, $\sigma_1(u_0, h, \theta_1)$ and $\sigma_1(u_1, h, \theta_1)$ and their corresponding posterior probability densities and marginalizations are similarly defined as in (2.20) and (2.21).)

The problem, as we see in Figure (2.3), is that $\rho_2(h|\neg u_1)$ actually peaks farther from the center of the prior distribution than the $\rho_2(h|u_1)$, with no longer of a tail towards the bottom than $\rho_2(h|u_1)$ has towards the top: again, this is so regardless of whether $C(u) = \frac{1}{3} \times length(u)$ or $C(u) = \frac{4}{3} \times length(u)$, or whether $\lambda = 4$ or $\lambda = 8$. 43
2.4 Solutions Within the Model

2.4.1 Varying the Parameters

I think we can see from Figures (2.2) and (2.3) that the cost roughly determines the location of \( \rho_2 \)'s maximum posterior probability density for \( h \) given an utterance \( u \), analogous to determining the mean of a normal distribution; furthermore, the faster the cost increases as a function of the length of the utterance, the faster the negation of an adjective’s antonym moves towards the extremes of the degree scale, compared to the adjective itself. Thus in Figure (2.2) we see that regardless of whether \( \lambda = 4 \) or \( \lambda = 8 \), when \( C(u) = \frac{4}{3} \times \text{length}(u) \), \( \neg u_2 \) has a much more extreme meaning than \( u_1 \); not as much so when \( C(u) = \frac{1}{3} \times \text{length}(u) \).

In fact, with regards to \( \text{ALT} = \{ u_0, u_1, u_2, \neg u_1, \neg u_2 \} \), we might note that our definition of \( \rho_0(h|\neg u_1, \theta_1) \) is equivalent to the definition of \( \rho_0(h|u_2, \theta_2) \) except for having substituted \( < \) for \( \leq \) (which substitution was necessitated by the intuition that nothing can be both ‘tall’
and ‘not tall’). Similar remarks apply to \( \rho_0(h|\neg u_2, \theta_2) \) and \( \rho_0(h|u_1, \theta_1) \). This has the result that if \( C(\neg u_1) = C(u_2) \) and \( C(u_1) = C(\neg u_2) \), \( \sigma_1(\neg u_1, h, \theta_1, \theta_2) \) for \( \theta_1, \theta_2 = i, j \) must equal \( \sigma_1(u_2, h, \theta_1, \theta_2) \) for \( \theta_1, \theta_2 = j, i \), so long as \( h \neq i, j \). Since \( \rho_2 \) assumes \( \theta_1 \) and \( \theta_2 \) are uniformly distributed over identical ranges, it follows that \( \rho_2(h|\neg u_1) \) will be identical to \( \rho_2(h|u_2) \) and \( \rho_2(h|\neg u_2) \) will be identical to \( \rho_2(h|u_1) \). Thus, even if we assume that \( C(\neg u_1) = C(u_2) \) and \( C(u_1) = C(\neg u_2) \), the best we could do is to predict that from not tall we infer the same thing as we infer from short, and likewise for not short and tall. This points up that given this ALT, even if we assumed a non-linear cost function such that each additional word in a sentence imposed a decreasing marginal cost, we would still not get the desired effect, even at the limit where the marginal cost of ‘not ’ is nothing. Predicting that not tall has a weaker meaning than short, as it actually seems to, requires that the former have a lower cost than the latter, even though it is longer—in short, it requires that ‘not’ have a negative cost. Similar remarks apply when \( \text{ALT} = \{u_0, u_1, \neg u_1\} \): even if we assume \( C(\neg u_1) = C(u_1) \) so that the cost of ‘not’ is nothing, we get the results seen in Figure (2.4) where the costs of \( u_0, u_1 \) are fixed to 0, 1, respectively, and the choice parameter varies as before: ‘not tall’ will have an equally strong, but opposite interpretation, as ‘tall’, so long as the cost of ‘not’ is 0. It is only by fixing the cost of ‘not’ to \(-\frac{2}{3}\) that we find ‘not tall’ is interpreted in a more plausible manner.

Nor will it help to alter the choice parameter. In Figure (2.5) we see that as the choice parameter increases exponentially, \( \rho_2 \)’s maximum posterior densities for all non-null signals moves increasingly slowly towards the extremes; as the choice parameter decreases to 0, \( \rho_2 \)’s posterior probability distribution approximates that of the prior probability distribution,
since when \( \lambda = 0 \), \( \sigma_1(u) \) is a constant function for all messages \( u \), for any \( \text{ALT} \).\(^5\) No matter the value of \( \lambda \), however, so long as one signal is more expensive than another, the first will be pushed farther to the extreme than the other, whether in the same direction, as is the problem with \( \neg u_1 \) versus \( u_2 \) and \( \neg u_2 \) versus \( u_1 \) when \( \text{ALT} = \{ u_0, u_1, u_2, \neg u_1, \neg u_2 \} \), or in opposite directions, as is the problem with \( \neg u_1 \) versus \( u_1 \) when \( \text{ALT} = \{ u_0, u_1, \neg u_1 \} \).

\(^5\)Whether there is a limit as \( \lambda \) approaches inf remains undetermined.
Another alternative is to allow $p(\theta_1)$ or $p(\theta_2)$, or both, to be either non-uniform, or dependent, or both. (There are uniform dependent distributions, like the heads/non-heads and tails/non-tails probabilities of a single coin.) But this leads to the question of why there would be the right kind of non-uniform probabilities, and the natural answer seems like the stimulus generalization picture.

### 2.4.2 Alternative Utterances

If varying the parameters will not solve either problem, what else can be done without modifying the model? Perhaps one option to address our problem with ‘short’ and ‘not tall’ is to differentiate the set of alternative utterances $\text{ALT}$ used in comparing ‘short’ and ‘not tall’: ‘not tall’ is normally only contrasted with ‘tall’. The idea would be that the natural context to our hearing ‘not tall’ is when the set of alternative utterances available to the speaker is limited to either ‘tall’ or ‘not tall’. It would, after all, seem odd if after asking someone to describe John, they replied that ‘John is not tall, bespectacled, and with a touch of gray in his hair.’ Perhaps the oddness is due to the fact that normally, when a speaker knows enough about someone’s height to include it in a description of them, they know enough to call them ‘tall’ or ‘short’; if they say nothing at all, we assume they lack that information.\(^6\) In any case, it seems more natural to hear someone described as ‘not tall’ when the speaker is first asked ‘Is John tall?’ (There are counterexamples such as ‘not well’, ‘not

---

\(^6\)Another possibility is that they know enough to call them ‘of roughly average height’, or some such description. I suspect this will not appreciably change the results of the model, and so leave this unexplored for now.
right’, and so on, but these seem to be lexicalized.) Thus, when we ask ourselves whether if from hearing something called ‘not tall’ we can infer that it is ‘short’, we are comparing what we would infer about something’s height from a speaker choosing to call it ‘not tall’ when the speaker is choosing from among ‘tall’, ‘not tall’, or saying nothing at all, to what we would infer about something’s height from a speaker choosing to call it ‘short’ when the speaker is choosing from among ‘short’, ‘tall’, or saying nothing at all.\(^7\)

In terms of the model this means that we ought to compare \(\rho_2\)’s posterior probability density function for \(h\) upon hearing ‘not tall’—that is, \(\rho_2(h|\neg u_1)\)—given that \(\text{ALT} = \{u_0, u_1, \neg u_1\}\), with \(\rho_2(h|u_2)\) given that \(\text{ALT} = \{u_0, u_1, u_2\}\), as in the original model. As we see in Figures (2.3) and (2.6), the problem with this option is that when \(\sigma_1\) has to choose among just ‘tall’, ‘not tall’, or nothing at all, \(\rho_2\) will actually interpret ‘not tall’ in a stronger manner than how \(\rho_2\) would interpret ‘short’ when \(\sigma_1\)’s choices are ‘short’, ‘tall’, or nothing at all, given the same linear cost function and choice parameter under both \(\text{ALTs}\), and assuming the uniformity of \(\theta_1\) and \(\theta_2\); thus, in order to obtain the desired effect, we again have to make the cost of ‘not’ negative. This seems due to the fact that \(\neg u_1\) is longer than \(u_2\), as numerical results when we fix \(C(\neg u_1) = C(u_2)\) suggest that \(\rho_2(h|\neg u_1)\) when the \(\text{ALT} = \{u_0, u_1, \neg u_1\}\) is in fact identical to \(\rho_2(h|u_2)\) when the \(\text{ALT} = \{u_0, u_1, u_2\}\).\(^8\) In fact, if we think that when \(\text{ALT} = \{u_0, u_1, \neg u_1\}\), there are two thresholds, one for \(u_1\) and one for \(\neg u_1\), so that we think of ‘tall’ and ‘not tall’ as partially defined predicates along the lines of Soames et al.

---

\(^7\)I discuss below the possibility of eliminating the null signal from the first case.

\(^8\)We suspect a proof is possible by noting that \(\rho_0(h|\neg u_2, \theta_2)\) in (2.15) is the same as \(\rho_0(h|u_1, \theta_1)\) in (2.1) and \(\rho_0(h|\neg u_1, \theta_1)\) in (2.14) is the same as \(\rho_0(h|u_2, \theta_2)\) in (2.2), except for the appropriate substitutions of \(\theta_1\) for \(\theta_2\) and \(>\) or \(<\) for \(\ge\) or \(\le\).
Figure 2.6: Lassiter and Goodman Model Three Signals Normal Distribution $\lambda = 4., 8$. Low Cost High Cost

(1999), then the threshold for $\neg u_1$ is equivalent to the threshold for $u_2$, and $\sigma_1(-u_1, h, \theta_1, \theta_2)$ and $\rho_2(h, \theta_1, \theta_2 | -u_1)$ when $\text{ALT} = \{u_0, u_1, -u_1\}$ will be identical to $\sigma_1(u_2, h, \theta_1, \theta_2)$ and $\rho_2(h, \theta_1, \theta_2 | u_2)$ when $\text{ALT} = \{u_0, u_1, u_2\}$, assuming $C(-u_1) = C(u_2)$. Note also that Figures (2.3) and (2.6) indicate that attempting to solve the second problem by comparing $\rho_2(h|u_1)$ given that $\text{ALT} = \{u_0, u_1, u_2\}$, with $\rho_2(h|-u_1)$ given that $\text{ALT} = \{u_0, u_1, -u_1\}$, will not work, as $\rho_2(h|u_1)$ in Figure (2.6) is weaker (even if only slightly) than $\rho_2(h|-u_1)$ in Figure (2.3).

With regards to the first problem, perhaps when we ask ourselves whether if from hearing something called ‘not tall’ we can infer that it is ‘short’, we are drawing upon what we would infer about something’s height from a speaker choosing to call it ‘not tall’ when the speaker is choosing from among ‘tall’ or ‘not tall’, instead of from among ‘tall’, ‘not tall’ or saying nothing at all, and comparing that with what we would infer about that object’s height from a speaker choosing to call it ‘short’ when the speaker is choosing from among ‘short’, ‘tall’, or saying nothing at all? Plausibly, the context to hearing ‘not tall’ is when
one has asked, or been asked, the question ‘Is John tall?’ In this case perhaps the null signal is not an available choice, since one is expected, after all, to answer the question. In terms of the model, then, the correct comparison would be between $\rho_2(h|\neg u_1)$ given that $\text{ALT} = \{u_1, \neg u_1\}$ and $\rho_2(h|u_2)$ given that $\text{ALT} = \{u_0, u_1, u_2\}$, instead of between $\rho_2(h|\neg u_1)$ given that $\text{ALT} = \{u_0, u_1, \neg u_1\}$ and $\rho_2(h|u_2)$ given that $\text{ALT} = \{u_0, u_1, u_2\}$. In this case where $\text{ALT} = \{u_1, \neg u_1\}$ there is no null signal to push the interpretation of $\neg u_1$ so far towards the low end of the height scale, so we should get a weaker interpretation of $\neg u_1$ than when $\text{ALT} = \{u_0, u_1, u_2\}$, under similar model settings. We might even hope that this would simultaneously solve the problem with ‘not tall’ and ‘tall’, by weakening the interpretation of ‘not tall’.

The problem is that when $\text{ALT} = \{u_1, \neg u_1\}$, we get implausibly weak readings of ‘tall’ and ‘not tall’, as we see in Figure (2.7). Here we vary the lower and upper bounds of the normal distribution in order to highlight something that seems even more problematic: when $h \geq \theta_1$ the denominator of $\sigma_1(u_1, h, \theta_1)$ is simply $\sigma_1^*(u_1, h, \theta_1)$, since in that case $\sigma_1^*(-u_1, h, \theta_1) = 0$ and there is no null signal. *Mutatis mutandis* for $\sigma_1(-u_1, h, \theta_1)$. Then when $h \geq \theta_1$, $\sigma_1(u_1, h, \theta_1) = 1.$ and $\sigma_1(-u_1, h, \theta_1) = 0$, and *vice-versa* when $h < \theta_1$. Then assuming the uniform probability density of $\theta_1$ it follows that

$$\rho_2(h|\neg u_1) \propto \int_a^h \sigma_1(-u_1, h, \theta_1) \times \phi(h) \times p(\theta_1) d\theta_1$$

$$\propto \frac{\phi(h)}{(b-a)} \times \int_a^h \sigma_1(-u_1, h, \theta_1) d\theta_1$$

(2.22)
and similarly

\[ \rho_2(h|u_1) \propto \phi(h) \times \frac{((b - a) - h)}{(b - a)}, \]  

(2.23)

so that \( \lambda \) and \( C(u) \) drop out of the system, and the separation of \( h \) posteriors is an artifact of the lower and upper bounds, which were imposed only in order to allow uniform distributions for \( \theta \) values.\(^9\) Even if we ignore this and hope that by moving \( a \) closer to \( \mu \) and \( b \) farther away, we could achieve the desired shapes for \( \rho_2(h|u_1) \) and \( \rho_2(\neg u_1) \) such that the former is farther from the mean and the latter is closer to the mean, there seem to be no such values that give plausible interpretations. Consider fixing \( a \approx -1 \) and \( b \approx 4 \), thereby making \((b - a)\) small enough to attain the required separation of interpretations, while still having \( u_1 \) mean something stronger than \( \neg u_1 \): this requires truncating the low end of the distribution much too close to the mean; no one thinks the height of persons is limited to one standard deviation below the mean, and we can’t make thinking so a condition on a plausible interpretation of an answer to the question ‘Is John tall?’

Similar comments apply to an attempt to solve either of the above problems by appeal to an \( \text{ALT} = \{u_0, \neg u_1\} \): in order to make such a strategy work, we would have to maintain that the cost coefficient (or the output of the cost function, if we opt for a non-linear cost function) is always lower when \( \text{ALT} = \{u_0, \neg u_1\} \) than when \( \text{ALT} = \{u_0, u_1\} \) or \( \text{ALT} = \{u_0, u_2\} \) and so \( \rho_2(h|\neg u_1) \) is always weaker than either \( \rho_2(h|u_2) \) (in the same direction) or \( \rho_2(h|u_1) \) (in

\(^9\)We leave out the normalizing denominator to simplify the presentation.
the opposite direction). We would have to further stipulate that \textbf{ALT} never contains \(\neg u_1\) and either \(u_1\) or \(u_2\). Even if we ignore the apparent artificiality of these stipulations, this strategy would involve making \(\rho_2(h, \theta_n|u_n)\) lexicalized—contrary to the assumption that \(\theta_n\) is contextually determined.

The fundamental problem, it seems, is that there is an asymmetry between a relative gradable adjective and its negation such that the central tendency of the distribution resulting from interpretation of the latter is closer than the central tendency of the distribution resulting from interpretation of the former is, to the central tendency of the prior distribution, (and similarly for its antonym, but on same, instead of opposite, sides of the central tendency of the prior distribution); this asymmetry does not seem able to be captured by merely altering the parameters of the foregoing model, nor by varying the set of alternative utterances that the receiver thinks the speaker is choosing from, as we have seen. Now, it is important to note that all of these mechanisms are, I think, part of what has traditionally been considered part of the pragmatics, and not the semantics. And it is instructive that a common assumption
driving all of these results is that $\theta_1$ and $\theta_2$ are uniformly distributed, as befitting the assumption that these values are pragmatically determined, and not semantically specified in advance. Nonetheless, it seems like a non-uniform probability distribution for $\theta$ values is a promising avenue for solving our problems, and we might thus attempt to derive non-uniform distributions for $\theta$ values while staying on the right side of the pragmatics-semantics divide. It is to such approaches that I turn next.

2.5 Extending the Model by Pragmatic Ascent, Part I

2.5.1 Abstract Scale Structures and Non-Uniform Threshold Priors

As noted above, when $\text{ALT} = \{u_0, u_1, u_2\}$ Lassiter and Goodman (2017) assume that $\rho_2$ has a uniform prior distribution for both $\theta_1$ and $\theta_2$, in order to reflect that the listener has ‘no relevant background knowledge about the resolution of free variables’. Lassiter and Goodman (2017) further justify uniform priors for at least scalar adjectives on the grounds of the preservation of interpretive flexibility: if $\theta_{\text{tall}}$ were biased towards the heights of adult American males, then $\text{tall skyscraper}$ would end up improperly interpreted, with implausibly weak meanings. I want to think of the scales as, in some sense, abstractions from actual measures of physical quantities, so the requisite interpretive flexibility may still be available even if we have non-uniform prior distributions for $\theta_1$ and $\theta_2$. Thus, I think that the degree scale is a scale of degrees of deviation from some measure of central tendency; since the
central tendency for the heights of skyscrapers is much greater than that for the heights of adult American males, if the most probable location for the threshold for ‘tall’ is 1 standard deviation above the central tendency, then for adult American males the threshold is most likely around 6’0”, while for skyscrapers it will be much higher.

While thinking of degree scales as measures of physical quantities might help ensure that the denotations of scalar adjectives are tightly connected to the truth conditions of the sentences in which they occur, I nonetheless think it is a mistake to think of scales as measures of physical quantities, since words like ‘fast’ are relative, but would have upper-end closed scales if it stood for actual velocity, even velocity relative to a reference frame, since velocity in all inertial reference frames is bounded by the speed of light. Similar remarks apply to words like ‘cold’ and physical quantities like temperature: ‘cold’ acts like a relative gradable adjective, but temperature has an absolute zero. If anything, these facts seem to reflect that the concepts behind words like ‘fast’ and ‘cold’ were formed prior to our knowledge of relativistic physics or modern thermodynamics.

Thus absent at least part of their motivation, how might we imagine a listener as having pragmatically derived non-uniform prior distributions for $\theta_1$ and $\theta_2$? One possibility is that we think of interpretation as happening at the level of $\rho_4$ instead of $\rho_2$. As we see in Figure (2.1), when $\text{ALT} = \{u_0, u_1, u_2\}$, given $u_1$, $\rho_2$ seems to have a posterior distribution for $\theta_1$ of the intuitively correct shape and location: the standard for ‘tall’ is very unlikely to be at or below the average height of the objects in the relevant comparison class, becoming monotonically increasingly probable as we reach a little less than 2 standard deviations above the average height, and then monotonically decreasing from there up. No one, after
all, thinks that the standard for tallness could be less than the average height; likewise no one thinks that the standard for tallness could exclude 99.999999% of the population. Similarly for \( \rho_2 \)'s posterior distribution for \( \theta_2 \) given \( u_2 \). Thus, if the \( \rho_2 \) posterior distributions for \( \theta_1 \) and \( \theta_2 \) become the \( \rho_4 \) prior distributions for \( \theta_1 \) and \( \theta_2 \), we might hope that the resulting \( \rho_4 \) posterior distributions for \( h \) given \( u_1, u_2, \neg u_1, \neg u_2 \) have the appropriate distributions to solve our two problems.

Now, implementing this requires that we define \( \sigma_3(u_n|h, \theta_1, \theta_2) \). How to do so? Remember that \( \rho_2 \) thinks of \( \sigma_1 \) as thinking of \( \rho_0 \) as a literal receiver who given a non-null utterance assigns a probability to a height conditional upon a given threshold value; thus we had

\[
\text{INFO}(u_0, h) = \ln(\rho_0(h|u_0)) \quad (2.4a \text{ revisited})
\]
\[
\text{INFO}(u_1, h, \theta_1) = \ln(\rho_0(h|u_1, \theta_1)) \quad (2.4b \text{ revisited})
\]
\[
\text{INFO}(u_2, h, \theta_2) = \ln(\rho_0(h|u_2, \theta_2)) \quad (2.4c \text{ revisited})
\]

Thus, in order to remain consistent with the original model, we can think of \( \sigma_3 \) as marginalizing out \( \theta_1 \) and \( \theta_2 \) for \( u_0 \), \( \theta_2 \) for \( u_1 \) and \( \neg u_1 \), and \( \theta_1 \) for \( u_2 \) and \( \neg u_2 \):
\[\rho_2(h|u_0) = \int_a^b \int_a^b \rho_2(h, \theta_1, \theta_2|u_0) \, d\theta_1 \, d\theta_2 \] (2.25a)

\[\rho_2(h, \theta_1|u_1) = \int_a^b \rho_2(h, \theta_1, \theta_2|u_1) \, d\theta_2 \] (2.25b)

\[\rho_2(h, \theta_2|u_2) = \int_a^b \rho_2(h, \theta_1, \theta_2|u_2) \, d\theta_1 \] (2.25c)

\[\rho_2(h, \theta_1|\neg u_1) = \int_a^b \rho_2(h, \theta_1, \theta_2|\neg u_1) \, d\theta_2 \] (2.25d)

\[\rho_2(h, \theta_2|\neg u_2) = \int_a^b \rho_2(h, \theta_1, \theta_2|\neg u_2) \, d\theta_1 \] (2.25e)

This gets us back \(\rho_2(h|u_0)\), just as we had \(\rho_0(h|u_0)\), but for \(u_1, \neg u_1\) and \(u_2, \neg u_2\) we need to think of \(\sigma_3\) as also normalizing by \(\theta_2\) and \(\theta_1\), respectively, to get back \(\rho_2(h|u_1, \theta_1), \rho_2(h|\neg u_1, \theta_1), \rho_2(h|u_2, \theta_2), \rho_2(h|\neg u_2, \theta_2)\):

\[\rho_2(h|u_1, \theta_1) = \frac{\rho_2(h, \theta_1|u_1)}{\int_a^b \int_a^b \rho_2(h, \theta_1, \theta_2|u_1) \, dh \, d\theta_2} \] (2.26a)

\[\rho_2(h|u_2, \theta_2) = \frac{\rho_2(h, \theta_2|u_2)}{\int_a^b \int_a^b \rho_2(h, \theta_1, \theta_2|u_2) \, dh \, d\theta_1} \] (2.26b)

\[\rho_2(h|\neg u_1, \theta_1) = \frac{\rho_2(h, \theta_1|\neg u_1)}{\int_a^b \int_a^b \rho_2(h, \theta_1, \theta_2|\neg u_1) \, dh \, d\theta_2} \] (2.26c)

\[\rho_2(h|\neg u_2, \theta_2) = \frac{\rho_2(h, \theta_2|\neg u_2)}{\int_a^b \int_a^b \rho_2(h, \theta_1, \theta_2|\neg u_2) \, dh \, d\theta_1} \] (2.26d)

Then \(\sigma_3\) can proceed just as \(\sigma_1\) did, once we relativize informativity to a given receiver level.

In fact this allows a general definition for \(\sigma_n^*\):
\[ \sigma_n^*(h_{u_0}) = e^{\lambda \times (\ln(f_0^h f_0^h \rho_{n-1}(h_{u_0}, \theta_1, \theta_2)dh_1dh_2) - C(u_0))} \] (2.27a)

\[ \sigma_n^*(u_1|h, \theta_1) = e^{\lambda \times (\ln(f_0^h f_0^h \rho_{n-1}(h_{u_1}, \theta_1, \theta_2)dh_1dh_2) - C(u_1))} \] (2.27b)

\[ \sigma_n^*(u_2|h, \theta_2) = e^{\lambda \times (\ln(f_0^h f_0^h \rho_{n-1}(h_{u_2}, \theta_1, \theta_2)dh_1dh_2) - C(u_2))} \] (2.27c)

\[ \sigma_n^*(-u_1|h, \theta_1) = e^{\lambda \times (\ln(f_0^h f_0^h \rho_{n-1}(h_{-u_1}, \theta_1, \theta_2)dh_1dh_2) - C(-u_1))} \] (2.27d)

\[ \sigma_n^*(-u_2|h, \theta_2) = e^{\lambda \times (\ln(f_0^h f_0^h \rho_{n-1}(h_{-u_2}, \theta_1, \theta_2)dh_1dh_2) - C(-u_2))} \] (2.27e)

Then, we can define \( \sigma_n \) as before, and \( \rho_n \) for \( n \geq 4 \) becomes:

\[ \rho_n(h, \theta_1, \theta_2|u_0) \propto \sigma_{n-1}(u_0|h, \theta_1, \theta_2) \times \rho_{n-2}(h|u_0) \times \rho_{n-2}(\theta_1, \theta_2|u_0) \] (2.28a)

\[ \rho_n(h, \theta_1, \theta_2|u_1) \propto \sigma_{n-1}(u_1|h, \theta_1, \theta_2) \times \rho_{n-2}(h|u_1) \times \rho_{n-2}(\theta_1, \theta_2|u_1) \] (2.28b)

\[ \rho_n(h, \theta_1, \theta_2|u_2) \propto \sigma_{n-1}(u_2|h, \theta_1, \theta_2) \times \rho_{n-2}(h|u_2) \times \rho_{n-2}(\theta_1, \theta_2|u_2) \] (2.28c)

\[ \rho_n(h, \theta_1, \theta_2|-u_1) \propto \sigma_{n-1}(-u_1|h, \theta_1, \theta_2) \times \rho_{n-2}(h|-u_1) \times \rho_{n-2}(\theta_1, \theta_2|-u_1) \] (2.28d)

\[ \rho_n(h, \theta_1, \theta_2|-u_2) \propto \sigma_{n-1}(-u_2|h, \theta_1, \theta_2) \times \rho_{n-2}(h|-u_2) \times \rho_{n-2}(\theta_1, \theta_2|-u_2) \] (2.28e)

Now, for \( n = 2 \) we can define
\[ \rho_2(h, \theta_1, \theta_2|u_0) \propto \sigma_1(u_0|h, \theta_1, \theta_2) \times \rho_0(h|u_0) \times \rho_0(\theta_1, \theta_2|u_0) \]  
(2.29a)

\[ \rho_2(h, \theta_1, \theta_2|u_1) \propto \sigma_1(u_1|h, \theta_1, \theta_2) \times \rho_0(h|u_0) \times \rho_0(\theta_1, \theta_2|u_1) \]  
(2.29b)

\[ \rho_2(h, \theta_1, \theta_2|u_2) \propto \sigma_1(u_2|h, \theta_1, \theta_2) \times \rho_0(h|u_0) \times \rho_0(\theta_1, \theta_2|u_2) \]  
(2.29c)

\[ \rho_2(h, \theta_1, \theta_2|\neg u_1) \propto \sigma_1(\neg u_1|h, \theta_1, \theta_2) \times \rho_0(h|u_0) \times \rho_0(\theta_1, \theta_2|\neg u_1) \]  
(2.29d)

\[ \rho_2(h, \theta_1, \theta_2|\neg u_2) \propto \sigma_1(\neg u_2|h, \theta_1, \theta_2) \times \rho_0(h|u_0) \times \rho_0(\theta_1, \theta_2|\neg u_2) \]  
(2.29e)

since \( \rho_0(h|u_0) = \phi(h) \) and for \( n = 0 \) we can have

\[ \rho_0(h, \theta_1, \theta_2|u_0) = \rho_0(h|u_0) \times p(\theta_1, \theta_2) \]  
(2.30a)

\[ \rho_0(h, \theta_1, \theta_2|u_1) = \rho_0(h|u_1, \theta_1) \times p(\theta_1, \theta_2) \]  
(2.30b)

\[ \rho_0(h, \theta_1, \theta_2|u_2) = \rho_0(h|u_2, \theta_2) \times p(\theta_1, \theta_2) \]  
(2.30c)

\[ \rho_0(h, \theta_1, \theta_2|\neg u_1) = \rho_0(h|\neg u_1, \theta_1) \times p(\theta_1, \theta_2) \]  
(2.30d)

\[ \rho_0(h, \theta_1, \theta_2|\neg u_2) = \rho_0(h|\neg u_2, \theta_2) \times p(\theta_1, \theta_2) \]  
(2.30e)

where \( \rho_0(h|u_0), \rho_0(h|u_1, \theta_1), \rho_0(h|u_2, \theta_2), \rho_0(h|\neg u_1, \theta_1), \) and \( \rho_0(h|\neg u_2, \theta_2) \) are defined as before. Note that

58
\[
\rho_0(\theta_1, \theta_2 | u_1) = \frac{\int_a^b \rho_0(h, \theta_1, \theta_2 | u_1) dh}{\int_a^b \int_a^b \rho_0(h, \theta_1, \theta_2 | u_1) dh d\theta_1 d\theta_2} \\
= \frac{\int_a^b \rho_0(h|u_1, \theta_1) \times p(\theta_1, \theta_2) dh}{\int_a^b \int_a^b \rho_0(h|u_1, \theta_1) \times p(\theta_1, \theta_2) dh d\theta_1 d\theta_2} \\
= \frac{\int_a^b \int_a^b p(\theta_1, \theta_2) \times \rho_0(h|u_1, \theta_1) dh}{\int_a^b \int_a^b p(\theta_1, \theta_2) \times \rho_0(h|u_1, \theta_1) dh d\theta_1 d\theta_2} \\
= p(\theta_1, \theta_2)
\]

since \(\int_a^b \rho_0(h|u_1, \theta_1) dh = 1\) and \(\int_a^b \rho_0(h|u_1, \theta_1) \times p(\theta_1, \theta_2) d\theta_1 d\theta_2 = 1\); \textit{mutatis mutandis} for \(\rho_0(\theta_1, \theta_2 | u_2)\), \(\rho_0(\theta_1, \theta_2 | -u_1)\), \(\rho_0(\theta_1, \theta_2 | -u_2)\), and \(\rho_0(h|u_0)\). Thus, (2.29) and (2.30) are consistent with the original definitions of \(\rho_0\) and \(\rho_2\). Note also that the process of marginalization and normalization defined in (2.25) and (2.26) applied to the base case distributions defined in (2.30) are consistent with the original model given a uniform prior distribution for \(\theta_1\) and \(\theta_2\) from \(a\) to \(b\):

\[
\int_a^b \int_a^b \rho_0(h, \theta_1, \theta_2 | u_0) d\theta_1 d\theta_2 = \int_a^b \int_a^b \rho_0(h|u_0) \times p(\theta_1, \theta_2) d\theta_1 d\theta_2 \\
= \int_a^b \int_a^b \rho_0(h|u_0) \times \frac{1}{(b-a)^2} d\theta_1 d\theta_2 \\
= \rho_0(h|u_0)
\]

(2.32a)
\[
\frac{\int_a^b \rho_0(h, \theta_1, \theta_2 | u_1) d\theta_2}{\int_a^b \int_a^b \rho_0(h, \theta_1, \theta_2 | u_1) dh d\theta_2} = \frac{\int_a^b \rho_0(h | u_1, \theta_1) \times p(\theta_1, \theta_2) d\theta_2}{\int_a^b \int_a^b \rho_0(h | u_1, \theta_1) \times p(\theta_1, \theta_2) dh d\theta_2} = \frac{\rho_0(h | u_1, \theta_1) \times \int_a^b \frac{1}{(b-a)^2} d\theta_2}{\int_a^b \rho_0(h | u_1, \theta_1) dh d\theta_2} = \rho_0(h | u_1, \theta_1)
\]  

(2.32b)

(Here we omit the corresponding proofs for \(u_2\), \(\neg u_1\), and \(\neg u_2\).) Hence, as defined in (2.27), \(\sigma_n^*\) for \(n = 1\) is the same as \(\sigma_1^*\) in the original model. Hence our entire extended receiver and speaker model is a consistent extension of Lassiter and Goodman’s original model.

Note also that we might replace (2.28) and (2.29) with a single slightly different definition of \(\rho_n\) for \(n > 0\), which uses the null utterance’s \(h\) posterior \(\rho_{n-2}(h | u_0)\) as the prior for calculating the \(h, \theta_1, \theta_2\) posterior at level \(n\), for all of the signals:

\[
\rho_n(h, \theta_1, \theta_2 | u_0) \propto \sigma_{n-1}(u_0 | h, \theta_1, \theta_2) \times \rho_{n-2}(h | u_0) \times \rho_{n-2}(\theta_1, \theta_2 | u_0) \\
\rho_n(h, \theta_1, \theta_2 | u_1) \propto \sigma_{n-1}(u_1 | h, \theta_1, \theta_2) \times \rho_{n-2}(h | u_0) \times \rho_{n-2}(\theta_1, \theta_2 | u_1) \\
\rho_n(h, \theta_1, \theta_2 | u_2) \propto \sigma_{n-1}(u_2 | h, \theta_1, \theta_2) \times \rho_{n-2}(h | u_0) \times \rho_{n-2}(\theta_1, \theta_2 | u_2) \\
\rho_n(h, \theta_1, \theta_2 | \neg u_1) \propto \sigma_{n-1}(\neg u_1 | h, \theta_1, \theta_2) \times \rho_{n-2}(h | u_0) \times \rho_{n-2}(\theta_1, \theta_2 | \neg u_1) \\
\rho_n(h, \theta_1, \theta_2 | \neg u_2) \propto \sigma_{n-1}(\neg u_2 | h, \theta_1, \theta_2) \times \rho_{n-2}(h | u_0) \times \rho_{n-2}(\theta_1, \theta_2 | \neg u_2)
\]  

(2.33a) (2.33b) (2.33c) (2.33d) (2.33e)

This eliminates the need for the separate \(\rho_2\) definitions in (2.29), and remains consistent with the original model. However, it does not help the strategy of ascending the hierarchy.
of senders and receivers to solve our first problem.$^{10}$

The difficulty is that on neither version of the hierarchy of senders and receivers have we introduced the required asymmetry, since at every level the posterior \((\theta_1, \theta_2)\) distributions for \(u_1\) and \(u_2\) will be approximately mirrored by those of \(\neg u_1\) and \(\neg u_2\), respectively. Thus, with regards to when \(\ALT = \{u_0, u_1, u_2, \neg u_1, \neg u_2\}\), as we see in Figure(2.8), at any level of the hierarchy of receivers above level 0, \(\neg u_1\) and \(\neg u_2\) will still be interpreted to mean something stronger than \(u_2\) and \(u_1\), respectively. Even shifting from (2.28) to (2.33) will not help, since the prior \(h\) distribution at any receiver level will be the posterior \(h\) distribution at the previous receiver level, which is symmetric with respect to the mean.

What about when \(\ALT = \{u_0, u_1, \neg u_1\}\)? Again, we can define an extended hierarchy of senders and receivers that is consistent with Lassiter and Goodman’s original model. For

$^{10}$Note that we might define a variation of (2.33) by using, at every level \(n > 0\), \(\rho_0(h|u_0)\) instead of \(\rho_{n-2}(h|u_0)\). Note further that we might define another option by using, at every level \(n > 0\), \(\rho_0(\theta_1, \theta_2|u_n)\) instead of \(\rho_0(\theta_1, \theta_2|u_n)\). These remain consistent extensions of the model.
the senders we marginalize out \( \theta_1 \) for \( u_0 \) and normalize by \( \theta_1 \) for \( u_1 \) and \( \neg u_1 \):

\[ \sigma_n^*(u_0|h) = e^{\lambda (\ln(\int h^0 \rho_{n-1}(h,\theta_1|u_0)d\theta_1) - C(u_0))} \]  
\[ \sigma_n^*(u_1|h,\theta_1) = e^{\lambda (\ln(\int h^1 \rho_{n-1}(h,\theta_1|u_1)d\theta_1) - C(u_1))} \]  
\[ \sigma_n^*(\neg u_1|h,\theta_1) = e^{\lambda (\ln(\int h^1 \rho_{n-1}(h,\theta_1|\neg u_1)d\theta_1) - C(\neg u_1))} \]

(2.34a)  
(2.34b)  
(2.34c)

Again, we can define \( \sigma_n \) as before, and \( \rho_n \) for \( n \geq 4 \) becomes:

\[ \rho_n(h,\theta_1|u_0) \propto \sigma_{n-1}(u_0|h,\theta_1) \times \rho_{n-2}(h|u_0) \times \rho_{n-2}(\theta_1|u_0) \]  
\[ \rho_n(h,\theta_1|u_1) \propto \sigma_{n-1}(u_1|h,\theta_1) \times \rho_{n-2}(h|u_1) \times \rho_{n-2}(\theta_1|u_1) \]  
\[ \rho_n(h,\theta_1|\neg u_1) \propto \sigma_{n-1}(\neg u_1|h,\theta_1) \times \rho_{n-2}(h|\neg u_1) \times \rho_{n-2}(\theta_1|\neg u_1) \]

(2.35a)  
(2.35b)  
(2.35c)

For \( n = 2 \) we can define

\[ \rho_2(h,\theta_1|u_0) \propto \sigma_1(u_0|h,\theta_1) \times \rho_0(h|u_0) \times \rho_0(\theta_1|u_0) \]  
\[ \rho_2(h,\theta_1|u_1) \propto \sigma_1(u_1|h,\theta_1) \times \rho_0(h|u_0) \times \rho_0(\theta_1|u_1) \]  
\[ \rho_2(h,\theta_1|\neg u_1) \propto \sigma_1(\neg u_1|h,\theta_1) \times \rho_0(h|u_0) \times \rho_0(\theta_1|\neg u_1) \]

(2.36a)  
(2.36b)  
(2.36c)
since $\rho_0(h|u_0) = \phi(h)$ and for $n = 0$ we can have

\[
\rho_0(h, \theta_1|u_0) = \rho_0(h|u_0) \times p(\theta_1) \quad (2.37a)
\]
\[
\rho_0(h, \theta_1|u_1) = \rho_0(h|u_1, \theta_1) \times p(\theta_1) \quad (2.37b)
\]
\[
\rho_0(h, \theta_1|\neg u_1) = \rho_0(h|\neg u_1, \theta_1) \times p(\theta_1) \quad (2.37c)
\]

where $\rho_0(h|u_0)$, $\rho_0(h|u_1, \theta_1)$, and $\rho_0(h|\neg u_1, \theta_1)$ are defined as before.\footnote{The proofs of consistency with the original model, assuming a uniform distribution $p(\theta_1)$, are similar to the ones already given, except we do not need to integrate over $\theta_2$. We could also replace (2.35) and (2.36) with a single definition of $\rho_n$ for $n > 0$, just as we proposed replacing (2.28) and (2.29) with (2.33). Doing so does not introduce the required asymmetry, and so does not solve our second problem. Note also that we could use $\rho_0(h|u_0)$ instead of $\rho_{n-2}(h|u_0)$ at every level $n > 0$, just as we proposed in Footnote (10) for when $\text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\}$. Note again that we might define a further option if we use $\rho_0(\theta_1|u_n)$ instead of $\rho_{n-2}(\theta_1|u_n)$ at every level $n > 0$. Again, these remain consistent extensions of the original model.} Again, as we see in Figure (2.9), where we define the sender and receiver hierarchy as in (2.27)-(2.30), when $\text{ALT} = \{u_0, u_1, \neg u_1\}$, at any level of the hierarchy of receivers above level 0, $\neg u_1$ will be interpreted to mean something stronger than, but in the opposite direction from, $u_1$, since at any level the posterior $\theta_1$ distribution given $u_1$ will be approximately mirrored by the posterior $\theta_1$ distribution given $\neg u_1$.

Now, we could require that when $\text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\}$, the negation of the antonym of a relative gradable adjective is always interpreted at a lower level of the sender-receiver hierarchy than the adjective itself. Since interpretation at higher levels are stronger, that...
would allow $u_1$ and $u_2$ to mean something stronger than $\neg u_2$ and $\neg u_1$, respectively, thus solving our first problem. A similar strategy might be applied when $\text{ALT} = \{u_0, u_1, \neg u_1\}$, but instead requiring that the negation of a relative gradable adjective be interpreted at a lower level of the sender-receiver hierarchy than the adjective itself, thus solving our second problem. In fact we could impose a general requirement that the negations of relative gradable adjectives be interpreted at lower levels than the adjectives themselves, for any given $\text{ALT}$. But, absent any independent motivation why the negations of relative gradable adjectives are always interpreted at a lower level of the sender-receiver hierarchy than the adjective itself, this seems *ad hoc*. And in fact, there seems to be some evidence for metalinguistic negation, and thus we say ‘Shaq isn’t big, he’s *HUGE*!’, and in the denial that Shaq is big one does not seem to be saying the same thing as what one is saying in the denial that *I* am big; instead, we are perhaps saying that the probability of a speaker calling Shaq ‘big’ is low, compared the probability of a speaker calling Shaq ‘huge’, since a receiver presuming the speaker to be choosing between the former and the latter would have a less accurate estimate.
of Shaq’s size upon hearing the former compared to the latter. Or, perhaps we are saying that a receiver would get a less accurate impression of Shaq’s size from the utterance that he is ‘big’ than from the utterance that he is ‘huge’. Importantly, however, in the former case the speaker seems to be talking about the encoding probabilities of a lower level sender, and in the latter case the speaker seems to be talking about the decoding probabilities of a lower level receiver. Thus in either case, appeal to metalinguistic negation should lead us to expect the negations of relative gradable adjectives to be interpreted at a higher level, not a lower level, than the corresponding un-negated forms.

Nor will ascending the hierarchy in combination with the earlier strategy of varying the set of alternative utterances solve our first problem, given that we are comparing \( \rho_n(h|u_2) \) when \( ALT = \{u_0, u_1, u_2\} \) with \( \rho_n(h|\neg u_1) \) when \( ALT = \{u_0, u_1, \neg u_1\} \), since the lower cost of \( u_2 \) will ensure that it receives a weaker interpretation than \( \neg u_1 \) given the same parameter settings. (When \( ALT = \{u_0, u_1, u_2\} \) the hierarchy is defined as in (2.27)-(2.30), or (2.27), (2.33), and (2.30), but with the sender and receiver definitions for \( \neg u_1 \) and \( \neg u_2 \) deleted.) Again, at any given level \( n \), making \( \neg u_1 \) have a weaker reading than \( u_2 \) requires that \( \neg \) have a negative cost. And, requiring that interpretation of \( \rho_n(h|u_2) \) when \( ALT = \{u_0, u_1, u_2\} \) happens at a level higher than interpretation of \( \rho_{n-k}(h|\neg u_1) \) when \( ALT = \{u_0, u_1, \neg u_1\} \) seems ad hoc.

### 2.5.2 Privileging Non-negated Posteriors

In order for ascension of the sender-receiver hierarchy to solve either of our problems, it might seem helpful to privilege the posterior \((\theta_1, \theta_2)\) distributions of \( u_1 \) and \( u_2 \) when \( ALT = \{u_0, u_1, u_2, \neg u_1, \neg u_2\} \), and the posterior \( \theta_1 \) distributions of \( u_1 \) when \( ALT = \{u_0, u_1, \neg u_1\} \). Af-
ter all, $\rho_2(\theta_1|\neg u_1)$ is the mirror image of $\rho_2(\theta_1|u_1)$ about the median of $\phi(h)$ when $C(\neg u_1) = C(u_1)$, and even further to the left when $C(\neg u_1) > C(u_1)$. If $\text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\}$, then, using the posterior $\rho_{n-2}(\theta_1, \theta_2|u_1)$ instead of $\rho_{n-2}(\theta_1, \theta_2|\neg u_1)$ as the prior for calculating $\rho_n(h, \theta_1, \theta_2|\neg u_1)$ might seem to pull $\rho_n(h, \theta_1, \theta_2|\neg u_1)$ to the right; in the same way, using the posterior $\rho_{n-2}(\theta_1, \theta_2|u_2)$ instead of $\rho_{n-2}(\theta_1, \theta_2|\neg u_2)$ as the prior for calculating $\rho_n(h, \theta_1, \theta_2|\neg u_2)$ might seem to pull $\rho_n(h, \theta_1, \theta_2|\neg u_2)$ to the left.

Thus we might propose that in the definition of $\rho_n$ for $n \geq 4$, (2.28d) and (2.28e) should be replaced by:

\[
\rho_n(h, \theta_1, \theta_2|\neg u_1) \propto \sigma_{n-1}(\neg u_1|h, \theta_1, \theta_2) \times \rho_{n-2}(h|\neg u_1) \times \rho_{n-2}(\theta_1, \theta_2|u_1) \tag{2.38a}
\]
\[
\rho_n(h, \theta_1, \theta_2|\neg u_2) \propto \sigma_{n-1}(\neg u_2|h, \theta_1, \theta_2) \times \rho_{n-2}(h|\neg u_2) \times \rho_{n-2}(\theta_1, \theta_2|u_2) \tag{2.38b}
\]

and $\rho_2$ and $\rho_0$ should be as in (2.29) and (2.30), respectively.

A second option when $\text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\}$ is to think of $\rho_n$ for $n \geq 4$ as gaining from $u_1$ information about $\theta_1$ but not $\theta_2$, and only gaining information about $\theta_1$ from $u_1$, and likewise for $u_2$ and $\theta_2$, and then assuming the independence of $\theta_1$ and $\theta_2$. Thus we might propose that in (2.28) we replace $\rho_{n-2}(\theta_1, \theta_2|u_0)$, $\rho_{n-2}(\theta_1, \theta_2|u_1)$, $\rho_{n-2}(\theta_1, \theta_2|u_2)$, $\rho_{n-2}(\theta_1, \theta_2|\neg u_1)$, and $\rho_{n-2}(\theta_1, \theta_2|\neg u_2)$ with

\[
\rho_{n-2}(\theta_1|u_1) \times \rho_{n-2}(\theta_2|u_2) \tag{2.39}
\]
We see the results from (2.38) and (2.39) in (2.10) and (2.11) and compare them to the results from the original model shown in (2.12). On either option, the negation of the antonym of a relative gradable adjective is still interpreted more strongly than the adjective itself; we only discuss (2.38) but similar comments apply to (2.39). Counterintuitively, for \( n > 2 \) using the posterior from \( \rho_{n-2}(\theta_1, \theta_2|u_1) \) to calculate \( \rho_n(h, \theta_1, \theta_2|\neg u_1) \) actually results in \( \rho_n(h|\neg u_1) \) being even more extreme than in the default hierarchy; similarly for using the posterior from \( \rho_{n-2}(\theta_1, \theta_2|u_2) \) to calculate \( \rho_n(h, \theta_1, \theta_2|\neg u_2) \) for \( n > 2 \). A further problem is that both of these strategies seem poorly pragmatically motivated: it seems, intuitively, that from hearing something called ‘not tall’ one can in fact gain information about how tall something has to be to count as tall: something has to be at least taller than the thing that was called ‘not tall’. Thus, there seems to be no independent reason to use \( \rho_{n-2}(\theta_1, \theta_2|u_1) \) instead of \( \rho_{n-2}(\theta_1, \theta_2|\neg u_1) \) when calculating \( \rho_n(h, \theta_1, \theta_2|\neg u_1) \).\(^{12}\)

\(^{12}\)Note also that we might apply either (2.38) or (2.39) to the definition of \( \rho_n \) for \( n > 0 \) in (2.33). If we choose the former, (2.33d) and (2.33e) should be replaced by:

\[
\begin{align*}
\rho_n(h, \theta_1, \theta_2|\neg u_1) & \propto \sigma_{n-2}(\neg u_1|h, \theta_1, \theta_2) \times \rho_{n-2}(h|u_0) \times \rho_{n-2}(\theta_1, \theta_2|u_1) \\
\rho_n(h, \theta_1, \theta_2|\neg u_2) & \propto \sigma_{n-1}(-u_2|h, \theta_1, \theta_2) \times \rho_{n-2}(h|u_0) \times \rho_{n-2}(\theta_1, \theta_2|u_2)
\end{align*}
\]

(2.40a, b)

and \( \rho_0 \) should be as in (2.30). Since \( \rho_2(\theta_1, \theta_2|u_1) = p(\theta_1, \theta_2) = \rho_2(\theta_1, \theta_2|\neg u_1) \), and similarly for \( u_2 \) and \( \neg u_2 \), this will still be consistent with the original model. If we choose the latter, then we replace \( \rho_{n-2}(\theta_1, \theta_2|u_0), \rho_{n-2}(\theta_1, \theta_2|u_1), \rho_{n-2}(\theta_1, \theta_2|u_2), \rho_{n-2}(\theta_1, \theta_2|\neg u_1), \) and \( \rho_{n-2}(\theta_1, \theta_2|\neg u_2) \) in (2.33) with (2.39) and again \( \rho_0 \) should be as in (2.30). Since \( \rho_2(\theta_1|u_1) = \rho_2(\theta_2|u_1) = \frac{1}{\alpha^b} \), this is also consistent with the original model.

The results for (2.40) do not significantly differ from those for (2.38). The results for the latter option for the same parameter settings as in Figure (2.39) solve our first problem only for \( n = 6 \); at levels higher than that the interpretations for ‘not tall’ become too close to ‘tall’, and at levels below that the interpretations
Figure 2.10: Lassiter and Goodman Model Three Signals with Not Normal Distribution
Lambda 4 Low Cost Theta Posterior Source = Signal Specific Signal 0, Signal 1, Signal 2, Signal 1, Signal 2 Receiver Hierarchy

Figure 2.11: Lassiter and Goodman Model Three Signals with Not Normal Distribution
Lambda 4 Low Cost Theta Posterior Source = Signal 1, Signal 2 Receiver Hierarchy
Will using $\rho_{n-2}(\theta_1|u_1)$ instead of $\rho_{n-2}(\theta_1|\neg u_1)$ to calculate $\rho_n(h, \theta_1|\neg u_1)$ when $\text{ALT} = \{u_0, u_1, \neg u_1\}$ solve our second problem? Here we would be proposing to replace (2.35c) with:

$$
\rho_n(h, \theta_1|\neg u_1) \propto \sigma_{n-1}(-u_1|h, \theta_1) \times \rho_{n-2}(h|\neg u_1) \times \rho_{n-2}(\theta_1|u_1)
$$

(2.41)

Another option when $\text{ALT} = \{u_0, u_1, u_2\}$, analogous to the approach in (2.39) when $\text{ALT} = \{u_0, u_1, \neg u_1, \neg u_2\}$, is to think of $\rho_n$ as only gaining information about $\theta_1$ from $u_1$; thus we would replace $\rho_{n-2}(\theta_1|u_0)$, $\rho_{n-2}(\theta_1|u_1)$, and $\rho_{n-2}(\theta_1|\neg u_1)$ in (2.35) with

$$
\rho_{n-2}(\theta_1|u_1)
$$

(2.42)

for ‘not tall’ are too close to ‘short’. Again, this solution seems $ad$ $hoc$. Both of these options also share the problem of privileging $u_1$ and $u_2$ over $\neg u_1$ and $\neg u_2$. 
The results for (2.41) and (2.42) are shown in Figures (2.13) and (2.14), respectively; as we see, neither of these approaches solves our second problem. As we see in Figure (2.13), even if we use $\rho_{n-2}(\theta_1|u_1)$ instead of $\rho_{n-2}(\theta_1|\neg u_1)$ to calculate $\rho_n(h,\theta_1|\neg u_1)$, $\rho_n(h|\neg u_1)$ is still stronger but in the opposite direction from $\rho_n(h|u_1)$. And again, it seems that we can gain information about just how tall someone has to be to count as tall, from an utterance of ‘not tall’; so there seems to be independent reason to treat $\rho_{n-2}(\theta_1|\neg u_1)$ on a par with $\rho_{n-2}(\theta_1|u_1)$.

Now, one way to treat $\rho_{n-2}(\theta_1|\neg u_1)$ on a par with $\rho_{n-2}(\theta_1|u_1)$ when $\text{ALT} = \{u_0, u_1, \neg u_1\}$ is to use the average of the two as the prior when calculating the posterior probabilities for $h$ and $\theta_1$. Thus, in (2.35) we might replace $\rho_{n-2}(\theta_1|u_0)$, $\rho_{n-2}(\theta_1|u_1)$, and $\rho_{n-2}(\theta_1|\neg u_1)$ with

\[\rho_{n-2}(\theta_1|\neg u_1) + \rho_{n-2}(\theta_1|u_1)\]
\[
\frac{\rho_{n-2}(\theta_1|u_1) + \rho_{n-2}(\theta_1|\neg u_1)}{2}
\] (2.43)

A similar strategy when \( \text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\} \) would be to claim that we gain information about \( \theta_1 \) but not \( \theta_2 \) from \( u_1 \) and \( \neg u_1 \) and information about \( \theta_2 \) but not \( \theta_1 \) from \( u_2 \) and \( \neg u_2 \). We then assume that \( \theta_1 \) and \( \theta_2 \) are independently distributed. Thus in (2.28) we might replace \( \rho_{n-2}(\theta_1, \theta_2|u_0), \rho_{n-2}(\theta_1, \theta_2|u_1), \rho_{n-2}(\theta_1, \theta_2|u_2), \rho_{n-2}(\theta_1, \theta_2|\neg u_1) \), and \( \rho_{n-2}(\theta_1, \theta_2|\neg u_2) \) with

\[
\frac{\rho_{n-2}(\theta_1|u_1) + \rho_{n-2}(\theta_1|\neg u_1)}{2} \times \frac{\rho_{n-2}(\theta_2|u_2) + \rho_{n-2}(\theta_2|\neg u_2)}{2}
\] (2.44)

or, if we find it acceptable to gain information about \( \theta_1 \) from \( u_2 \) and \( \neg u_2 \) and information about \( \theta_2 \) from \( u_1 \) and \( \neg u_1 \):
\[
\frac{\rho_{n-2}(\theta_1|u_1) + \rho_{n-2}(\theta_1|\neg u_1) + \rho_{n-2}(\theta_1|u_2) + \rho_{n-2}(\theta_1|\neg u_2)}{4} \times \\
\frac{\rho_{n-2}(\theta_1|u_1) + \rho_{n-2}(\theta_1|\neg u_1) + \rho_{n-2}(\theta_2|u_2) + \rho_{n-2}(\theta_2|\neg u_2)}{4}
\]  
(2.45)

Note that we could also define versions of (2.43), (2.44), and (2.45) that incorporate information about \(\theta_1\) and \(\theta_2\) from \(u_0\).\(^{14}\) We see the results for (2.43), (2.44) and (2.45) in Figures (2.15), (2.16), and (2.17); as we see, the negation of a relative gradable adjective is still interpreted more strongly than the relative gradable adjective itself, and the negation of a relative gradable adjective is still interpreted more strongly than the antonym of the relative gradable adjective. Thus, neither of our two problems are solved.

What about appealing to alternative utterances while privileging the \(\theta_1\) posterior distributions from \(u_1\) when \(\text{ALT} = \{u_0, u_1, \neg u_1\}\), perhaps in combination with applying (2.39) to (2.28a), (2.28b), and (2.28c) when \(\text{ALT} = \{u_0, u_1, u_2\}\)? As noted in Footnote (13), this solves the second problem for certain parameter settings at certain levels if we use (2.42) applied to the hierarchy defined in Footnote (11), and in combination with the appeal to alternative utterances, will solve our first problem, since the interpretation of \(\neg u_1\) will be weaker than the interpretation of \(u_2\). Again, however, this is too \textit{ad hoc}, the interpretation of \(u_0\) has an implausible asymmetry that seems to be an artifact of the model, and the entire

\(^{14}\)And again, when \(\text{ALT} = \{u_0, u_1, \neg u_1\}\) we could apply (2.43) to the modified hierarchy outlined in Footnote (11), and when \(\text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\}\) we could apply (2.44) or (2.45) to the modified hierarchy defined in (2.33). Neither of the latter two models solves our first problem, and the former model does not solve our second problem.
Figure 2.15: Lassiter and Goodman Model Two Signals with Not Normal Distribution Lambda 4 Low Cost Theta Posterior Source = Signal 1, Not 1 Receiver Hierarchy

prospect of privileging the $\theta_1$ posterior distribution from $u_1$ and ignoring that of $\neg u_1$ seems poorly pragmatically motivated.

\section{2.6 Modifying the Model: Assuming $\theta_1 > \theta_2$}

What might we gather from the failure of ascending the sender-receiver hierarchy to solve either of our problems? We have been trying to induce a desired asymmetry between the interpretation of relative gradable adjectives and their negations; that is, we have been trying to make $\rho_n(h|\neg u_1)$ sufficiently weaker than $\rho_n(h|u_2)$ in the same direction, and sufficiently weaker than $\rho_n(h|u_2)$ in the opposite direction, in keeping with the intuition that upon hearing something called ‘not tall’ we would shift our probability distribution for their height downward, but not as far downward as we would shift it downward if we heard that thing called ‘short’, and not as far downward as we would shift it \textit{upward} if we heard that thing called ‘tall’. \textit{Mutatis mutandis} for $\rho_n(h|\neg u_2)$ and ‘not short’. We have been trying to create this asymmetry by using non-uniform, asymmetric distributions for $\theta_1$ and...
Figure 2.16: Lassiter and Goodman Model Three Signals with Not Normal Distribution
Lambda 4 Low Cost Theta Posterior Source = Signal 1, Not 1, Signal 2, Not 2 Receiver Hierarchy

Figure 2.17: Lassiter and Goodman Model Three Signals with Not Normal Distribution
Lambda 4 Low Cost Theta Posterior Source = Signal 1, 2, Not 1, Not 2, Signal 1, 2, Not 1, Not 2 Receiver Hierarchy
\(\theta_2\) derived at higher levels of the sender-receiver hierarchy. However, even if we ignored 
\(\rho_{n-2}(\theta_1, \theta_2|\neg u_1)\) and \(\rho_{n-2}(\theta_1, \theta_2|\neg u_2)\) and instead used \(\rho_{n-2}(\theta_1, \theta_2|u_1)\) and \(\rho_{n-2}(\theta_1, \theta_2|u_2)\) to calculate \(\rho_n(h|\neg u_1)\) and \(\rho_n(h|\neg u_2)\), respectively, for \(n \geq 4\) in order to introduce the desired asymmetry between relative gradable adjectives and their negations, \(\rho_n(h|\neg u_1)\) and \(\rho_n(h|\neg u_2)\) were still not sufficiently weaker than \(\rho_n(h|u_1)\) and \(\rho_n(h|u_2)\) — in fact, they were stronger. Thus we might infer that we need to introduce asymmetric, non-uniform distributions for \(\theta_1\) and \(\theta_1\) at even lower levels: that is, we should depart from Lassiter and Goodman’s original model at the level of \(\rho_2\) and let \(\rho_2\) draw on non-uniform prior \(\theta_1\) and \(\theta_2\) distributions. That, of course, would require that \(\rho_0\) have a non-uniform distributions for \(\theta_1\) and \(\theta_2\).

How might we think of \(\rho_0\) as having pragmatically driven non-uniform distributions for \(\theta_1\) and \(\theta_2\)? We might think of the listener as assuming that \(\theta_1 > \theta_2\) — lest he think that anything can be both short and tall. This cannot be solely a matter of the meaning of ‘short’ and ‘tall’ on the Kennedy semantics, since nothing in that semantics requires that the threshold for ‘tall’ be greater than the threshold for ‘short’. However, if we grant that the receiver knows that each threshold is chosen so as to ensure that the object ‘stands out’ in the context of utterance, perhaps then the receiver would infer that relative to the statistical distribution of heights of the comparison class, the height required to stand out as having greater height than some preponderance of the comparison class must be greater than the height required to stand out as having lesser height than some preponderance of the comparison class, so the threshold for ‘tall’ must be greater than the threshold for ‘short’. More explicitly, if the threshold for ‘tall’ were chosen so as to ensure the object stands out relative to the
comparison class, that threshold would have to be significantly above most of the heights
the comparison class; then if the threshold for ‘short’ were at or above the threshold for ‘tall’,
then most of the heights of the comparison class would fall below the threshold for ‘short’,
and there would be no guarantee that anything called ‘short’ would stand out relative to the
comparison class: even things of slightly more than average height would be ‘short’.

What is the distribution for $\theta_1$ and $\theta_2$ if we assume $\theta_2 < \theta_1$? We might imagine two urns,
one with balls for each possible $\theta_1$ value, and the other with balls for each possible $\theta_2$ value.
A machine spins the urns around and stops randomly, if it is the $\theta_1$ urn we draw a ball to
get a $\theta_1$ value, and then every ball in the $\theta_2$ urn with a value greater than our $\theta_1$ value drops
out of that urn, and we then pick a $\theta_2$ ball from the $\theta_2$ urn; likewise if it is the $\theta_2$ urn that
appears in front of us, but with the $\theta_1$ balls less than our $\theta_2$ value dropping out of the $\theta_1$
urn. Where $\phi^{\theta_1}(\theta_1)$ and $\phi^{\theta_2}(\theta_2)$ are the distributions of balls, respectively, in the $\theta_1$ and $\theta_2$
urns, we can define the probabilities $Pr^{\theta_1}(\theta_1, \theta_2)$, $Pr^{\theta_2}(\theta_1, \theta_2)$ of $(\theta_1, \theta_2)$ when picking first
from the $\theta_1$, $\theta_2$ urns, respectively, as:

$$Pr^{\theta_1}(\theta_1, \theta_2) = p(\theta_2|\theta_2 < \theta_1) \times \phi^{\theta_1}(\theta_1)$$

$$= \begin{cases} 
\frac{\phi^{\theta_2}(\theta_2)}{\int_0^{\theta_1} \phi^{\theta_2}(\theta_2) d\theta_2} \times \phi^{\theta_1}(\theta_1) & \text{if } \theta_2 < \theta_1 \\
0 & \text{otherwise}
\end{cases} \quad (2.46a)$$

76
\[ P_{\theta_2}(\theta_1, \theta_2) = p(\theta_1 | \theta_2 < \theta_1) \times \phi^{\theta_2}(\theta_2) \]

\[
= \begin{cases} 
\frac{\phi^{\theta_1}(\theta_1)}{\int_{\theta_2} \phi^{\theta_1}(\theta_1)d\theta_1} \times \phi^{\theta_2}(\theta_2) & \text{if } \theta_2 < \theta_1 \\
0 & \text{otherwise}
\end{cases}
\] (2.46b)

Then the overall distribution is

\[
p(\theta_1, \theta_2) = \begin{cases} 
\frac{P_{\theta_2}(\theta_1, \theta_2) + P_{\theta_2}(\theta_1, \theta_2)}{2} & \text{if } \theta_2 < \theta_1 \\
0 & \text{otherwise}
\end{cases}
\] (2.47)

We can then think of \( \rho_0 \) as deriving \( p(\theta_1, \theta_2) \) and then using it in (2.30). Of course, what \( p(\theta_1, \theta_2) \) turns out to be depends on just what \( \phi^{\theta_1} \) and \( \phi^{\theta_2} \) are. We should note that initial normal distributions in the urns allows for using the full normal state distribution for \( \phi(h) \), instead of a truncated normal state distribution, since truncated normal distributions were only required since uniform distributions are defined only over finite range. If we think of \( \phi^{\theta_1} \) and \( \phi^{\theta_2} \) as uniform, however, we seem to hew closer to Lassiter and Goodman’s original model, even if that model’s limitation to uniform \( \theta \) distributions was a semantically unmotivated technical necessity.

Overall, for \( \text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\} \), we find that when assuming that the listener assumes \( \theta_2 < \theta_1 \), we come closer to solving our first problem when we define the senderreceiver hierarchy as in (2.28) and (2.29), or some such previously defined variant thereof, as opposed to (2.33) or some such previously defined variant thereof. More specifically, the most plausible \( h \) posteriors occur not for (2.28) itself but for its variants defined in (2.38),

77
(2.39), (2.44), and (2.45), which are shown in Figures (2.18), (2.19), (2.20), and (2.21), respectively. Note that for (2.28) and (2.38), at each level \( n \) in the receiver hierarchy, for any signal contributing a posterior to be a prior at level \( n + 2 \), we marginalize out \( h \) and \( \theta_1 \) to define \( \phi_{n}^{\theta_2} \) and marginalize out \( h \) and \( \theta_1 \) to to define \( \phi_{n}^{\theta_1} \), and then calculate (2.47) as the prior contributed by that signal to be used at level \( n + 2 \). For (2.39), (2.44) and (2.45), at each level \( n \), we calculate the average for \( \theta_1 \) and \( \theta_2 \), and then instead of multiplying them as in (2.44) and (2.45), we use those as \( \phi_{n}^{\theta_1} \) and \( \phi_{n}^{\theta_2} \) and calculate (2.47) as the prior to be used by all signals at level \( n + 2 \). Furthermore, we find that starting with \( \phi_{0}^{\theta_1} \) and \( \phi_{0}^{\theta_2} \) as normally distributed, as opposed to uniformly distributed, tends to yield \( h \) posteriors that come closer to solving our first problem, and \( \theta_1, \theta_2 \) posterior distributions that are more plausible. Under these conditions, the best results tend to occur at level 4; at higher levels the interpretations are overly narrow.

Although having the listener assume that \( \theta_2 < \theta_1 \) seems to solve our first problem, there are two problems: First, among (2.38), (2.39), (2.44), or (2.45), the most plausible \( \theta_1 \) and
Figure 2.19: Lassiter and Goodman Model Three Signals with Not Normal Distribution
Lambda 4 Low Cost Theta Posterior Source = Signal 1, Signal 2 Theta Distr Rel = True
Theta Distr Normal Receiver Hierarchy

Figure 2.20: Lassiter and Goodman Model Three Signals with Not Normal Distribution
Lambda 4 Low Cost Theta Posterior Source = Signal 1, Not 1, Signal 2, Not 2 Theta Distr Rel = True
Theta Distr Normal Receiver Hierarchy
\( \lambda = 4.0, C(u_0) \approx 0.0, C(u_n) \approx 1.0, C(\neg u_n) \approx \ldots \), \( \mu = 0.0, \sigma = 1.0 \), \( \text{num states} = 160 \), \( \theta \text{ distribution type} = \text{normal} \), \( \theta \text{ distribution rel} = \text{True} \), \( \theta \text{ posterior source} = \text{signal1, 2, not1, not2, signal1, 2, not1, not2} \), \( \theta \text{ pragmatic sender type} = \text{hsensitive} \).

\[ \theta_2 \] priors (used to calculate \((h, \theta_1, \theta_2)\) posteriors) occur with (2.38) or (2.39); see Figures (2.22)-(2.24). In Figure (2.22), we do see plausible priors resulting from Equation (2.47) at level 2 and 4, from the model defined in Equation (2.39). (We omit the figures for the model defined in Equation (2.38) for brevity, since each signal has a unique \((\theta_1, \theta_2)\) distribution; however, the distributions are similar to those for the model defined in Equation (2.39).) In Figure 2.22, the green and blue lines in Sub-figure (2.22a) result from projecting the surface in Sub-figure (2.22b) onto the right and left walls, respectively, thus the green line represents the marginalization out of \(\theta_1\) from the result of Equation (2.47) at level 2, and the blue line represents the marginalization out of \(\theta_2\) from the result of Equation (2.47) at level 2. Similar remarks apply to Figures (2.23) and (2.24). In contrast, we see in Figures (2.23) and (2.24) the priors resulting from Equation (2.47) from the models defined in Equations (2.44) and (2.45), respectively. The projections show an implausible ‘two-hump’ shape. We thus seem forced back to model (2.38) or (2.39); however, these face the earlier objection that they privilege the information about the threshold from the un-negated form over the negated

Figure 2.21: Lassiter and Goodman Model Three Signals with Not Normal Distribution Lambda 4 Low Cost Theta Posterior Source = Signal 1, 2, Not 1, Not 2, Signal 1, 2, Not 1, Not 2 Theta Distr Rel = True Theta Distr Normal Receiver Hierarchy
form, *contra* our intuitions that we can learn about the threshold from negated forms just as well as from un-negated forms. Thus, not privileging the un-negated forms forces us to define the model in ways that depend on less plausible $\theta_1$ and $\theta_2$ priors.

Second, assuming the listener assumes that $\theta_2 < \theta_1$ does nothing at all to solve our second problem, when $\text{ALT} = \{u_0, u_1, \neg u_1\}$, since in that case there is no $\theta_2$ to be less than $\theta_1$. Now, we might appeal to alternative utterances and claim that we are comparing $\rho_n(h|\neg u_1)$ given that $\text{ALT} = \{u_0, u_1, \neg u_1\}$ with $\rho_n(h|u_1)$ given that $\text{ALT} = \{u_0, u_1, u_2\}$, and assuming the receiver assumes $\theta_2 < \theta_1$ in the latter case. However, under the same parameter settings, at any level, $\rho_n(h|\neg u_1)$ when $\text{ALT} = \{u_0, u_1, \neg u_1\}$ remains roughly as strong as, but in the opposite direction from, $\rho_n(h|u_1)$ when $\text{ALT} = \{u_0, u_1, u_2\}$ and we assume the receiver is assuming $\theta_2 < \theta_1$, as we can see if we compare Figures (2.9), (2.13), (2.14), or (2.15) to Figures (2.25) or (2.26) or (2.27). Here, Figure (2.25) is the result of (2.28), (2.29), and (2.30), applied to $\text{ALT} = \{u_0, u_1, u_2\}$, Figure (2.26) is the result of (2.39) applied to $\text{ALT} = \{u_0, u_1, u_2\}$, and Figure (2.27) is the result of (2.45), appropriately modified to account for the absence of negated forms in calculating the average $\theta_1$ and $\theta_2$ distributions, applied to $\text{ALT} = \{u_0, u_1, u_2\}$. In all three cases the receiver assumes $\theta_2 < \theta_1$. Note that since the forms are only un-negated when $\text{ALT} = \{u_0, u_1, u_2\}$, (2.38) does not analogize.\textsuperscript{15}

This inability to extend the technique of having the receiver assume that $\theta_2 < \theta_1$ to solve our second problem should be especially worrisome, and make the technique seem especially

\textsuperscript{15}We let $\phi_0^{\theta_1}$ and $\phi_0^{\theta_2}$ be uniform over the ranges indicated in Figures (2.25) or (2.26) or (2.27) in order to allow direct comparison with Figures (2.9), (2.13), (2.14), and (2.15), in which we assumed that $p(\theta_1)$ was uniform over the same indicated ranges. If we assume normal distributions for the comparison, the results are no different.
Figure 2.22: Lassiter and Goodman Model Three Signals with Not Normal Distribution Lambda 4 Low Cost Theta Posterior Source = Signal 1, Signal 2 Theta Distr Rel = True Theta Distr Normal Receiver Level 2, 4 Theta Distribution
Figure 2.23: Lassiter and Goodman Model Three Signals with Not Normal Distribution
Lambda 4 Low Cost Theta Posterior Source = Signal 1, Not 1, Signal 2, Not 2 Theta Distr
Rel = True Theta Distr Normal Receiver Level 2, 4 Theta Distribution
Figure 2.24: Lassiter and Goodman Model Three Signals with Not Normal Distribution
Lambda 4 Low Cost Theta Posterior Source = Signal 1, 2, Not 1, Not 2, Signal 1, 2, Not 1, Not 2 Theta Distr Rel = True Theta Distr Normal Receiver Level 2, 4 Theta Distribution
ad hoc, given that the first and second problems seem fundamentally related: both grow out of the fact that a relative gradable adjective and its negation are such that the central tendency of the distribution resulting from interpretation of the latter is closer than the central tendency of the distribution resulting from interpretation of the former is, to the central tendency of the prior distribution. Since their central tendencies are on opposite sides of the central tendency of the prior distribution, and the interpretation of the antonym of a relative gradable adjective is the mirror-image of the interpretation of the adjective itself, it follows that the antonym of a relative gradable adjective and its negation are such that the central tendency of the distribution resulting from interpretation of the latter is also closer than the central tendency of the distribution resulting from interpretation of the former is, to the central tendency of the prior distribution, but on the same side as, instead of opposite sides of, the central tendency of the prior distribution. Perhaps put more simply, the second problem grows out of an asymmetry between the interpretation of a relative gradable adjective and its negation such that the latter is interpreted more weakly than the former; the first problem then arises simply since the interpretation of the antonym of a relative gradable adjective is the mirror image of the interpretation of the adjective itself, so the interpretation of the negation of a relative gradable adjective is also weaker than that of the adjective’s antonym.
Figure 2.25: Lassiter and Goodman Model Three Signals Normal Distribution Lambda 4
Low Cost Theta Posterior Source = Signal Specific Signal 0, Signal 1, Signal 2, Theta Distr Rel = True Theta Distr Uniform Receiver Hierarchy

Figure 2.26: Lassiter and Goodman Model Three Signals Normal Distribution Lambda 4
Low Cost Theta Posterior Source = Signal 1, Signal 2 Theta Distr Rel = True Theta Distr Uniform Receiver Hierarchy
2.7 Extending the Model by Pragmatic Ascent, Part II

There is an alternative definition of the sender-receiver hierarchy that is at levels $\leq 2$ at least numerically equivalent to Lassiter and Goodman’s original model. Here we define the initial receiver level $\rho_0$ as before in (2.30). We think of the sender, for every signal other than the null signal, as trying to convey information about both $h$ and the threshold of application of that signal. For the null signal, we think of the sender as before: trying to convey information only about $h$. Thus for $\text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\}$ we can define an $h, \theta$-sensitive sender as:
To see that for \( n = 1 \), (2.48) is equivalent to (2.27) if we assume \( p(\theta_1, \theta_2) \) is uniform over \([a, b]\), consider (2.27b); here the result of marginalizing and normalizing the \( u_1 \) output of level 0 is:

\[
\frac{\int_a^b \rho_0(h, \theta_1, \theta_2|u_1)\,d\theta_2}{\int_a^b \int_a^b \rho_0(h, \theta_1, \theta_2|u_1)\,d\theta_1\,d\theta_2} = \frac{\int_a^b \rho_0(h, \theta_1, \theta_2|u_1)\,d\theta_2}{\int_a^b \int_a^b \rho_0(h, u_1|\theta_1) \times p(\theta_1, \theta_2)\,d\theta_1\,d\theta_2} = \frac{\int_a^b \rho_0(h, \theta_1, \theta_2|u_1)\,d\theta_2}{\int_a^b \frac{1}{(b-a)^2} \int_a^b \rho_0(h|u_1, \theta_1)\,dhd\theta_2} = (b-a) \times \int_a^b \rho_0(h, \theta_1, \theta_2|u_1)\,d\theta_2
\]  

(2.49)

so assuming \( p(\theta_1, \theta_2) \) is uniform over \([a, b]\) (2.27b) for \( n = 1 \) is
\[ \sigma^*_1(u_1, h, \theta_1) = \frac{\left( \int_{a}^{b} \int_{a}^{b} \rho_0(h, \theta_1, \theta_2|u_1) d\theta_2 \right)^{\lambda}}{e^{\lambda C(u_1)}} \]

\[ = \frac{(b-a) \times \int_{a}^{b} \rho_0(h, \theta_1, \theta_2|u_1) d\theta_2)^{\lambda}}{e^{\lambda C(u_1)}} \]

\[ = (b-a)^{\lambda} \times \frac{\left( \int_{a}^{b} \rho_0(h, \theta_1, \theta_2|u_1) d\theta_2 \right)^{\lambda}}{e^{\lambda C(u_1)}} \] (2.50)

and likewise for (2.27c)-(2.27e)

Now (2.48a) for \( n = 1 \) is:

\[ \sigma^*_1(u_0, h) = \frac{\left( \int_{a}^{b} \int_{a}^{b} \rho_0(h, \theta_1, \theta_2|u_0) d\theta_1 d\theta_2 \right)^{\lambda}}{(b-a)^2} \]

\[ = \frac{1}{(b-a)^{2\lambda}} \times \frac{\left( \int_{a}^{b} \int_{a}^{b} \rho_0(h, \theta_1, \theta_2|u_0) d\theta_1 d\theta_2 \right)^{\lambda}}{e^{\lambda C(u_0)}} \] (2.51)

and (2.48b) for \( n = 1 \) is

\[ \sigma^*_1(u_1, h, \theta_1) = \frac{\left( \int_{a}^{b} \rho_0(h, \theta_1, \theta_2|u_1) d\theta_2 \right)^{\lambda}}{e^{\lambda C(u_1)}} \]

\[ = \frac{1}{(b-a)^{\lambda}} \times \frac{\left( \int_{a}^{b} \rho_0(h, \theta_1, \theta_2|u_1) d\theta_2 \right)^{\lambda}}{e^{\lambda \times C(u_1)}} \] (2.52)

and likewise for (2.48c) - (2.48e). But since \( \sigma_1(h, \theta_1, \theta_2, u_1) \) is:

89
\[
\sigma_1(u_1|h, \theta_1, \theta_2) = \frac{\sigma_1^*(u_1, h, \theta_1)}{\sigma_1^*(u_0, h) + \sigma_1^*(u_1, h, \theta_1) + \sigma_1^*(u_2, h, \theta_2) + \sigma_1^*(-u_1, h, \theta_1) + \sigma_1^*(-u_2, h, \theta_2)} \\
= \frac{\sigma_1^*(u_1, h, \theta_1)}{\sigma_1^*(u_0, h) + \sigma_1^*(u_1, h, \theta_1) + \sigma_1^*(u_2, h, \theta_2) + \sigma_1^*(-u_1, h, \theta_1) + \sigma_1^*(-u_2, h, \theta_2)} \times \frac{(b - a)^{2\lambda}}{(b - a)^{2\lambda}}
\]

(2.53)

then when (2.51) is in the denominator of (2.53) it becomes equivalent to (2.27a) for \(n = 1\) under the assumption that \(p(\theta_1, \theta_2)\) is uniform over \([a, b]\). Likewise when (2.52) is in the numerator and denominator of (2.53) it becomes (2.50), which is itself equivalent to (2.27b) for \(n = 1\) under the assumption that \(p(\theta_1, \theta_2)\) is uniform over \([a, b]\); and similarly for all of the other terms in the denominator of (2.53). Similar remarks apply to \(\sigma_1(u_0, h, \theta_1, \theta_2)\), \(\sigma_1(u_2, h, \theta_1, \theta_2)\), \(\sigma_1(-u_1, h, \theta_1, \theta_2)\), and \(\sigma_1(-u_2, h, \theta_1, \theta_2)\). Since the definition of the receiver remains unchanged, for \(n = 2\) the interpretation is no different from the original model.

We see the results from our \(h, \theta\)-sensitive sender-receiver hierarchy when \(\text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\}\) in Figure (2.28), and when \(\text{ALT} = \{u_0, u_1, \neg u_1\}\) in Figure (2.29). (When \(\text{ALT} = \{u_0, u_1, \neg u_1\}\) the sender model consists of the appropriately modified versions of (2.48a), (2.48b), and (2.48d).) For similar reasons as with the original sender model, varying the parameters of the overall model—that is, the cost coefficient or the choice parameter—will not help to solve either of our problems, and neither will appeal to a non-linear cost function that monotonically increases with length. With regards to cost, only by assuming that negation has a negative cost can we solve either problem.

On the other hand, appealing to alternative utterances—comparing \(\rho_n(h|u_2)\) when \(\text{ALT} = \{u_0, u_1, u_2\}\) with \(\rho_n(h|\neg u_1)\) when \(\text{ALT} = \{u_0, u_1, \neg u_1\}\)—appears at least close to solving our
\( \lambda = 4.0, C(u_0) \approx 0.0, C(u_n) \approx 1.0, C(\neg u_n) \approx 1.33 \)

Figure 2.28: Lassiter and Goodman Model Three Signals with Not Normal Distribution
Lambda 4 Low Cost Theta Sensitive Theta Posterior Source = Signal Specific Signal 0, Signal 1, Signal 2, Signal Not 1, Signal Not 2 Receiver Hierarchy

\( \lambda = 4.0, C(u_0) \approx 0.0, C(u_n) \approx 1.0, C(\neg u_n) \approx 1.33 \)

Figure 2.29: Lassiter and Goodman Model Two Signals with Not Normal Distribution
Lambda 4 Low Cost Theta Sensitive Theta Posterior Source = Signal Specific Signal 0, Signal 1, Signal Not 1 Receiver Hierarchy
first problem. (When $\text{ALT} = \{u_0, u_1, u_2\}$ the sender model consists of the appropriately modified versions of (2.48a), (2.48b), and (2.48c). We show the results of our $h, \theta$-sensitive sender-receiver hierarchy when $\text{ALT} = \{u_0, u_1, u_2\}$ in Figure (2.30).) And similarly, comparing $\rho_n(h|u_1)$ when $\text{ALT} = \{u_0, u_1, u_2\}$ with $\rho_n(h|\neg u_1)$ when $\text{ALT} = \{u_0, u_1, \neg u_1\}$, or when $\text{ALT} = \{u_0, \neg u_1\}$ appears to close to solving our second problem. However, as we see when comparing (2.29) and (2.30), at the higher levels $n \geq 4$ where

$$\argmax_h \rho_n(h|\neg u_1) > \argmax_h \rho_n(h|u_2)$$  \hspace{1cm} (2.54)$$

(and thus $\neg u_1$ has the desired weaker interpretation than $u_2$), the interpretation of $\neg u_1$ is too strong: there is almost no probability that $h$ is greater than the median. However, intuitively there seems to be no contradiction at all in saying ‘Feynman is not tall, but he’s taller than average.’ We could try lowering the cost, but as we see in Figure (2.31), at these parameter settings $\rho_n(h|u_1)$ is much too weak, especially at the higher levels where $\rho_n(h|\neg u_1)$
is appropriately weak. If you think I’m choosing between ‘tall’, ‘not tall’, or nothing at all, (perhaps I know it’s quite appealing to you to date someone tall, and someone short is utterly unacceptable, so that ALT is ‘tall’, ‘not tall’, or nothing at all), you still do not interpret ‘tall’ to mean ‘not short’, as in Figure (2.31). In this regard it seems important to remember that according to the proposed solution, when we ask ourselves whether if from hearing something called ‘not tall’ we can infer that it is ‘short’, we are comparing what we would infer about something’s height from a speaker choosing to call it ‘not tall’ when we assume the speaker is choosing from among ‘tall’, ‘not tall’, or saying nothing at all, to what we would infer about something’s height from a speaker choosing to call it ‘short’ when we assume the speaker is choosing from among ‘short’, ‘tall’, or saying nothing at all. But getting the right results from our $h, \theta$-sensitive sender model also then requires that what we would infer about something’s height from a speaker choosing to call it ‘tall’ when we assume the speaker is choosing from among ‘tall’, ‘not tall’, or nothing at all, is entirely too weak. Furthermore, at these parameter settings, at levels $n$ such that $\rho_n(h|\neg u_1)$ is appropriately weak, $\rho_n(\theta_1|u_1)$ is simply implausible: upon hearing someone called ‘tall’, we do not think that the speaker’s threshold for tallness is most likely below the median height for persons; this problem persists even at higher cost settings.

Now, we might try to think of these comparisons between ‘not tall’, ‘tall’, and ‘short’ as resulting from a process of lexicalization such that $\rho_n(h, \theta_1|\neg u_1)$ when ALT = \{u_0, u_1, \neg u_1\} becomes lexically fixed and $\rho_n(h, \theta_1|u_1)$ and $\rho_n(h, \theta_2|u_2)$ when ALT = \{u_0, u_1, u_2\} become lexically fixed. We might then keep Lassiter and Goodman’s pragmatic hierarchy, in one of the versions defined above, and use the new lexicalized items at $\rho_0$, and thus not end up with
an overly weak interpretation of $u_1$ when $\text{ALT} = \{u_0, u_1, \neg u_1\}$. However, we then seem to have given up on the Kennedy-style threshold semantics altogether. Furthermore, it seems \textit{ad hoc} to think that lexicalization does not happen for $u_1$ when $\text{ALT} = \{u_0, u_1, \neg u_1\}$ but only when $\text{ALT} = \{u_0, u_1, u_2\}$.

What about the various ways of privileging the $\theta_1$ and $\theta_2$ posterior distributions, in combination with our $h, \theta$-sensitive sender? That is, we could use (2.39) or (2.38) in combination with (2.48) when $\text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\}$, or (2.41) or (2.42) in combination with appropriately modified versions of (2.48a), (2.48b), and (2.48d) when $\text{ALT} = \{u_0, u_1, \neg u_1\}$.

However, there is a certain incoherence about the idea, since the posterior $\theta$ distributions are generated by appealing to the unnegated signals, but for each negated signal the sender attempts to convey information about $\theta$ for that signal. Consider $\neg u_1$: in a sense, the receiver at $n$ ignores what he at $(n-2)$ thought he could infer about $\theta_1$ given that the $(n-3)$ sender used $\neg u_1$ and instead draws on what he at $(n-2)$ thought he could infer about $\theta_1$ given that the $(n-3)$ sender used $u_1$, even as he assumes at $n$ that the sender used $\neg u_1$. Even worse,
he continues to think that the \((n - 3)\) sender (and the \((n - 1)\) sender) was trying to convey information about \(\theta_1\) using \(\neg u_1\); at \(n\) the receiver thinks ‘Here is the probability that the sender at \((n - 1)\) would use \(\neg u_1\) given \((h, \theta_1, \theta_2)\), trying as he is to convey information about \(h\) and \(\theta_1\), and here is the probability of \(\theta_1\) at \((n - 2)\) given that the sender at \((n - 3)\) did not use \(\neg u_1\) but instead used \(u_1\).’

Now, one might think that since the level \(n\) receiver thinks of the \((n - 1)\) sender as thinking of the \((n - 2)\) receiver as gaining information from the \((n - 3)\) sender’s utterance of \(\neg u_1\) conditional on the \((n - 4)\) receiver’s posterior \(u_1\) probabilities for \(\theta_1\), so also he ought to gain information from the \((n - 1)\) sender’s utterance of \(\neg u_1\) conditional on the \((n - 2)\) receiver’s posterior \(u_1\) probabilities for \(\theta_1\). However, of course, this inductive step has to bottom out at a justification for why we should think of the level 2 receiver as gaining information from the level 1 sender’s utterance of \(\neg u_1\) conditional on the level 0 receiver’s posterior \(u_1\) probabilities for \(\theta_1\), instead of his posterior \(\neg u_1\) probabilities for \(\theta_1\). And there seems no good reason for doing so, since by stipulation at level 0 all signals carry the same information about all thresholds. (And of course, all this is not to mention the earlier noted problem that privileging \(\rho_n(\theta_1|u_1)\) when \(\text{ALT} = \{u_0, u_1, \neg u_1\}\) (and additionally \(\rho_n(\theta_2|u_2)\) when \(\text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\}\)) seems poorly pragmatically motivated.) Finally, the resulting posterior distributions solve neither our first nor our second problem:

We should note that the rest of our options do not produce solutions: neither our alternative hierarchy using \(\rho_n(h|u_0)\) for \(n \geq 2\) along the lines of (2.33), when \(\text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\}\) or \(\text{ALT} = \{u_0, u_1, \neg u_1\}\), nor the other ways of not privileging \(\theta\) sources as in (2.44) or (2.45) when \(\text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\}\) or (2.43) when \(\text{ALT} = \{\).
Equation: \( \lambda = 4.0, C(u_0) \approx 0.0, C(u_n) \approx 1.0, C(\neg u_n) \approx 1.33 \)

Figure 2.32: Lassiter and Goodman Model Three Signals with Not Normal Distribution Lambda 4 Low Cost Theta Sensitive Theta Posterior Source = Signal 1, Signal 2 Receiver Hierarchy

Figure 2.33: Lassiter and Goodman Model Three Signals with Not Normal Distribution Lambda 4 Low Cost Theta Sensitive Theta Posterior Source = Signal Specific Signal 0, Signal 1, Signal 2, Signal 1, Signal 2 Receiver Hierarchy
Figure 2.34: Lassiter and Goodman Model Two Signals with Not Normal Distribution
Lambda 4 Low Cost Theta Sensitive Theta Posterior Source = Signal Specific Signal 0, Signal 1, Signal 1 Receiver Hierarchy

Figure 2.35: Lassiter and Goodman Model Two Signals with Not Normal Distribution
Lambda 4 Low Cost Theta Sensitive Theta Posterior Source = Signal 1 Receiver Hierarchy
\{u_0, u_1, \neg u_1\}, produce acceptable results in our \(h, \theta\)-sensitive sender-receiver hierarchy. We do observe results closer to the intuitively correct interpretation when we use truncated normal distributions for \(\theta\) at level \(n = 0\), especially when we use (2.39) or (2.44); see Figures (2.36) and (2.37). However, even here at the lower levels \(\neg u_1\) and \(u_2\) are interpreted too similarly, and at higher levels where \(\neg u_1\) and \(u_2\) have sufficiently distinct interpretations, \(\neg u_1\) has too strong an interpretation and too narrow an interpretation. *Mutatis mutandis* for \(\neg u_2\) and \(u_1\). We also notice some seeming artifacts in the interpretations of \(u_2\) and \(u_1\) at \(n = 8, 10\). We might search for model parameters that eliminate these problems; or we might try having the receiver assume that \(\theta_1 > \theta_2\). But this seems like the only addition of so many epicycles in order to save the model, and should instead motivate us to find an alternative model.

What about privileging the \(\theta_1\) posterior distributions of \(u_1\) given that \(\text{ALT} = \{u_0, u_1, \neg u_1\}\), in combination with the strategy of appealing to alternative utterances, and ascending the sender-receiver hierarchy using our \(h, \theta\)-sensitive sender? As we see in Figure (2.34) and
(2.35), the resulting interpretations of \( \neg u_1 \) under either privileging method are too strong for this strategy to succeed.

### 2.8 An Alternative Bayes-Grice Model

#### 2.8.1 Motivation: Unrealistic Sender Behavior

If our \( h, \theta \)-sensitive sender model does not solve our two problems, perhaps there are other sender models that we might pursue? In this regard we note one particular problem with the original sender model, and see how we might modify the sender model in light this problem. Consider when \( ALT = \{ u_0, u_1, \neg u_1 \} \). If \( \theta_1 \leq h \) then \( \rho_0(h|\neg u_1, \theta_1) = 0 \), per (2.14). Then per (2.17), \( \sigma_1^*(\neg u_1, h, \theta_1) = 0 \). Since \( \text{INFO}(u_0, h) = \ln(\rho_0(h|u_0)) = \ln(\phi(h)) \), if we let \( C(u_0) = 0 \) (since the production cost of saying nothing is nothing) then \( U_{\sigma_1}(u_0, h) = \ln(\phi(h)) \). Then we have:
\[
\sigma_1(u_1|h, \theta_1) = \frac{\sigma_1^\ast(u_1, h, \theta_1)}{\sigma_1^\ast(u_0, h) + \sigma_1^\ast(u_1, h, \theta_1) + \sigma_1^\ast(\neg u_1, h, \theta_1)}
\]

\[
= \frac{\sigma_1^\ast(u_1, h, \theta_1)}{\sigma_1^\ast(u_0, h) + \sigma_1^\ast(u_1, h, \theta_1)}
\]

\[
= \frac{\left(\frac{\phi(h)}{\int_{\theta_1}^\infty \phi(h)dh}\right)^\lambda}{\left(\frac{\phi(h)}{\int_{\theta_1}^\infty \phi(h)dh}\right)^\lambda} + e^{\lambda \ln(\phi(h))}
\]

\[
= \frac{1}{1 + e^{\lambda C(u_1)} \times (\frac{\phi(h)}{\int_{\theta_1}^\infty \phi(h)dh})^\lambda}
\]

\[
= \frac{1}{1 + e^{\lambda C(u_1)} \times \left(\int_{\theta_1}^\infty \phi(h)dh\right)^\lambda}
\]

Else, if \(\theta_1 > h\), then \(\rho_0(h|u_1, \theta_1) = 0\) and \(e^{\lambda (\ln(\rho_0(h|u_1, \theta_1)) - C(u_1))} = 0\) so \(\sigma_1(u_1|h, \theta_1) = 0\). Notice, however, that for a fixed \(\theta_1\), \(\sigma_1(u_1|h, \theta_1)\) is constant for all \(h \geq \theta_1\), since \(\phi(h)\) drops out. It follows that, as can be seen in Figure (2.38), for a fixed \(\theta_1\), no matter how much greater \(h\) is than \(\theta\), \(\sigma_1\) has the same probability of sending \(u_1\). This seems counterintuitive; if the

---

16Likewise, for a fixed \(\theta\) we cannot make \(\sigma_1\) certain for every \(h \geq \theta\) to send \(u_1\)—that is, for every \(h \geq \theta\)
question under discussion is whether John is tall or not, for a given threshold for ‘tall’, we
would expect that the taller John is, the more likely John is to be called ‘tall’. Since $\sigma_1$ is
an abstraction in the mind of $\rho_2$ this is not directly an objection to the model, but it does
highlight the extent to which the model, if it accurately predicts listener’s interpretation
of gradable adjectives, depends on the listener having assumptions about the speaker that
do not match up to what intuitively seems like actual speaker behavior: it seems intuitive
that actual speakers do not use gradable adjectives with determinate thresholds below which
the adjective is certain never to apply and above which the adjective is always equally (but
perhaps not certainly) likely to apply.\footnote{I do not think it will be enough to merely point out that the threshold may shift from context to context depending on factors in addition to the comparison class; if that defense of thresholds is not to capitulate and become a probabilistic model it must maintain that those factors are a part of the truth conditions of the utterance, since the point of thresholds is to provide contextually sensitive truth conditions. Even more problematic for such a view will be the trade-off of vagueness for massive ambiguity, since different uses of ‘tall’ will require different thresholds from context to context; this problem remains even if we find a way to keep these other factors out of the truth conditions of ‘tall’. A probabilistic model has in a sense built in such factors by assuming that they affect sending probabilities to a greater or lesser degree as the height moves away from a tipping point. Perhaps it is odd that I talk about truth conditions, since I am giving a semantic theory not in terms of truth conditions; but the point remains even if we don’t have truth conditions in our semantic theory: as long as we require precise sending thresholds we will require lexical ambiguity instead of vagueness.} Again, even if speakers actually don’t behave as $\sigma_1$ senders, so long as receivers behave as $\rho_2$ receivers, the model has done its job; we should then look for the effects of systematic error between actual receiver assumptions and actual speaker behavior. Another option is to think of actual receivers as $\rho_4$ receivers, and see if

\begin{align*}
\text{we cannot make } & \sigma_1(u_1|h, \theta_1, \theta_2) \text{ close to 1—unless } \lambda \text{ is sufficiently high and } C(u) \text{ sufficiently low.}
\end{align*}
the model predicts probabilistic $\sigma_3$ senders even for a fixed threshold.\textsuperscript{18}

It is important to distinguish the foregoing feature from the fact that for a fixed $h$, if $h$ is relatively high but $\theta_1$ is relatively low, there is a low probability of $\sigma_1$ using $u_1$. (This can also be seen in Figure (2.38).) This may also seem counterintuitive, but if ‘tall’ meant, roughly, ‘not extremely short’, then even if someone was very tall, we would seem unlikely to call them ‘tall’, as the cost of saying so would outweigh the minimal information gained about $h$. It is one thing for any given threshold to be such that the probability of calling someone ‘tall’ monotonically increases above that threshold (even if to some bound much less than 1); it is another thing for any given threshold to be such that the probability of calling someone ‘tall’ approaches 1 sufficiently far above the threshold. The former seems desirable, the latter does not; the model thus far lacks both features.

\textsuperscript{18}Here, $\rho_2$’s expected sending probabilities $\sigma_1(u_0|h)$, $\sigma_1(u_1|h)$, and $\sigma_1(\neg u_1|h)$ are derived by marginalizing over $\theta_1$:

\begin{align}
\sigma_1(u_0|h) &\propto \int_{-\infty}^{\infty} \sigma_1(u_0|h, \theta_1) \times \rho_0(\theta_1|u_0) \, d\theta_1 \\
\sigma_1(u_1|h) &\propto \int_{-\infty}^{h} \sigma_1(u_1|h, \theta_1) \times \rho_0(\theta_1|u_1) \, d\theta_1 \\
\sigma_1(\neg u_1|h) &\propto \int_{h}^{\infty} \sigma_1(\neg u_1|h, \theta_1) \times \rho_0(\theta_1|\neg u_1) \, d\theta_1
\end{align}

Note that we must also marginalize over $\theta_2$ when ALT includes $u_2$ or $\neg u_2$. 

102
2.8.2 A Distance-Weighted Informativity Model

In light of the foregoing problem with the sender model, we present here a modified version of Lassiter and Goodman (2017). The level 0 receiver $\rho_0$ decodes as before, as in equations (2.1), (2.2), and (2.3). The intuition behind our modified model is that in deciding which signal to encode a given state as, a sender might care not just about the probability that the receiver decodes that signal as the observed state, but about the probability that the receiver will decode that signal as a state nearby the observed state; the nearer a given decoding state is to the observed state, the more valuable it is to the sender that that signal is decoded as that state. Put another way, the informativity of a signal relative to an observed state and a given threshold for that signal is not to be measured as the probability that the receiver will, using that threshold for that signal, decode that signal as that state, but as the distance-weighted probability that the receiver will, using that threshold for that signal, decode that signal as a state nearby the observed state. We assume here that the weight is a Gaussian function of the distance from the observed state, with the rate of weighting drop-off determined by $\sigma_w$, which we set in our model to 1.0:

$$f_h(x) = \frac{1}{\sqrt{2\sigma_w^2\pi}} \times e^{\frac{-(x-h)^2}{2\sigma_w^2}}$$ (2.57)

We can then replace the definitions of informativity in (2.4a)-(2.4c) and (2.16a)- (2.16b) with:
\[ \text{INFO}(u_0, h) = \ln\left( \int_{-\infty}^{\infty} f_h(x) \times \rho_0(x|u_0) \, dx \right) \quad (2.58a) \]

\[ \text{INFO}(u_1, h, \theta_1) = \ln\left( \int_{-\infty}^{\infty} f_h(x) \times \rho_0(x|u_1, \theta_1) \, dx \right) \quad (2.58b) \]

\[ \text{INFO}(u_2, h, \theta_2) = \ln\left( \int_{-\infty}^{\infty} f_h(x) \times \rho_0(x|u_2, \theta_2) \, dx \right) \quad (2.58c) \]

\[ \text{INFO}(-u_1, h, \theta_1) = \ln\left( \int_{-\infty}^{\infty} f_h(x) \times \rho_0(x|-u_1, \theta_1) \, dx \right) \quad (2.58d) \]

\[ \text{INFO}(-u_2, h, \theta_2) = \ln\left( \int_{-\infty}^{\infty} f_h(x) \times \rho_0(x|-u_2, \theta_2) \, dx \right) \quad (2.58e) \]

and then make the corresponding generalization to levels \( n \) in order to allow for ascent of our sender-receiver hierarchy. To see the effect of distance-weighting, consider Figure (2.39); the area under the red dashed line is greater than the area under the red dotted line, so for \( h_0 = -1.5 \), \( u_0 \) is more informative than \( u_1 \). In contrast, for \( h_1 = 1.5 \), the area under the blue dotted line is greater than the area under the blue dashed line, so \( u_1 \) is more informative than \( u_0 \).\(^{19}\)

If we assume the rest of the model is the same as in Figure (2.38), (that is, we follow (2.34)-(2.30)), the \( \sigma_1 \) results are as in Figure (2.40). As we see there, for a fixed \( \theta_1 \) value, the prob-

\(^{19}\)Note that we could also define a modified \( h, \theta \)-sensitive definition of informativity, assuming the sender desires to convey information about both the height and threshold using ‘tall’. We could do so by defining a weighting function for \( \theta \):

\[ f_{\theta_i}(x) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \times e^{\frac{(x-\theta_i)^2}{2\sigma_i^2}} \quad (2.59) \]

and then define informativity as a combination of the two:
ability that \( \sigma_1 \) will use ‘tall’ monotonically increases as \( h \) increases, as desired. The receiver hierarchy is shown in Figure (2.41); as we see there, our second problem remains even as we ascend the sender-receiver hierarchy: ‘not tall’ means something stronger, but in the opposite direction from, ‘tall’. We see in Figure (2.42) the results when \( \text{ALT} = \{u_0, u_1, u_2, -u_1, -u_2\} \); there too ascending the sender-receiver hierarchy leaves our first problem unresolved, as ‘not tall’ continues to be interpreted more strongly than ‘short’. For the same reasons as before,

We leave this possibility unexplored in the interest of space.
neither changing the choice parameter $\lambda$ nor altering the cost function (unless ‘not’ is assigned a negative cost) will solve either problem. As we see comparing Figure (2.41) with Figure (2.43), neither will appealing to alternative utterances—that is, comparing $u_1$ when $\text{ALT} = \{u_0, u_1, u_2\}$ to $\neg u_1$ when $\text{ALT} = \{u_0, u_1, \neg u_1\}$ to solve our first problem, and comparing $u_2$ when $\text{ALT} = \{u_0, u_1, u_2\}$ to $\neg u_1$ when $\text{ALT} = \{u_0, u_1, \neg u_1\}$ to solve our second problem—solve either problem.

Where do we go from here? We could of course pursue our various options above: First, we might try to define the receiver hierarchy in a unified manner for $n > 0$ as in (2.33) when $\text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\}$ or as in Footnote (11) when $\text{ALT} = \{u_0, u_1, \neg u_1\}$, instead of giving separate definitions for $n = 2$ and $n > 2$ as in (2.28) and (2.29) when $\text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\}$ or as in (2.35) and (2.36) when $\text{ALT} = \{u_0, u_1, \neg u_1\}$. Second, we might try our various ways of privileging the $\theta$-posteriors of the various signals as in (2.38)-(2.45). Third, we might might try having the receiver assume that $\theta_1 > \theta_2$ as in §2.6. Along with that, we might attempt to use a normal as opposed to a uniform prior
\[ \lambda = 4.0, C(u_0) \approx 0.0, C(u_n) \approx 1.0, C(¬u_n) \approx 1.33 \]

\[ \mu = 0.0, \sigma = 1.0, \text{num states} = 160, \text{theta distribution type} = \text{uniform}, \text{theta posterior source} = \text{signal specific}, \text{pragmatic sender type} = \text{modified sensitive} \]
θ-distributions at \( n = 0 \). Fourth, we might attempt to define an \( h, \theta \)-sensitive version of our distance-weighted sender model. We might hope that by some permutation of these various approaches we will solve both of our problems; furthermore, we could try to appeal to alternative utterances along with some permutation of these various approaches. However, these start to seem like only so many epicycles lacking independent motivation except to our problems, and it seems fair to look in other directions.

## 2.9 Non-Uniform \( \rho_0 \theta \) Priors

As noted at the beginning of §2.5, there seems to be good reason to think of degree scales for gradable adjectives as abstractions from actual measures of physical properties: cases like the speed of light, or absolute zero, seem to be cases in which the physical property ostensibly being talked about has an upper or lower limit that would seem to entail that the corresponding degree scale is upper or lower closed, contra the fact that ‘perfectly fast’ and
‘perfectly cold’ are unavailable. We might then instead think of degree scales as degrees of
deviation from some measure of central tendency, and the prior distribution of the property
being talked about as distributions over degrees of deviation from the central tendency, at
least for relative gradable adjectives. This has the effect of transforming the distribution of
the property being talked about into a version of the standard normal distribution, as the
degrees of deviation are standardized. We might also distinguish, as suggested by Kennedy
(2007), between whether the degree scale is bounded or unbounded, and whether it is open
or closed: the normal distribution is unbounded, but we can also define a truncated normal
distribution that is bounded:

$$\phi^*(h) = \frac{1}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} \phi\left(\frac{h-\mu}{\sigma}\right)$$ (2.61)

where $\phi(x)$ and $\Phi(x)$ are the probability density and cumulative distribution functions,
respectively, for the standard normal distribution, and $-\infty \leq a < b \leq \infty$. We can let
$b = \infty$ so that $\Phi\left(\frac{b-\mu}{\sigma}\right) = \Phi(\infty) = 1$, and let $a = 0$. If the degree scale is defined only for
$0 < h$ then even if $\phi^*(0) > 0$, $\Phi^*(\infty) = 1$, so we still have a proper probability distribution.
We would then have a lower bounded, but open degree scale, in accordance with Kennedy
(2007)’s observation that relative gradable adjectives have degree scales that are both lower
and upper open. We might then claim that the unavailability of ‘perfectly tall’ is due to
the openness of the scale on both ends, there being no endpoints to move probability mass
onto or off of. On the other hand, we might let the degree scale be defined for $0 \leq h$;
again, even if $\phi^*(0) > 0$, $\Phi^*(\infty) = 1$, so we again have a proper probability distribution. We
would in this case have a lower bounded, but closed degree scale. This would allow us to
offer different explanations for the unavailability of ‘perfectly cold’ and ‘perfectly tall’: The former has a degree scale which is lower unbounded, there being intuitively at least (contra the deliverances of modern thermodynamics) no limit on the number of standardized degrees of deviation below the central tendency that an object’s temperature might be. In contrast, the latter has a degree scale which is lower bounded, since whatever the central tendency of the heights of objects in the comparison class, there is a lower limit to the number of standardized degrees of deviation below that central tendency that an object’s height might be: the central tendency will be at a finite height, and no object (or at least no object that is the kind of object that might have height) can have a height less than or equal to 0. Nevertheless, the degree scale is also lower open, since (again) no object (or at least no object that is the kind of object that might have height) can have a height less than or equal to 0.\textsuperscript{20}

\textsuperscript{20}In the case of ‘completely tall’ we might distinguish between a lower bounded scale that has 0 probability density at the lower bound, and a lower bounded scale that has undefined probability density at the lower bound; if we think of the intuition that no object that is the kind of object that can have height has 0 height as driving the topology of the scale, then it seems we will want to model the degree scale as the former kind of scale. On the other hand, if we think of the intuition that no object that is the kind of object that can have height has 0 height as deriving from the topology of the scale, we might want to model the degree scale as the latter kind of scale. Either way, however, if we think of the effect of ‘perfectly’ as forcing the receiver to place all of the probability mass on the endpoint of the degree scale, the calculation of meaning will go wrong: in the former case, the probability mass will be 0, in the latter case, the probability mass will be undefined. Note that this understanding of ‘completely’ requires appealing to generalized functions, as we will be attempting to define a probability density function that sums to 1, but is everywhere except the endpoint equal to 0, and infinitely dense at the endpoint.
Now, the original motivation from Lassiter and Goodman (2017) for uniform prior distributions for threshold values was the preservation of interpretive flexibility; if $\theta_{tall}$ were biased towards human heights, then *tall skyscraper* would end up improperly interpreted. If, however, for relative gradable adjectives we can think of degree scales as degrees of deviation from some measure of central tendency, and the prior distribution of the property being talked about as distributions over degrees of deviation from the central tendency, then it seems we can have non-uniform prior distributions for threshold values without losing interpretive flexibility: 1.5 standard deviations above the average height of men, will be much less than 1.5 standard deviations above the average height of skyscrapers.

We show in Figure (2.44) the resulting receiver hierarchy when the hierarchy is defined as in (2.27)-(2.30) and $\text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\}$, and $p(\theta_1) = \mathcal{N}(1.5, 1.)$ and $p(\theta_2) = \mathcal{N}(-1.5, 1.)$. As we see there, our first problem seems better addressed than on any of the previous approaches, insofar as at all levels $n$, $\rho_n(h|u_1)$ is stronger than $\rho_n(h|\neg u_2)$, and $\rho_n(h|u_2)$ is stronger than $\rho_n(h|\neg u_1)$.
Our second problem is more difficult to resolve. We see in Figure (2.45) the resulting receiver hierarchy when $\text{ALT} = \{u_0, u_1, \neg u_1\}$ and the hierarchy is defined as in (2.34)-(2.37), and $p(\theta_1) = N(1.5, 1)$. When $n = 0$ $\rho_n(h|\neg u_1)$ is appropriately weak and $\rho_n(h|u_1)$ is appropriately strong. However, at higher levels $\rho_n(h|\neg u_1)$ is nearly as strong as $\rho_n(h|\neg u_1)$. For that matter, we might also quibble with the interpretation when $\text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\}$.

Looking again at Figure (2.44) at higher levels $\rho_n(h|\neg u_1)$ and $\rho_n(h|\neg u_2)$ seems too strong, and there seems an odd shift in $\rho_n(h|u_1)$ and $\rho_n(h|u_1)$ as we ascend the hierarchy, becoming weaker at $n = 2$ and then stronger then on. Now, we could specify that interpreters never ascend that high in the sender-receiver hierarchy, or we could hope that some permutation of the foregoing approaches along with our now appropriately specified prior $\rho_0$ distributions for $\theta_1$ and $\theta_2$ would give us the interpretations we want. To review, those approaches included:

1. Varying the parameters in a appropriate manner.

2. Appealing to alternative utterances
3. Defining the receiver hierarchy in a unified manner for \( n > 0 \) as in (2.33) when \( \text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\} \) or as in Footnote (11) when \( \text{ALT} = \{u_0, u_1, \neg u_1\} \), instead of giving separate definitions for \( n = 2 \) and \( n > 2 \) as in (2.28) and (2.29) when \( \text{ALT} = \{u_0, u_1, u_2, \neg u_1, \neg u_2\} \) or as in (2.35) and (2.36) when \( \text{ALT} = \{u_0, u_1, \neg u_1\} \).

4. Privileging the \( \theta \)-posteriors of the various signals as in (2.38)-(2.45).

5. Having the receiver assume that \( \theta_1 > \theta_2 \) as in § 2.6. Along with that, we might attempt to use a normal as opposed to a uniform prior \( \theta \)-distributions at \( n = 0 \).

6. Using an \( h, \theta \)-sensitive version of our sender-receiver hierarchy.

7. Using an \( h \)-sensitive or \( h \)- and \( \theta \)-sensitive version of our distance-weighted sender model.

However, I think the most important objection to attempting to resolve either of our problems by appropriately specified prior \( \rho_0 \) distributions for \( \theta_1 \) and \( \theta_2 \) is that doing so seems to anticipate a simpler model that entirely lacks \( \theta \) values at all. To see this, consider again the Bayesian reasoning that a hypothetical receiver engages in, when he assumes that \( \text{ALT} = \{u_0, u_1, \neg u_1\} \). Let us write \( P_{\rho_n} \) for \( \rho_n \)'s probability function, in order to distinguish the Bayesian component of our interpretive model from the recursive sender-receiver hierarchy to which we have married it. The probability that \( \rho_n \) assigns to any particular height \( h \) given utterance \( u_0 \) is:

\[
P_{\rho_n}(h|u_0) = \frac{P_{\rho_n}(u_0|h) \times P_{\rho_n}(h)}{P_{\rho_n}(u_0)}
\]

(2.62)

Now, according to our recursive sender-receiver hierarchy\(^{21}\):

\(^{21}\)Note that this raises the possibility of defining a different hierarchy, one that takes the \( n - 2 \) utterance
1. $P_{\rho_n}(h)$ is taken to be $\rho_{n-2}(h|u_0)$.

2. $P_{\rho_n}(u_0|h)$ is taken to be the probability that $\rho_n$ assigns to $\sigma_{n-1}$ uttering $u_0$ given that the height is $h$—that is, $\sigma_{n-1}(u_0|h)$. $\sigma_{n-1}(u_0|h)$ is then taken to be the sum, for every $\theta_1$, of the product of the probability of $\sigma_{n-1}$ uttering $u_0$ given that the height is $h$ and the threshold is $\theta_1$, (that is, $\sigma_{n-1}(u_0,h,\theta_1)$), and the probability that $\rho_n$ assigns to the threshold being $\theta_1$ (which is in our recursive sender-receiver hierarchy defined as $\rho_{n-2}(\theta_1|u_0)$). As an integral that sum is $\int_a^b \sigma_{n-1}(u_0|h,\theta_1) \times \rho_{n-2}(\theta_1|u_0)d\theta_1$.

3. $P_{\rho_n}(u_0)$ is taken to be the probability that $\rho_n$ assigns to $\sigma_{n-1}$ uttering $u_0$—that is, $\sigma_{n-1}(u_0)$. The probability that $\rho_n$ assigns to $\sigma_{n-1}$ uttering $u_0$ at all is then taken to be the sum, for every $h$ and $\theta_1$, of product of $\sigma_{n-1}(u_0,h,\theta_1)$, $\rho_{n-2}(\theta_1|u_0)$, and the probability that $\rho_n$ assigns to the height being $h$, which in our recursive sender-receiver hierarchy is taken to be $\rho_{n-2}(h|u_0)$. As an integral, that sum is $\int_a^b \int_a^b \sigma_{n-1}(u_0|h,\theta_1) \times \rho_{n-2}(\theta_1|u_0) \times \rho_{n-2}(h|u_0)d\theta_1dh$.

Since in the recursive sender-receiver hierarchy $P_{\rho_n}(h|u_0)$ is taken to be $\rho_n(h|u_0)$, we thus have that:

---

probabilities into account. That is, in items 2 and 3, instead of using $\rho_{n-2}(\theta_1|u_0)$, use $\rho_{n-2}(\theta_1)$, which is $\sum_{u\in\text{ALT}} \rho_{n-2}(\theta_1|u) \times \rho_{n-2}(u)$, where $\rho_{n-2}(u)$ is defined as $\sigma_{n-3}(u)$, similarly to $\sigma_{n-1}(u)$ in item 3. Likewise in items 1 and 3 instead of using $\rho_{n-2}(h|u_0)$, use $\rho_{n-2}(h)$, which is $\sum_{u\in\text{ALT}} \rho_{n-2}(h|u) \times \rho_{n-2}(u)$, where again $\rho_{n-2}(u)$ is defined as $\sigma_{n-3}(u)$, similarly to $\sigma_{n-1}(u)$ in item 3. The latter change will not, I think, for $n=2$ be equivalent to the original hierarchy assuming $\rho_0$ has uniform prior probabilities for each utterance and $p(\theta_1)$ is uniform. We could, however, just use $\phi(h)$ all the way up the hierarchy and be consistent with the original model, an additional wrinkle we have not yet considered.
\[ \rho_n(h|u_0) = \frac{\int_a^b \sigma_{n-1}(u_0|h, \theta_1) \times \rho_{n-2}(\theta_1|u_0)d\theta_1 \times \rho_{n-2}(h|u_0)}{\int_a^b \int_a^b \sigma_{n-1}(u_0|h, \theta_1) \times \rho_{n-2}(\theta_1|u_0) \times \rho_{n-2}(h|u_0)d\theta_1dh} \]  

(2.63)

Of course, we can move \(\rho_{n-2}(h|u_0)\) within the integral, and this is equivalent to (2.35a), only adding the marginalization over \(\theta_1\) and the normalization over \(\theta_1\) and \(h\):

\[ \rho_n(h, \theta_1|u_0) \propto \sigma_{n-1}(u_0|h, \theta_1) \times \rho_{n-2}(h|u_0) \times \rho_{n-2}(\theta_1|u_0) \]  

(2.35a, revisited)

Similar remarks apply to \(u_1\) and \(-u_1\), and mutatis mutandis when \(\text{ALT} = \{u_0, u_1, u_2, -u_1, -u_2\}\).

We might therefore express our model for interpretation when \(\text{ALT} = \{u_0, u_1, -u_1\}\) as:

\[
\begin{align*}
\rho_n(h|u_0) &= \frac{\sigma_{n-1}(u_0|h)}{\sigma_{n-1}(u_0)} \times \rho_{n-2}(h|u_0) \quad (2.64a) \\
\rho_n(h|u_1) &= \frac{\sigma_{n-1}(u_1|h)}{\sigma_{n-1}(u_1)} \times \rho_{n-2}(h|u_1) \quad (2.64b) \\
\rho_n(h|-u_1) &= \frac{\sigma_{n-1}(-u_1|h)}{\sigma_{n-1}(-u_1)} \times \rho_{n-2}(h|-u_1) \quad (2.64c)
\end{align*}
\]

Now, consider again \(u_0\): when \(n = 2\), \(\rho_{n-2}(h|u_0) = \phi(h)\), reflecting \(\rho_2\)'s prior knowledge of the height distribution of objects in the comparison class. This will remain fixed invariant of the model parameters or even of the various options regarding \(\text{ALT}\), or how to calculate the \(\theta\)-posteriors for higher levels of \(n\), or any of the other options in the foregoing list. We might then wonder what \(\frac{\sigma_{n-1}(u_0|h)}{\sigma_{n-1}(u_0)}\) looks like, or perhaps what it must look like in order to have plausible interpretations of \(u_0\) when \(n = 2\). But note here that \(\sigma_{n-1}(u_0)\) is simply a
normalizing factor to ensure that \( \int_a^b \rho_n(h|u_0)dh = 1 \), which we can ignore insofar as we focus on the relative probabilities that \( \rho_2 \) assigns to various values of \( h \) given that \( \rho_2 \) thinks that \( \sigma_1 \) utters \( u_0 \). So we can focus instead simply on \( \sigma_{n-1}(u_0|h) \)—that is, simply on how the probability that \( \rho_2 \) assigns to \( \sigma_{n-1} \) using \( u_0 \) varies, as \( h \) varies. *Mutatis mutandam* for \( u_1 \) and \( \neg u_1 \).

Let us set \( \rho_n(h|u_i) \) to \( \rho_0(h|u_0) \) so that we can consistently see the effect of \( \sigma_{n-1}(u_i|h) \) on \( \rho_n(h|u_i) \) for various levels \( n \). Thus, our hierarchy is defined as in (2.34)-(2.37), except that \( \rho_{n-2}(h|u_0) \), \( \rho_{n-2}(h|u_1) \), and \( \rho_{n-2}(h|\neg u_1) \) in (2.35) are replaced by \( \rho_0(h|u_0) \), which is just \( \phi(h) \). We again set \( p(\theta_1) = \mathcal{N}(1.5,1) \). We see the resulting receiver hierarchy in Figure (2.46), and the resulting sender hierarchy in Figure (2.47). As one might expect, as the probability mass of \( \sigma_{n-1}(u_i|h) \) shifts away from the central tendency of \( \phi(h) \), \( \rho_n(h|u_i) \) shifts in the same direction away from the central tendency of \( \phi(h) \). This makes sense: the less likely a speaker is to call someone ‘not tall’ unless that person is well and truly short, the more likely we should be to consider someone they call ‘not tall’ to be well and truly short. Thus, \( \sigma_1(\neg u_1|h) \) has significant probability mass to the right of the inflection point around \(-.75\sigma\), and \( \rho_2(h|\neg u_1) \) is plausibly weak. (See Figure (2.48) for an enlarged view of \( \sigma_1(\neg u_1|h) \).) However, as we ascend the sender hierarchy the probability mass of \( \sigma_{n-1}(u_i|h) \) shifts away from the central tendency of \( \phi(h) \). Thus \( \sigma_9(\neg u_1|h) \) has proportionally less of its probability mass to the right of the main inflection point, which itself has moved to around \(-.9\sigma\), and \( \rho_{10}(h|\neg u_1) \) is implausibly strong. Similar remarks apply to \( \sigma_{n-1}(u_1|h) \) and \( \rho_n(h|u_1) \): as we ascend the sender-receiver hierarchy, the probability mass of \( \sigma_{n-1}(u_1|h) \) moves towards the central tendency of \( \phi(h) \), and \( \rho_n(h|u_1) \) becomes weaker. We thus see that
Figure 2.46: Lassiter and Goodman Model Two Signals with Not Normal Distribution Lambda 4 Low Cost Theta Posterior Source = Signal Specific Signal 0, Signal 1, Signal Not 1 Theta Distr Offset Normal Mod 2 Receiver Hierarchy

generating plausible interpretations for relative gradable adjectives under the Bayes-Grice framework seems to require that the receiver assume roughly sigmoid encoding behavior on the part of the sender, whether or not there is a threshold in the underlying semantics. We might then wonder if there is a way to generate such encoding behavior independently of thresholds at all, and it is to such attempts that I turn next. This may seem to reverse the order of things: we have been assuming a semantics and trying to generate plausible interpretations. Instead we are now assuming we know what a plausible interpretation looks like, and working backwards to a semantics, going by way of a model of sending behavior.
\( \lambda = 4.0, C(u_0) \approx 0.0, C(u_n) \approx 1.0, C(\neg u_n) \approx 0.33 \)

\( \sigma_n(u_0|h) \)
\( \sigma_n(u_1|h) \)
\( \sigma_n(\neg u_1|h) \)

Figure 2.47: Lassiter and Goodman Model Two Signals with Not Normal Distribution Lambda 4 Low Cost Theta Posterior Source = Signal Specific Signal 0, Signal 1, Signal Not 1 Theta Distr Offset Normal Mod 2 Sender Hierarchy

\( \lambda = 4.0, C(u_0) \approx 0.0, C(u_n) \approx 1.0, C(\neg u_n) \approx 1.33 \)

\( \sigma_n(u_0|h) \)
\( \sigma_n(u_1|h) \)
\( \sigma_n(\neg u_1|h) \)

Figure 2.48: Lassiter and Goodman Model Two Signals with Not Normal Distribution Lambda 4 Low Cost Theta Posterior Source = Signal Specific Signal 0, Signal 1, Signal Not 1 Theta Distr Offset Normal Mod 2 Sender Hierarchy Enlarged

118
Chapter 3

Game Theory and Probabilistic Signaling

3.1 Introduction

In Chapter 2 I argued that at least one prominent model due to Lassiter and Goodman (2017) of how we engage in the statistical inferences that we typically do engage in upon hearing a sentence containing a vague predicate such as ‘Feynman is tall’, when married to a common kind of truth-conditional semantics for that sentence such as is found in Kennedy (2007), generates the wrong predictions about the negations of such sentences. That is, such a model predicts that upon hearing ‘Feynman is not tall’, listeners will infer that Feynman is likely to be shorter than if he had been called ‘short’. This seems to be much too strong an interpretation, and I argue in Chapter 2 that getting the model to generate the intuitively correct results, requires that listeners have a fixed distribution for the threshold for tallness, at or above which ‘Feynman is tall’ is true, and below which it is false. This in turn suggested a simpler model of interpretation, on which listeners simply consider the probability that the speaker would have said ‘Feynman is tall’ given that Feynman is a given height, and
the prior probability of Feynman being that height, to generate a posterior probability for
Feynman being that height. Of course, it is implausible that listeners do this unless speakers
actually do increase the probability that Feynman will be called ‘tall’ as Feynman’s height
increases.

However, on the most plausible understanding of what language is, according to which it is
a convention in the game-theoretic sense, just as driving on the left-hand side of the road is
a convention, such encoding behavior is at least initially unexpected, since as Lipman (2009)
proves, probabilistic encoding behavior is never a strict Nash equilibrium. Thus, in this
chapter I explore some attempts to explain in terms of game theory how such probabilistic
encoding behavior might arise, and I extend in two ways the explanation that I find most
successful.

3.2 The Puzzle of Game Theory and Vagueness

In D. Lewis (1969), the author introduced the notion of a signaling game: we imagine a
sender who observes states of nature and chooses a signal to send to a receiver. Given a
signal, the receiver chooses an action to perform. The sender can observe and send a signal,
but cannot act. The receiver can receive a signal and perform an action, but cannot observe.
If the action performed is appropriate to the observed state, the sender and receiver receive
a reward. If not, they receive nothing. Obviously, the challenge for sender and receiver is
to coordinate their choices of signals and actions so as to maximize their rewards. Suppose
there are two possible states of nature, two signals, and two actions. We can represent the
game in extensive form as in Figure 3.1, where N is nature, \( n_1 \) and \( n_2 \) are the possible states of nature, S is the sender, \( m_1 \) and \( m_2 \) are the messages she can send, R is the receiver, and \( a_1 \) and \( a_2 \) are the actions he can take. The 1’s and 0’s at the end of the paths represent rewards given the preceding choices. Thus, \( n_1 \rightarrow m_1 \rightarrow a_1, n_2 \rightarrow m_2 \rightarrow a_2 \) is one arrangement that leads to maximal payoffs; but not the only one: \( n_1 \rightarrow m_2 \rightarrow a_1, n_2 \rightarrow m_1 \rightarrow a_2 \) will do just as well. This is a coordination game, where each player prefers that the other players play a corresponding strategy to the one he plays, no matter which one he chooses.\(^1\) For the former arrangement, that one-one correspondence maps \( n_1 \rightarrow m_1 \rightarrow a_1 \) to \( m_1 \rightarrow a_1 \) and \( n_2 \rightarrow m_2 \rightarrow a_2 \) to \( m_2 \rightarrow a_2 \); for the latter arrangement, it maps \( n_1 \rightarrow m_2 \rightarrow a_1 \) and \( n_2 \rightarrow m_1 \rightarrow a_2 \).

Lewis went on to analyse conventions of language use as conventions of games of pure coordination. To see what this means, consider a coordination game we’ll call *Left-Right*, where player 1 and 2 always receive equal payoffs \((x_1, y_2)\). We present the game in tabular form in Figure 3.2. We might think of this game as analogous to choosing which side of the road to drive on: neither I nor the driver of the oncoming car care which particular side we drive on, just that he choose right if I choose right, and he choose left if I choose left, for otherwise

\(^1\)In a *pure* coordination game, for any action, each player receives the same payoff, (although different actions can have different payoffs.)
Now, a *pure* strategy is one where players always only choose one strategy or another; a mixed strategy is one where a player probabilistically chooses between strategies—in the game above, sometimes choosing A and other times choosing B. A *Nash equilibrium* is an assignment of pure or mixed strategies to each player in the game such that no player is better off by unilaterally deviating from the strategy assigned to him. The game above then has two pure strategy Nash equilibria. Each of these is also a *strict* Nash equilibria, which is an assignment of pure or mixed strategies to each player in the game such that any player is actually *worse* off by unilaterally deviating from the strategy assigned to him. There is one mixed strategy Nash equilibrium in this game: \( s = (0.5A + 0.5B) \).

With these preliminaries in mind we can say that for Lewis, a strict Nash equilibrium of a game of pure coordination is a convention just in case each player adheres to that equilibrium because (i) the rules and payoff structure of game, (ii) the strategic rationality of each player, and (iii) the intention of each player to adhere to that equilibrium, are all items of common knowledge, in the sense that every player knows these items, knows that every other player knows these items, knows that every other player knows that every other player knows them, and so on. The value of such a notion of convention is clear, for we can now explain in formal terms why we drive on the right (or left) side of the road: this is also a convention in a pure
coordination game.

Lewis’ analysis of conventions of language as conventions of a pure coordination game has been highly influential, and a number of linguists and philosophers have attempted to extend the game-theoretic approach to account for puzzles of language such as scalar implicatures and free choice permissions, and more general features of language such as compositionality and context-sensitivity. It would seem reasonable to think that such an approach could also shed light on vagueness. However, Lipman (2009) proves that vagueness is in one sense puzzling on a game-theoretic approach: it is never part of the strict Nash equilibrium of a signaling game of pure coordination, and thus can never be explained as arising from a convention.

What is vagueness in a signaling game? Intuitively, the taller something is relative to a comparison class, the more likely it is to be called tall. Thus, in a signaling game vagueness is not leaving some states unexpressed, nor is it having some states be expressed by more than 1 message; it is having some states be probabilistically expressed by at least one message. Given this notion of vagueness in a signaling game, the proof that it is never part of a strict Nash equilibrium of a signaling game of pure coordination is as follows: let \((m, a^*)\) be a pair of (perhaps mixed, perhaps pure) strategies such that

\[
U_{S,R}(m^*, a^*) = U^* \tag{3.1}
\]

where \(U\) is the supremum of \(U_{S,R}\), the utility function for the sender and receiver. Then \((m^*, a^*)\) is a Nash equilibrium, since any deviation from \(m^*\) cannot raise the payoff for
Now, one response to this problem is to draw on the resources of evolutionary game theory and attempt to explain vagueness in signaling games as a result of the manner in which the
strategies of such games evolve. Even if vagueness is never part of the strict Nash equilibrium of a signaling game, perhaps evolutionary game theory can tell us how players starting from non-optimal strategies might come to engage in vague signaling. There are at least two flavors of evolutionary game theory of interest to us: the replicator dynamics interpretation, and the best response dynamics interpretation. As for the replicator dynamics interpretation, let us return to the *Left-Right* game above. Suppose we initially have a population of individuals, each of whom is either an *A*-player or a *B*-player. The game is played over multiple rounds, and after each round a player can produce an offspring of only the same player type—*A*-players can only create *A*-players, and *B*-players can only create *B*-players. Assume that the average number of offspring of any player of any type after any round is equal to the expected utility of a player of that type when playing against a random player, and that any new player has one and only one parent. If \( a \) and \( b \) are the number of *A*-players and *B*-players, respectively, then since \( 1 \times a/(a+b) + 0 \times b/(a+b) = a/(a+b) \), the average number of offspring of an *A*-player is equal to the proportion of *A*-players, and likewise for *B*-players: \( 0 \times a/(a+b) + 1 \times b/(a+b) = b/(a+b) \). (That is, the number of offspring is calculated just as expected utility would be.)

If we start with 50 *A*-players and 50 *B*-players at round 0, then after round 1 we will have 75 *A*-players and 75 *B*-players, after round 2, 112.5 *A*-players and 112.5 *B*-players, and so on. Evolutionary game theory typically assumes that populations are infinitely large and that random variation is non-existent. Thus, we can say that if the initial distribution is 50:50, it will indefinitely remain so. If we start with 100 *A*-players and 0 *B*-players at round 0, then after round 1 we will have 150 *A*-players and 0 *B*-players, after round 2, 225 *A*-players and 0
$B$-players, and so on; thus, we see that if the initial distribution is 100:0, it will indefinitely remain so. Likewise if the initial distribution is 0:100.

In general, suppose $a > b$. Then since

$$EU(A) = 1 \times a/(a + b) + 0 \times b/(a + b) > EU(B) = 0 \times a/(a + b) + 1 \times b/(a + b) \quad (3.3)$$

and the average number of offspring is equal to the expected utility $EU$, it follows that after each round the percent increase of the $A$-player population will be greater than that of the $B$-player $A$- or $B$-players is a geometric function of the number of rounds, the population will eventually converge at the limit to 100% $A$-players. Likewise if $b > a$.

There are thus only three steady states of the population distribution $A\%$ to $B\%$: 100\%:0\%, 50\%:50\%, and 0\%:100\%. Imagine, however, an invasion of a population in the second distribution by an arbitrarily small group of $A$-players: as now $a > b$, the population will eventually converge to only $A$-players. The second distribution is not an evolutionarily stable state, defined in John Maynard Smith (1982) as a population state such that the initial distribution is restored by natural selection after a limited disturbance. The first distribution is such a state, however: so long as the number of $B$-player invaders is strictly less than the total of an $A$-player only population, the population will eventually converge again to an $A$-player only population. Likewise for the third distribution.

Under the replicator dynamics interpretation of evolutionary game theory, the notion of an evolutionary stable state gives rise to the notion of an evolutionary stable strategy, for which there are two distinct definitions: First, following J Maynard Smith and Price (1973) we can
say that a strategy $s$ is evolutionarily stable if and only if for every player $p$:

$$\forall t \neq s : (U_p(s, s) > U_p(t, s) \lor (U_p(s, s) = u(t, s) \land U_p(s, t) > U_p(t, t)))$$ (3.4)

In words: $s$ is such that it is always either better to stay with $s$ if the other player stays too or both the same to switch if he stays and better to stay if he switches. Second, following Thomas (1985), we can say that a strategy $s$ is evolutionarily stable if and only if:

$$\forall t \neq s : (U_p(s, s) \geq U_p(t, s) \land U_p(s, t) \geq U_p(t, t))$$ (3.5)

In words: $s$ is such that it is never better to switch if the other player stays and always better to stay if he switches. There is a third definition in John Maynard Smith (1982) that is equivalent to the first definition: we can say that a strategy $s$ is evolutionarily stable if and only if:

$$\forall t \neq s : (U_p(s, s) \geq U_p(s, t) \land U_p(s, s) = U_p(t, s) \rightarrow U_p(s, t) > U_p(t, t))$$ (3.6)

In words, $s$ is such that it is never better to switch if the other player stays and if ever the same to switch if he stays then better to stay if he switches.\footnote{The equivalence follows from propositional logic and the fact that $a = b$ and $a > b$ are never both true.}

The first and the third definitions are not equivalent to the second definition; both pure strategy Nash equilibria in the game above are evolutionarily stable strategies under the first and third definitions, but not under the second definitions, since for neither A nor B is
it better to stay if the other player switches. The mixed strategy Nash equilibrium for the
game above is not an evolutionarily stable strategy under any of the definitions: let \( s = (0.5A + 0.5B) \) and \( t = (0.7A + 0.3B) \); then since \( U_p(s,s) = 0.50, U_p(t,s) = 0.50, U_p(t,t) = 0.58, \) and \( U_p(s,t) = 0.50, \) it follows that \( U_p(s,s) = U_p(t,s) \) and hence \( U_p(s,s) > U_p(t,s) \), but \( (U_p(s,t) > U_p(t,t)) \). The last fact by itself entails that \( s \) is not an evolutionarily stable strategy under the second definition, and the three facts together entail that \( s \) is not an evolutionarily stable strategy under the first definition, (and hence the third one too). Since each of the three definitions entails that an evolutionary stable strategy under any of the three definitions is a Nash equilibrium, this overall demonstrates that the evolutionarily stable strategies under any of the three definitions are a proper subset of the Nash equilibria.

Finally, a strict Nash equilibrium is automatically an evolutionarily stable strategy under the first definition by its first disjunct, and hence is an evolutionarily stable strategy under the third definition too. Not so for the second definition, since either strategy in the game above is a strict Nash equilibrium, but not an evolutionary stable strategy under the third definition. In the game illustrated in Figure 3.3, we see that \( B \) is a evolutionarily stable strategy under the first definition, but also a non-strict Nash equilibrium. Hence the strict Nash equilibria are a proper subset of the evolutionarily stable strategies, under the first and third definitions. (Of course, there is a clear correspondence between the evolutionarily stable states and the evolutionarily stable strategies: we can turn any evolutionarily stable state into a (perhaps mixed) evolutionarily stable strategy by interpreting the population distribution as the mixing proportions, and vice-versa.)

The appeal of evolutionarily stable strategies for attempts to explain vagueness in game-
Figure 3.3: B is a non-strict evolutionarily stable strategy

theoretic terms is that even if vagueness in signaling games is not part of the strict Nash equilibrium (and thus we cannot explain its existence as a feature of language in terms of its utility-maximization), perhaps there is a way to explain how it arises under the replicator dynamics. However, Selten (1978) proves that for asymmetric games, a strategy is evolutionarily stable under the replicator dynamics (for the first and third definition above) if and only if it is a strict Nash equilibrium. Since the strategy sets of the sender and receiver are distinct (that is, they choose between different sets of moves), signaling games are asymmetric games. Thus, if vagueness is not part of the strict Nash equilibrium, it is not predicted to arise either on the replicator dynamics!

3.3.2 Best Response Dynamics

We derived these notions of evolutionarily stable states and evolutionarily stable strategies by appeal to an interpretation of evolutionary game theory that made no mention of any player’s knowledge of the reward structure of the game or capacity to engage in rational deliberation and then arrive at the correct decision-theoretic outcome. Game theory has traditionally, however, been couched in precisely these terms, albeit with the rather unrealistic assumptions that each player has logical omniscience and the ability to reason to infinite strategic depth,

\[ \begin{array}{c|cc}
     & A & B \\
   \hline
   A & (0,0) & (1,1) \\
   B & (1,1) & (1,1) \\
\end{array} \]

The asymmetry is clearer if we think of Nature as a third player: one who picks a state that the sender observes.
and the further assumption that that is common knowledge.

Enter best-response dynamics, which takes a middle ground between the replicator dynamics and traditional game theory. To illustrate the difference, consider the game of Rochambeau in Figure 3.4:

The single Nash equilibrium of this game is the mixed strategy

\[ s = \frac{1}{3} \times Rock + \frac{1}{3} \times Paper + \frac{1}{3} \times Scissors \tag{3.7} \]

However, \( s \) is not evolutionarily stable under the replicator dynamics: suppose we introduce a small a population of rock players in a population in the equilibrium state. The next round of the game will see an increased reproduction rate for paper players and a decreased rate for scissors players; this will continue until the excess of paper players leads to an increased reproduction rate for scissors players, eventually leading to an excess of rock players, and so on. This will lead to a cycle of paper, then scissors, then rock dominating the population, with the dominance in one round always greater than that in the previous one.

In fact we can see that \( s \) is not an evolutionarily stable strategy under either definition. In Figure 3.5 we let \( s = \frac{1}{3} \times Rock + \frac{1}{3} \times Paper + \frac{1}{3} \times Scissors \) and \( t = \frac{1}{2} \times Rock + \frac{1}{4} \times Paper + \frac{1}{4} \times Scissors \).
\begin{array}{c|ccc}
(s,s) & 1/3 & 1/3 & 1/3 \\
1/3 & 1/9 & 1/9 & 1/9 \\
1/3 & 1/9 & 1/9 & 1/9 \\
1/3 & 1/9 & 1/9 & 1/9 \\
\end{array}
\times
\begin{array}{c|ccc}
Rock & 1 & 0 & 2 \\
Paper & 2 & 1 & 0 \\
Scissors & 0 & 1 & 2 \\
\end{array}

\begin{array}{c|ccc}
(t,s) & 1/3 & 1/3 & 1/3 \\
1/2 & 1/6 & 1/6 & 1/6 \\
1/4 & 1/12 & 1/12 & 1/12 \\
1/4 & 1/12 & 1/12 & 1/12 \\
\end{array}
\times
\begin{array}{c|ccc}
Rock & 1 & 0 & 2 \\
Paper & 2 & 1 & 0 \\
Scissors & 0 & 1 & 2 \\
\end{array}

\begin{array}{c|ccc}
(t,t) & 1/2 & 1/4 & 1/4 \\
1/2 & 1/4 & 1/8 & 1/8 \\
1/4 & 1/8 & 1/16 & 1/16 \\
1/4 & 1/8 & 1/16 & 1/16 \\
\end{array}
\times
\begin{array}{c|ccc}
Rock & 1 & 0 & 2 \\
Paper & 2 & 1 & 0 \\
Scissors & 0 & 1 & 2 \\
\end{array}

\begin{array}{c|ccc}
(s,t) & 1/2 & 1/4 & 1/4 \\
1/3 & 1/6 & 1/12 & 1/12 \\
1/3 & 1/6 & 1/12 & 1/12 \\
1/3 & 1/6 & 1/12 & 1/12 \\
\end{array}
\times
\begin{array}{c|ccc}
Rock & 1 & 0 & 2 \\
Paper & 2 & 1 & 0 \\
Scissors & 0 & 1 & 2 \\
\end{array}

Figure 3.5: \( u(s,s) = u(t,s) = u(t,t) = u(s,t) = 1 \)
We see that even though $u(t,s) = u(s,s)$, it’s false that $u(s,t) > u(t,t)$.

On the other hand, we can imagine engaging in the following line of reasoning: suppose I notice my opponent is playing rock half the time and paper and scissors each one quarter of the time. I reason that I ought to play paper half the time (and scissors and rock a quarter each.) I do so. But then if my opponent performs similar reasoning in the next round, then he will play scissors half the time and rock and paper a quarter each. In the round after that, I will play rock half the time and paper and scissors a quarter each. Thus, we will cycle endlessly. There seems to be some kind of stability here, where each player is successively playing best responses to the other’s last move. Hofbauer and Sigmund (1998) thus give the following definition for 2-player games where both strategy sets are the same: a strategy $s$ is evolutionarily stable under the best response dynamics iff for all $t \neq s$:

1. $u(s,s) > u(t,s)$, or

2. $u(s,s) = u(t,s)$ and there is some $t' \neq s$ such that $u(t',s) = u(s,s)$ and $u(t',t) > u(t,t)$

All this talk of best respose dynamics might lead us to think we can explain vagueness in terms of its being a best response to some other strategy. However, Jäger (2007a) proves that for asymmetric games, a strategy is evolutionarily stable under the best response dynamics if and only if it is a strict Nash equilibrium. Thus again, if vagueness is not part of the strict Nash equilibrium, it is not predicted to arise on the best response dynamic!
3.3.3 Two Solutions

In this section I present two attempts to explain the origin of vagueness; importantly, both rely on an assumption of what is called bounded rationality: in contrast to the assumptions of classical game theory, players are not assumed to perfectly know the structure of the game and its payoffs and be able to reason to infinite strategic depth.

3.3.3.1 The Shakes

A second attempt to explain the emergence of vagueness in signaling games is due to Franke et al. (2010). They also assume a similarity relation imposed on the states observed by the sender and compute the payoff as a function of the similarity between encoded and decoding states; instead of defining similarity as the absolute difference between the encoded and decoding states and then specifying a payoff function, they specify a Gaussian similarity relation, and then identify the payoff with the similarity:

$$U_{S,R}(t, t') = \text{sim}(t, t') = \exp \left( \frac{-(t - t')^2}{2\sigma^2} \right), \quad (3.8)$$

where $t$ and $t'$ are the encoded and decoding states, respectively. (They call these sim-max games.) Furthermore, they assume that sender and receiver do not deterministically choose payoff maximal strategies; instead, sender and receiver use what is called a logit probabilistic choice rule:
\[
P_S(t, m) = \frac{\exp(\lambda \times EU_S(t, m))}{\sum_{t'} \exp(\lambda \times EU_S(t, m'))}
\]

(3.9)

\[
P_R(m, t) = \frac{\exp(\lambda \times EU_R(m, t))}{\sum_{t'} \exp(\lambda \times EU_R(m, t'))}
\]

(3.10)

where in the case at hand the expected utility functions \(EU\) are given by:

\[
EU_S(t, m) = \sum_{t'} P_R(m, t') \times U_{S,R}(t, t')
\]

(3.11)

\[
EU_R(m, t) = \sum_{t'} P_S(t', m) \times U_{S,R}(t, t').
\]

(3.12)

As for \(\lambda\), it is a rational choice constant: as \(\lambda\) approaches 0, \(P_S(t, t^*)\) approaches uniform random choice of strategy; as \(\lambda\) approaches \(\infty\), \(P_S(t, t^*)\) approaches deterministic choice of expected payoff-maximal strategy. It is important to note the adoption of a probabilistic choice rule is Franke, et al’s departure from classical game theory; this in turn allows them to offer an explanation of vagueness in signaling games. That is, this is their adoption of a version of bounded rationality. If we assume sender and receiver use the same \(\lambda\) and are correct in their assessment of the others’ encoding and decoding probabilities, (an assumption reflected in fact that \(EU_S\) depends on \(P_R\), and \(EU_R\) depends on \(P_S\)) then there will always exist a quantal response equilibria:
Franke, et al simulate a 100 state, 3 signal sim-max game with $\sigma = .2$, $\lambda = 20$ and demonstrate that it has a quantal response equilibria on which $P_S(t, m_1)$ is near 1 for almost all of the lower third of states and smoothly transitions to 0 for the rest of the states, $P_S(t, m_2)$ is near 1 for almost all of the middle third of states and smoothly transitions down to 0 over the rest of the states, and $P_S(t, m_3)$ is near 1 for almost all of the upper third of states and smoothly transitions down to 0 for the rest of the states. $P_R(m_1, t)$, $P_R(m_2, t)$, and $P_R(m_3, t)$ are a roughly Gaussian distributions centered at the middle of the lower, middle, and upper third of states, respectively.

However, I find this model of the origins of vagueness unsatisfying, as it offers no explanation as to exactly why senders and receivers fail to deterministically choose payoff maximizing strategies. The equations describe probabilistic encoding and decoding behavior, and given sufficient data we might be able to specify values for $\lambda$ that accurately describe observed behavior; however, that is different than explaining just why $\lambda$ takes the values that it does.

Now, the authors do say that just as a single person at different times may make different choices under conditions that are indistinguishable relative to a given behavioral model, due to factors that fail to be captured by that model, so also such factors may sum over different
strategies such that even though one strategy is payoff maximizing according to the model and another is not payoff maximizing according to the model, the player at times treats the latter as if it were one of the former. This may be due in part to a player’s imperfect understanding of what is payoff maximizing, and it may be due in part to the model failing to adequately capture the player’s preferences. As large sums of small error factors are less likely than small sums of small error factors, we may even assume that the farther a strategy is from being payoff maximizing according to the model, the less likely a player is to mistake it from one that is payoff maximizing according to the model; this is the effect of \( \lambda \) in their model. However, to say this is again not to say what those error factors are, and it makes no predictions of when vagueness can be expected to arise.

### 3.3.3.2 Generalized Reinforcement

Another attempt to explain the origin of vagueness is due to O’Connor (2014). She presents a modified signaling game in which there is a similarity relation imposed on the states observed by the sender. The rationale for doing so is that vague predicates often apply to varying degrees to an object: a 50 kg bear is not as big as a 500 kg bear. Furthermore, in view of the similarity relation over states, the payoffs obtained by the sender and receiver are allowed to vary with the distance between the state encoded by the sender in the chosen message and the state that message is decoded as by the receiver. (To simplify matters, we assume in the remainder of this paper that the receiver’s actions are all to choose among the states observed by the sender, unless otherwise noted.) Again, this makes sense: the bigger the bear, the faster we ought to run from it. Thus, in a 20-state game with 2 messages, if the
sender observes state 5 and encodes it via message 0, the sender and receiver would obtain (the same) maximal payoff if message 0 is decoded as state 5, a smaller payoff if decoded as state 4 or 6, a yet smaller payoff if decoded as state 3 or 7, and so on. She calls this a contiguous signaling (CS) game.

To see the effect this has on the equilibria in signaling games, let us return to the game in Figure (3.1). In such a signaling game what Lewis called a signaling system is possible: a strategy profile such that there is a one-one sender function from states to signals and a one-one receiver function from signals to states which is also the inverse of the sender function; signaling systems thus effect perfect coordination. With more than 2 states, senders can also engage in partial pooling, where a signal can encode more than one state. When states and signals are equi-numerous, partial pooling is payoff dominated by signaling systems; but where states outnumber signals, players can do no better than partial pooling to achieve partial coordination. As Jäger (2007b) proves, the effect of imposing a similarity metric on states and then varying the payoffs with the distance between the encoded state and the decoding state is to reduce the number of partial pooling Nash equilibria. To see why, consider Figure 3.6a-3.6d.

Here Figure 3.6a is a pure strategy partial pooling equilibria of a 7-state, 2-signal CS game; Figures 3.6b-3.6d are pure strategy profiles that are partial pooling equilibria of the corresponding non-CS game (where payoffs do not vary with distance of the decoding state from the encoded state.), but which are not pure strategy partial pooling equilibria of the corresponding CS game. In Figure 3.6b the receiver decodes signal 2 as a state too far from the median of the states encoded by signal 2. In Figure 3.6c the sender would achieve
higher payoffs by encoding state 7 as signal 2, since state 6 is closer to state 7 than state 2. In Figure 3.6d the sender would achieve higher payoffs by encoding state 3 as signal 1, since state 2 is closer to state 3 than state 6. Thus for CS games, the pure strategy partial pooling equilibria require that the receiver decodes each signal as a state among the median states of that signals’ encoded states, and that the sender encodes states in convex-shaped, equal-sized groups.

However, this is still not sufficient to explain vagueness as a strict Nash partial pooling equilibrium of a CS game: the results of Lipman (2009) still apply, since the average distance between encoded states and decoding states is minimized by playing pure strategies: suppose the sender probabilistically encoded states 3, 4, and 5 as signals 1 and 2 at the rates of .75/.25, .5/.5, and .25/.75, respectively. Consider the 25% of the time that state 5 is encoded as signal 1: since state 2 is farther from state 5 than state 6, this results in a lower overall payoff. Similarly for the 25% of the time that state 3 is encoded as signal 2.4

Instead, O’Connor explains vagueness in signaling games as the result of a modified version of Herrnstein reinforcement learning, inspired by the phenomenon of stimulus generalization; she calls this generalized reinforcement learning. This learning method permits players to develop signaling arrangements that can maximize payoffs when states are numerous, signals are few, and the players must quickly develop such arrangements; however, it also results in persistently (and essentially) vague signaling.5

Under (non-generalized) Herrnstein reinforcement learning, due to Roth and Erev (1995),

4What if only state 4 was probabilistically encoded? Even here the mixed strategy does no better than 3.6a since the decoding state (2 or 4) is still always 2 states away. In general, games where the states cannot be evenly divided among the signals allow edge states to be split between two signals without loss—or
players reinforce choice of strategy in proportion to that strategy’s past payoff. To see how this work for a 7-state, 2-signal game, we imagine the sender with 7 urns, one for each state, in front of him. Each sender urn starts with one signal 0 ball, and one signal 1 ball. We imagine the receiver with two urns, one for each signal, in front of him. Each receiver urn starts with one state 0 ball, one state 1 ball, and so on up to state 6. (See Round 0 in Figure 3.7.) In each round of play of the game, the sender and receiver pick encoding signals and decoding states, respectively, by pulling balls from their urns. For example, if the sender observes state 2 in round 1, he picks a ball from urn 2; perhaps it is a signal 1 ball. The sender thus sends signal 1, and the receiver, having received signal 1, reaches into his signal 1 urn and picks a ball; perhaps it is a state 5 ball. Since states 2 and 5 are a medium distance apart, the players earn a moderate payoff; having earned this moderate payoff, the sender places additional signal 1 balls in his state 2 urn, and the receiver places additional state 5 balls in his signal 1 urn. (In fact, we can imagine the payoff as just being the balls themselves.) How many balls do they add? For this we require a *payoff function* that specifies the payoff as a function of the distance of the decoding state from the encoding state. In our own simulations we have found that the choice of function does not much alter the kind of resulting signaling arrangements that result, so long as the payoffs monotonically decrease with increasing distance. If we suppose that this payoff function specifies a payoff of 4 balls for a 3 state difference, then the sender and receiver urns will be as illustrated in Figure 3.7, Round 1. Then, in Round 2, if the sender observes state 2 again she will be likely gain—of payoff.

5Note that we decline to call these arrangements *signaling systems*, since they are not, except in the limiting case, representable as one-one sender state-to-signal functions and inverse receiver sender-to-state functions.
to pick signal 1, and the receiver will be more likely to pick state 5 upon receiving signal 1; the sender encoding probabilities are calculated by normalizing weights across signals:

$$P_S(t, m) = \frac{W_S(t, m)}{\sum_{m'} W_S(t, m')} \quad (3.15)$$

and likewise for the receiver decoding probabilities, which are calculated by normalizing weights across states:

$$P_R(m, t) = \frac{W_R(m, t)}{\sum_{t'} W_R(m, t')} \quad (3.16)$$

where $W_S$ and $W_R$ are the sender and receiver weight matrices (or arrays of urns), respectively. Note that the sender only adds signal 1 balls only to the state 2 urn, and the receiver only adds state 5 balls to the message 1 urn. Since the sender ball type will always match the receiver urn type, we can describe this as the sender and receiver reinforcing on the encoded state 2 and the decoding state 5, respectively.

What about generalized reinforcement learning? Here we imagine that the sender and receiver reinforce not just on the encoded state and decoding state, respectively, but also on nearby states. The rationale here is that a given situation tends to evince a given reaction in proportion to the past success of that reaction under similar situations: if running very quickly from a 500 kg bear was successful in the past, one will tend to run—albeit not as quickly—from a 250 kg bear too, and even less so from a 50 kg bear. For our games, this means the level of sender reinforcement monotonically decreases across states with distance
from the *encoded* state; the level of receiver reinforcement monotonically decreases across states with distance from the *decoding* state. Thus, in addition to a payoff function that specifies the payoff for the encoded state and decoding state, we also require a *generalization function* that specifies how quickly the payoff drops off across states; if we suppose that this generalization function specifies a 1-1 state-ball drop off, then after Rounds 0 and 1 our urns would be as in Figure 3.8, (assuming the payoff function remains unchanged from Figure 3.7). As before, the sender encoding and receiver decoding probabilities are calculated by normalizing the weights across signals and states, respectively.

The more widely we generalize our reinforcement across states, the more vague the resulting signaling arrangement will tend to be, for a fixed number of rounds of play, where vagueness is measured as the average, across states, of the probability difference between the first and second most probable signal encoding that state. (In fact, O’Connor reports that in the absence of generalization, reinforcement learning dynamics in signaling games approaches non-vague signaling arrangements as the number of rounds of play approaches infinity.)

We can define a measure of success by summing over all rounds of play the expected payoff of the sender and receiver strategies resulting from that round of play, and divide that total by the expected payoff of playing a payoff dominant strategy for all rounds of play—which for our games will be pure strategies in which the sender encodes states as equal-sized, convex regions, and in which the receiver maps signals to the median state of the states encoded by that signal. For a 200 state, 40 signal game, O’Connor reports that in shorter simulations of 100,000 rounds, generalizing more widely results in higher levels of success. However, at simulations run to 1,000,000 rounds, generalizing more widely results in almost no change

141
in success; at 10,000,000 rounds generalizing more widely results in lower success. This is because wider generalization allows players to more quickly reach signaling arrangements, but the resulting arrangements will have progressively lower expected payoffs the more widely the players generalize. In shorter simulations this is an advantage; in longer ones, a disadvantage. This effect is reduced as we shrink the number of states but maintain the state/signal ratio; in a 50 state, 10 signal game, wider generalization almost monotonically leads to lower success no matter the length of simulation. In contrast, this effect is increased as the number of signals decreases; for a 200 state, 10 signal game run to 1,000,000 rounds, wider generalization leads to progressively higher levels of success.

It is important to note that this explanation of vagueness as a result of a learning process that achieves higher shorter-term payoffs depends on a kind of bounded rationality and thus a departure from classical game theory: players are not assumed to adopt a strategy that achieves the maximal payoff or even one that maximizes expected payoff; instead, they play a strategy necessitated by the need to learn quickly the tendencies of the other player. Given infinite time, they would be better off by playing deterministic, non-vague strategies.

3.4 Symmetrization and Normal Distributions

Here we extend O’Connor’s model in two ways: first, we allow players to arbitrarily take on the role of either sender or receiver; second, we investigate the behavior that emerges when the states of nature are roughly normally distributed, as opposed to uniformly distributed. The former change more completely reflects the conditions under which language evolves;
the latter change reflects the intuition that vague predicates like ‘tall’ are often used with comparison classes that are normally distributed with respect to the feature in question: for example, we say of people that they are ‘tall’ or ‘short.

Thus, it will be helpful to review the results of our own implementation of that model. Consider, for example, a 20 state, 2 signal, 20 action game, with a Gaussian payoff function

\[
U^*_{S,R}(t, t') = \alpha \times \exp \left( \frac{-(t - t')^2}{2\sigma^2} \right) \tag{3.17}
\]

where \( t \) and \( t' \) are the encoded and decoding states, respectively, and where \( \alpha = 5 \) and \( \sigma \approx 2.548 \). Since this is the payoff function, \( U^*_{S,R} \) is the amount of reinforcement (number of balls) that the sender and receiver place on the encoded and decoding states, respectively. If \( t = t' \) then \( U^*_{S,R} = 5 \); as \( t - t' \) increases, \( U^*_{S,R} \) decreases. With \( \sigma \approx 2.548 \), the full width at half maximum of our payoff function is 6; that is, \( U^*_{S,R} = 2.5 \) for \( t - t' = 3 \). But sender and receiver also reinforce on states nearby the encoded and decoding states, respectively, according to the generalization function. Suppose we assume a Gaussian generalization function

\[
U_S(t'') = U^*_{S,R}(t, t') \times \exp \left( \frac{-(t - t'')^2}{2\sigma^2} \right) \tag{3.18}
\]

and

\[
U_R(t'') = U^*_{S,R}(t, t') \times \exp \left( \frac{-(t' - t'')^2}{2\sigma^2} \right) \tag{3.19}
\]

\[\text{For a Gaussian function, full width at half maximum} = 2\sqrt{2\ln 2} \sigma\]
with again full width at half maximum = 6; again \( t \) and \( t' \) are the encoded and decoding states, so that states farther from the former and latter get progressively less and less reinforcement from the sender and receiver, respectively.\(^7\)

The sender and receiver’s state-to-signal and signal-to-state probabilities for a typical sample of 10,000 rounds of play is as in Figure 3.9, where the blue line represents one signal and the green line represents the other. Note that the sender’s state to signal probabilities for each state must sum to 1 across all signals, and the receiver’s signal to state probabilities for each signal must sum to 1 across all states. To get an idea of the distribution of outcomes, we see the sender and receiver’s state-to-signal and signal-to-state probabilities for 30 samples of 10,000 rounds of play in Figure 3.10. Note that although across samples the signals tend to group into those that encode the lower states and those that encode the upper states, they are different signals from sample to sample; whether it is signal 0 or signal 1 in a given sample that comes to encode and be decoded as the lower (or higher) states is determined by the events that transpire as the game progresses. For comparison, we see the results from a 20 state, 2 signal, 20 action game with a linear payoff function

\[
U_{S,R}^*(t, t') = \begin{cases} 
\alpha |t - t'| + \beta & \text{if } |t - t'| \leq 5 \\
0 & \text{else}
\end{cases}
\]  

(3.20)

where \( \alpha = -5/6 \) and \( \beta = 5 \) (so that \( U_{S,R}^* = 2.5 \) for \( t - t' = 3 \)) and a Gaussian generalization function with full width at half height = 6 in Figure 3.11.

\(^7\)This is of course two functions, and leaves unexplored the possibility of specifying different \( \sigma \) values for the sender and receiver.
3.4.1 Symmetrization

We begin to depart from the existing model by noticing an intuitive shortcoming of the generalized reinforcement dynamics of signaling games studied thus far: in real life, the receiver often becomes the sender and vice versa. Language evolution is not a one-way street from senders to receivers. Furthermore, the assumption that the actions are identical to the states in fact makes possible an exchange of roles: the sender’s state-many by signal-many weight matrix can be transposed into a signal-many by state-many weight matrix that a receiver would use, and the receiver’s signal-many by state-many weight matrix can be transposed into state-many by signal-many weight matrix that a sender would use. So, we can simulate the evolution of games where players take on the sender or receiver role at random by having a single state-many by signal-many weight matrix for each player. If in a certain round a player is picked as the sender, their weight matrix is normalized across signals as before to derive their encoding probabilities; if they are picked as the receiver, their weight matrix is transposed and normalized across states to derive their decoding probabilities. This is the first modification we make to the generalized reinforcement model.

Note that it is not merely a matter of mathematical expediency that permits us to allow players to sometimes be the sender and other times the receiver, making use of the same (sometimes transposed) table of weights no matter which role they take on: if we have to keep separate matrices of state-to-signal and signal-to-state weights for each player, they will in essence be playing two separate games, with no more than random probability that they will have similar encoding and decoding probabilities for a given signal. That would
be as if our model attempted to explain how words evolve between you and I, but ended up predicting that it is only a matter of chance that you use ‘tall’ they way I use ‘tall’. Similar to how ‘hole’ in English means what’ ho:l’ means in Yucatec Mayan, the players would be speaking in two separate languages, like a parent who speaks to a child in a mother tongue and a child who speaks back in an adopted second language.

For a 20 state, 2 signal, 20 action, 2 player game with our symmetric generalized reinforcement learning, still assuming a Gaussian payoff function of height 5 and full width at half maximum = 6, and a Gaussian generalization function with full width at half maximum = 6, the players’ encoding and decoding probabilities after 10,000 rounds of play are as in Figure 3.12, with sample size = 30. After 100,000 rounds of play we have Figure 3.13.

Note also that we can take the sender and receiver’s state-to-signal and signal-to-state weights, respectively, resulting from non-symmetrized generalized reinforcement learning and transpose them to see how such a sender would receive, and how such a receiver would send. We see the results in Figure 3.14.

The main difference we observe here is that the sender resulting from a one-way game, if asked to act as a receiver, would on average have a ‘flatter’ signal to state decoding probability distribution than the receiver resulting from the same one-way game; and the receiver resulting from a one-way game, if asked to act as a sender, would on average have a ‘steeper’ state to signal probability function than the sender resulting from the same one-way game. If we designate one signal in our two-signal game as the ‘low-state’ signal and the other as the ‘high-state’ signal based on the average of the players’ peak state to signal probabilities for each signal $m$: 

146
\[ \sum_{i \in n} \arg \max_i P_i(t, m) \]

then for a given sample we can calculate the difference between the encoding behaviors of the receiver and sender resulting from that sample, and the difference between the decoding behaviors of the sender and receiver resulting from that signal. If we do so for the same 30 samples that we saw in Figure 3.10 then we have Figure 3.15a. If we do the same for the 30 samples of the symmetrized game we saw in Figure 3.12 then we have Figure 3.15b, and likewise for Figure 3.13 in Figure 3.15c. In each sub-figure the ‘low-state’ signal is in red, and the ‘high-state’ signal is indicated in blue; the steeper receiver-cum-sender curves from non-symmetrized games are apparent as the red and blue bulges above and below 0 in the middle of the top graph of Figure 3.15a, and the flatter sender-cum-receiver curves from non-symmetrized games are apparent as the red and blue curves above and below 0 in the bottom graph of Figure 3.15a. We also see in Figures 3.15b and 3.15c that in symmetrized games as the number of rounds increases, the player’s encoding and decoding probabilities tend to converge.

Allowing players to take on the sender or receiver role at random also allows the us to have more than 2 players in the game; the motivation here is that language evolves groups with more than 2 members. If we continue to assume that any iteration of the game involves only a single sender and a single receiver, the convergence of state-to-signal and signal-to-state probabilities across samples takes more iterations than in a 2 player game; this is unsurprising: with fewer turns for any given player, it takes more time to the probabilities to converge.
We might also allow for one-to-many sender-receiver relations, or many-to-one sender-receiver relations, or many-to-many sender-receiver relations. The first seems best motivated of the three modifications, since it seems common that one speaker is heard by many listeners; hence it is the only one of the three that we have simulated. To do so we calculate the reinforcement level by using a function, perhaps Gaussian, linear, or otherwise, on the standard deviation from the state of nature. The sender still reinforces on the state of nature, and the receivers each reinforce on their individual decoding states. The intuition at work here is that sender and receivers would reinforce more strongly in cases where all receivers have a similar decoding state that is close to the state of nature than in either cases where the receivers’ decoding states are on average close to the state of nature but still widely dispersed, or cases where the receivers’ decoding states are closely clustered together but still far from the state of nature. We want both closeness and concentration. (The function from standard deviation to reinforcement level takes the place of the reinforcement function from the one-to-one sender-receiver model.) We also require a function to specify the distribution of the number of receivers. We use a gamma distribution, with the user supplying a shape parameter and a scale parameter. Our findings are as expected, with the time (measured in iterations of the game) required for convergence greater than for 2-player symmetrized games, but less than for multi-player symmetrized games where only one-one sender-receiver relations are allowed.

For the second, we propose to calculate the signal by assigning signals to upper-open intervals (that is, signal 0 is $[0,1)$, signal 1 is $[1, 2)$, and so on. Given many signals, we reinforce on the signal into which the average of the signals falls. We then account for the distribution
of signals adjusting the level of reinforcement by using a function, perhaps Gaussian, linear, or otherwise, on the standard deviation of the signals from the average signal. For example, if the average signal is .3333333333 then we reinforce on signal 0, adjusting the level of reinforcement downwards if the standard deviation is large. The intuition at work is that if a bunch of people tell you ‘blorp’ and you pick state 5 and it matches the state of nature, you would reinforce more than if among the same number of people some tell you ‘blorp’, others tell you ‘blorn’, and still others tell you ‘blurp’. We want concentration, but there is no closeness at stake. Admittedly, our procedure seems to assume there will always be some metric of similarity between blorp, blorn, and blurp, which is conveniently provided here by the number of different letters among the signals. (I do not currently have a better solution.) We also require a function to specify the distribution of the number of senders.

For the third, we propose to simply combine the first and the second. This necessitates the specification of three more functions than a simple one-to-one sender-receiver model: two to determine the distribution of the numbers of senders and receivers, and one to adjust the level of reinforcement depending on the Euclidean distance or standard deviation of the signals from the average signal. Again, however, we leave these second and third modifications unexplored, and point them out only as justification for allowing players to take on the sender or receiver role at random (and for doing so by allowing the state-many by signal-many weight matrix to be transposed depending on the role a player takes in a given iteration of the game.)
3.4.2 Normal State Distributions

The second modification we make is to allow non-uniform distributions of states. The intuition here is that the property of members of a comparison class of interest that a vague predicate gives information about is often not uniformly distributed: adult human heights are not uniformly distributed, neither are the masses of stars or the lengths of ships. For a 20 state/20 action game we approximate a normal distribution with a truncated normal distribution centered at 10 and generate random variates of that distribution; a variate $v$ is assigned to one of our discrete states $t$ just in case $t < v < t + 1$.

For a 20 state, 2 signal, 20 action game, 2 player game with a truncated normal state distribution with $\sigma = 3$ and $\sigma = 6$ for 10,000 rounds of symmetrized generalized reinforcement learning we see the results of 30 samples in figures 3.16a and 3.16b, respectively. We see that as $\sigma$ decreases it is increasingly difficult to generate the kind of signaling arrangements that intuitively reflect vagueness: instead, the sender tends to use one signal to the exclusion of the other, and the receiver interprets both signals very similarly. Increasing the number of rounds in figure 3.16c does mitigate this effect.

One solution is to increase the payoffs for extreme states of nature; we might justify this choice on the grounds that ‘getting it right’ in the extreme cases can be very important. The difference between something 100dB and 110db is less significant than the difference between 110dB and 120dB, just as the difference between hitting another car at 25 versus 35 mph is less significant than the difference between hitting another car at 35 mph versus 45 mph. Of course, both of these examples involve essentially exponential scales: decibels are
logarithmic scale, and kinetic energy is a squared function of velocity. Here we increase the payoff of a round by the ratio of the probability of the most common state to the probability of the state of nature: if the state of nature in a given round is 20 times less probable than the most common state, the payoff for that round is increased by a factor of 20. (Thus there is no increased payoff for the central states.) For the same game settings as in figure 3.16a but using this extremal value compensation, we have the results in figure 3.17.
Figure 3.6
Figure 3.7: 7-state, 2-signal game after Rounds 0 and 1

Figure 3.8: 7-state, 2-signal game after Rounds 0 and 1
Figure 3.9: 10000 Rounds Non-symmetrized 1 Sample

Figure 3.10: 10000 Non-symmetrized 30 Samples

Figure 3.11: 10000 Rounds Non-symmetrized 30 Samples Linear Payoff

Figure 3.12: 10000 Rounds Symmetrized 30 Samples
Figure 3.13: 100000 Rounds Symmetrized 30 Samples

Figure 3.14: 10000 Rounds Non-symmetrized 30 Samples Roles Reversed
Figure 3.15

(a) 10000 Rounds Non-symmetrized 30 Samples Roles Reversed Comparison

(b) 10000 Rounds Symmetrized 30 Samples Roles Reversed Comparison

(c) 100000 Rounds Symmetrized 30 Samples Roles Reversed Comparison
(a) 10000 Rounds Symmetrized 30 Samples
Normal Distribution $\sigma = 3$

(b) 100000 Rounds Symmetrized 30 Samples
Normal Distribution $\sigma = 6$

(c) 1000000 Rounds Symmetrized 30 Samples
Normal Distribution $\sigma = 6$

Figure 3.16

Figure 3.17: 10000 Rounds Symmetrized 30 Samples Normal Distribution $\sigma = 3$ Extremal Compensation
Chapter 4

A Dynamic Probabilistic View

4.1 Introduction

I begin this chapter by showing in §4.2 the equivalence of a number of operations that have featured prominently in recent semantics and pragmatics; I then argue that these operations have a natural generalization to an equation, identical in form if not substance to one seen in Chapter 2, that generates statistical inference patterns from utterances of sentences containing vague predicates. I then argue that conformation to this equation is a requirement for any plausible Bayesian model of how, given utterances containing vague predicates, listeners draw the pre-theoretically expected statistical inference patterns from beliefs concerning the truth-conditions of such utterances. Then, in §4.3 I present three arguments against any such model: the first argument claims that under normal conditions, (where what Barker (2002) has termed ‘metalinguistic modes of use’ are not in play), utterances of sentences containing vague predicates have interpretations under which the thresholds, relative to which the truth-conditions of utterances of such sentences must be defined, turn out to be otiose to the model. The second argument is disjunctive; it claims that any plausible Bayesian model
of truth-condition driven inference must fall into one of two categories, both of which suffer fatal defects. Finally, the third argument claims that any such model, if it gives rise to plausible inferences from sentences such as ‘Feynman is tall’, must also give rise to implausible inferences from sentences such as ‘Feynman is not tall’.

I then present in §4.4 an alternative model of the manner in which utterances of sentences with vague predicates give rise to the expected probabilistic listener inference patterns, of both the metalinguistic and non-metalinguistic kind. On this model, in non-metalinguistic uses of vague predicates the truth-conditions of utterances of sentences containing vague predicates have been, at bottom, replaced by probabilistic encoding and decoding behavior. Bayesian calculations based on anything recognizable as the truth-conditions of such sentences, are used only to generate metalinguistic inference patterns, with a probability distribution for thresholds defined by taking the derivative of the sender’s encoding probabilities. Knowledge of the truth-conditions of an utterance containing a vague predicate is only reflected as knowledge of the encoding probabilities, conditional on first, the subject of the utterance’s degree of deviation, with respect to the property expressed by that vague predicate, from some measure of central tendency, and second, the threshold for application of that vague predicate, again expressed as some requisite degree of deviation, with respect to the property expressed by that vague predicate. The upshot is that knowledge of the truth-conditions of an utterance of a sentence containing a vague predicate does not play any role in generating the non-metalinguistic statistical inference patterns that we expect from utterances of sentences with vague predicates. Finally, on the assumption that it is knowledge of the meaning of vague predicates that allows us to engage in the the kind of
statistical inferences we typically engage in when we hear utterances of sentences containing vague predicates, I propose a probabilistic semantics for vague predicates, and demonstrate how we can integrate such a view with an existing degree-theoretic, truth-conditional semantics.

4.2 Showing Equivalence

4.2.1 From Barker to Stalnaker/Kaplan

One important body of semantic theory that has developed in the past thirty-odd years is that of dynamic semantics, on which the meaning of a given sentence is determined by its potential to change the conversational context in which it is uttered; one component of this is its potential to eliminate from consideration possible worlds that have not, prior to the utterance of said sentence, been ruled out by the common knowledge of the discourse participants. We say that $P$ is commonly known among the discourse participants just in case for every discourse participant $x$, $x$ knows that $P$, knows, for any discourse participant $y$, that $y$ knows that $P$, knows, for any discourse participant $y$, that $y$ knows, for any discourse participant $z$, that $z$ knows that $P$, and so on. The collection of all $P$ that are commonly known among the discourse participants is the common knowledge of the discourse participants; if we represent each such $P$ as a set of possible worlds, we can represent the common knowledge as the intersection of all such $P$, and this set of possible worlds is called the common ground of the discourse participants.
To see how a sentence token can change the conversational context in which it is uttered by eliminating from consideration possible worlds that have not, prior to the utterance of said sentence, been ruled out of the common ground of the discourse participants, consider that perhaps it is common knowledge among parties to the going conversation that Feynman is taller than 6'0" tall, but just exactly how tall he is, is not common knowledge. When I say to the discourse participants, ‘Feynman is 6'3" tall’, I am eliminating from the common ground of the discourse participants, all worlds in which Feynman is less than or greater than 6'3" tall. Notice that in this example, it seems that the only information that has in some rough sense been *directly* added to the common ground is information about Feynman’s height, in the sense of ‘direct’ such that if the common ground prior to my utterance were to contain only worlds where the threshold for tallness was greater than 6'3", then even though the discourse participants would learn something about ‘tall’ by my utterance—they would learn that ‘tall’ does not apply to Feynman—this is only learned indirectly: if the common ground prior to my utterance contained both worlds where the threshold for tallness is above 6'3" and also worlds where the threshold for tallness is below 6'3", then the common ground after my utterance will contain both worlds where ‘tall’ applies to Feynman, and worlds where ‘tall’ does not apply to Feynman.

On the other hand, there do seem to be cases where the information added to the common ground by an utterance of a sentence is directly about the words in the language of the sentence, even the words that are used, and not merely mentioned, in the sentence. For example, perhaps you and I are looking through the glass at the unique person whom ‘Batman’ is used to refer to, in the interrogation room with Harvey Dent. You and I both know that
that person in the black cape is the unique person whom ‘Batman’ is used to refer to, that
the other knows that that person in the black cape is the unique person whom ‘Batman’ is
used to refer to, that the other knows that the other knows . . . and so on. Thus, all worlds in
the common ground are worlds in which that person in the black cape is the unique person
whom ‘Batman’ is used to refer to. If you don’t know that ‘Bruce Wayne’ is used to refer to
the same person as ‘Batman’ is used to refer to, however, the common ground includes worlds
in which ‘Bruce Wayne’ is used to refer to someone other than the person in the black cape
behind the glass, and worlds in which ‘Bruce Wayne’ is in fact used to refer to the person
in the black cape behind the glass. I might say ‘That is Bruce Wayne,’ and thus eliminate
from consideration all possible worlds in the common ground in which ‘Bruce Wayne’ is not
used to refer to the same person as ‘Batman’ is used to refer to.¹ Having uttered this, I have
changed the context of future utterances in our discourse into one in which such possible
worlds have been eliminated from the common ground.

How can we extend this dynamic analysis from sentences to vague predicates? A key insight
from Barker (2002) is that a usage of a given vague predicate can serve not merely to add
information to the conversational common ground about the degree to which an object has
the property measured by that predicate, but also to give information about the degree

¹If I didn’t know that you knew that that person in the black cape is the unique person whom ‘Batman’
is used to refer to, I couldn’t utter ‘That is Bruce Wayne,’ and successfully inform you that ‘Bruce Wayne’
is used to refer to the same person as ‘Batman’ is used to refer to. Furthermore, if I didn’t know that you
knew that that person in the black cape is the unique person whom ‘Batman’ is used to refer to, I might in
fact feel the need to tell you ‘That is Batman,’ which I in fact do not do. Thus, the common ground helps
account for both what is said, and what is not said.
of possession of that property that is required, in the context of utterance, to warrant usage of that vague predicate. He calls the former and latter kinds of usage descriptive and metalinguistic usages, respectively. To use his example: I might utter ‘Feynman is tall’, and thereby add to the common ground information about Feynman’s actual height. If it is part of the conversational common ground that the degree of height required for application of ‘tall’, in the context of our conversation, is 180 cm, then I will have added to the common ground the information (all other things being equal—no protestations to the contrary from you, for example) that Feynman’s height is greater than 180 cm. This would be an example of a purely descriptive usage. On the other hand, it might be part of the common ground that Feynman’s height is 180 cm, and thus it might be perfectly well known to all participants in the conversation that that is his height. You might wonder, however, just how tall someone has to be around these parts to count as tall—and thus I might utter, ‘Well, certainly, Feynman is tall’, and thus add to the common ground the information that the degree of height required for tallness in our conversational context is certainly less than or equal to 180 cm. This would be an example of a purely metalinguistic usage of a predicate.²

Of course, it seems obvious that in most situations the common ground does not precisely specify the degree of height required for the application of a vague predicate (if there even is a precise degree of height so required), and also in most situations the common ground does not precisely specify the height of all entities in the domain of discourse (if there even is a precise degree of height—just how tall am I, given that my height varies from morning to evening?) Thus, Barker claims we can expect that in most contexts, usage of a vague predicate...
predicate will not be either purely descriptive or purely metalinguistic; instead, such usages will add both descriptive and metalinguistic information to the common ground.

How can we implement this in a formal semantics? In general, formal semantics analyzes semantic composition as functional application: composing two expressions together to form new meaningful expressions is analyzed as applying the meaning of one of the expressions, a function, to the meaning of the other expression, an argument, to yield the meaning of the compound expression. In dynamic semantics, where we typically take sentences to denote functions from sets of worlds to sets of worlds, it seems we ought to take a unary predicate (which semantically composes with a noun phrase to form a sentence) to denote a function from entities (which are taken to be the meanings of noun phrases) to a function from sets of worlds to sets of worlds. Thus, just as a unary predicate composed with a noun phrase yields a sentence, so also the denotation of a unary predicate, (a function from entities to a function from sets of worlds to sets of worlds), when applied to the denotation of a noun phrase (an entity), yields the denotation of a sentence (a function from sets of worlds to sets of worlds).

Now, as a feature of the context of utterance we might relativize the denotations of tokens of vague unary predicate types like ‘tall’ and ‘long’ to the degree of height or length required to count as tall or long in a given context, analogous to how we relativize the denotations of tokens of deictic pronoun types like ‘I’ and ‘you’ to the parameters of a given context. As R. Stalnaker (1998) argues, if we think of an assertion of a sentence in a particular discourse as changing the context, and also think of an assertion of a sentence as interpreted only relative to such features of context as the degrees of height, heat, etc., required for
application of vague predicates, and we further think that such features of context can be changed by prior utterances in a discourse, then we see that a given context is both changed by the assertion of prior sentences, and goes on to determine the interpretation of subsequent sentences. In discourses where information is exchanged, the change in context effected by the assertion of prior sentences can be represented by changes in sets of possible worlds, as the primary effect of assertions of sentences in such discourses is to change the information presupposed to be common knowledge among participants to the discourse. Since this continually updated information is then the source for such features of context as the degrees of height, heat, etc., required for application of vague predicates, which are necessary for subsequent interpretation of sentences, we can think of this information as the context relative to which interpretation occurs, and represent it too as a set of possible worlds.

Some further preliminaries: First, we define a function \( d \): given a world and a predicate denotation, \( d \) tells you to what degree one must possess the property measured by that predicate in order to qualify for the application of that predicate in that world: it tells you, for example, just how tall you have to be to count as tall in any given world. Second, let \( \text{tall} \) be a function that, given a degree and an entity, delivers the set of worlds in which that entity possesses at least that degree of tallness. The dynamic semantics that Barker (2002) gives for ‘tall’ is then:

\[
\llbracket \text{tall} \rrbracket = \lambda x_e. \lambda W_{(s,t)}. \{ w \in W | w \in \text{tall}(d(w)(\llbracket \text{tall} \rrbracket), x) \} \tag{4.1}
\]
For example, assuming the copula is semantically vacuous and that $[\text{Feynman}]$ is indeed Richard Feynman, we see that

$$[\text{Feynman is tall}] = \lambda W_{(s,t)} \{ w \in W | w \in \text{tall}(d(w)([\text{tall}]), \text{Feynman}) \} \quad (4.2)$$

Here, we have applied the function defined in (4.1) to the meaning of the term ‘Feynman’—that is, Richard Feynman, and derived our new function: a function that takes us from a context—a set of worlds—to a subset of that context set in which Feynman possesses the degree of tallness that is required for the application of ‘tall’ in the original context.

The presence of $[\text{tall}]$ on both sides of the equation in (4.1) may seem illicit; how can (4.1) tell us what function $[\text{tall}]$ is if we must use $[\text{tall}]$ as input before we can define $[\text{tall}]$? The key is to consider (4.1) as imposing a certain defining constraint on $[\text{tall}]$. Consider the following scenario: suppose we have in our context one world $w_1$ containing one object $x_1$. Then the denotation of $[\text{tall}]$ is either $\{(x_1, \{w_1\}, \{w_1\})\}$ or $\{(x_1, \{w_1\}, \emptyset)\}$. If $x_1$’s height is greater than or equal to $d(w_1)([\text{tall}])$, then $[\text{tall}]$ will be the former. If it is not, then it will be the latter. Likewise, if $[\text{tall}]$ is the former, then $d(w_1, [\text{tall}])$ must be less than or equal to $x_1$’s height; if $[\text{tall}]$ is the latter, then $d(w_1, [\text{tall}])$ must be greater than $x_1$’s height. Thus (4.1) does not really define a denotation; it instead defines a constraint on $[\text{tall}]$, and just what the denotation of $[\text{tall}]$ is, will depend on $d$—which will in turn, as the latter pair of conditionals shows, depend on $[\text{tall}]$.

To see how an utterance of (4.2), given (4.1), can give rise to only descriptive information, only metalinguistic information, or some mixture of both, consider an utterance of ‘Feynman
is tall’ in the contexts shown in Figure 4.1; here, \( \iota(\max(\lambda d. \text{tall}(d, \text{Feynman}))) \) is the greatest degree \( d \) such that \( w_n \) is in \( \text{tall}(d, \text{Feynman}) \)—that is, the maximal degree of height of Feynman in \( w_n \).

If it is part of the common ground that the standard for tallness is 179 cm, but it is unknown exactly how tall Feynman is, then we start with \( W = \{w_1, w_2, w_3\} \) and end with \( W' = \{w_2, w_3\} \). If it is part of the common ground that Feynman is exactly 180 cm tall, but unknown just what counts for tall around here, then \( W = \{w_2, w_5\} \) and \( W' = \{w_2\} \). But if we know only that his height is between 178 and 182 cm, and that the standard of tallness is between 179 and 181 cm, then \( W = \{w_1, w_2, w_3, w_4, w_5, w_6\} \) and \( W' = \{w_2, w_3, w_6\} \). In the first case, we gain only descriptive information, since we gain no information about what counts as tall around here, but only information about the language-independent part of the world—in particular, Feynman’s height. In the second case we gain only metalinguistic information, since we gain no information about how tall Feynman is. In the third case, however, we gain some information of both kinds—or rather, we enough about how Feynman’s height and the requisite standard of tallness are related so as to eliminate certain combinations of the two.\(^3\)

\(^3\)Note that if we thought that the standard for tallness around here was either 177, 178, or 179 cm,
Prior to Barker (2002), R. C. Stalnaker (1978) introduced the ‘†’ operator (although see D. K. Lewis (1973), p. 63 for an earlier version of the †. Lewis himself cites Kamp (1971) and Vlach (1973) for temporal analogues of † and †, respectively. See also Kaplan (1979), Kaplan (1978), and Kaplan (1989).) To see the point of the † operator, consider if I utter ‘That is either Zsa Zsa Gabor or Elizabeth Anscombe’ about someone talking in the next room. Assume $w_1$ is a world in which that person is Zsa Zsa Gabor, $w_2$ is a world in which that person is Elizabeth Anscombe, and $w_3$ is a world in which that person is Tricia Nixon. Assume further that for a given context of utterance, a token of proper names and demonstrative pronouns denotes the same individual in every possible world. Then if we are in $w_1$, the proposition I express is the same as I would have expressed (again in $w_1$) by uttering ‘Zsa Zsa Gabor is either Zsa Zsa Gabor or Elizabeth Anscombe’. Mutatis mutandis for $w_2$ and $w_3$. We could display the propositions variably expressed as in Figure (4.2), where along the vertical axis we list the worlds as they determine the content of my utterance—the proposition I express—and along the horizontal axis we list the worlds as they determine the truth or falsity of the proposition I express. Thus the proposition I express in $w_3$—that Tricia Nixon is either Zsa Zsa Gabor or Elizabeth Anscombe—is false in $w_1$, $w_2$, and $w_3$. Stalnaker calls a matrix such as in Figure (4.2) a propositional concept.

and Feynman’s height was either 176, 177, or 178 cm, then we would gain some information about how tall Feynman is, and about what the standard for tallness is around here.
Now, Stalnaker lists three constraints on assertion:

1. A proposition asserted is always true in some but not all of the worlds in the context set.

2. Any assertive utterance should express a proposition, relative to each possible world in the context set, and that proposition should have a truth value in each possible world.

3. The same proposition is expressed relative to each possible world in the context set.

The motivation for the third constraint is that otherwise, the receiver does not know which worlds to eliminate: in our example, if he is in $w_1$ he is to eliminate no worlds from the context set, and if he is in $w_3$ he is to eliminate all worlds from the context set; but without knowing which world in the context set he is in, he does not know which worlds to eliminate from the context set. The motivation for the second constraint is that otherwise, if at any world in the context set the proposition expressed is neither true nor false, due either to the utterance not expressing a function from worlds to truth-values at all, or expressing only a partial such function, then it will be unclear whether to eliminate or retain that world. The motivation for the first constraint is that since the point of assertion is to eliminate certain worlds from the context set, to assert something true in every world is to attempt something that’s already been done, and to assert something false in every world in the context set leaves the receiver with no information at all: *ex falso, quodlibet.*

In our case at hand, we can attempt to rectify the violation of the third constraint by

---

4Stalnaker says that ‘To assert something incompatible with what is presupposed is self-defeating; one wants to reduce the context set, but not to eliminate it altogether.’ I think the problem he sees with eliminating the context set altogether is that then no information is available to the receiver.
shrinking the context set to exclude \( w_3 \), but this ends up violating the first constraint: any proposition I would have been expressing with my utterance would have been true in every world in the context set. Instead, the proposition I intuitively seem to be expressing is the one which is true in every world in which it is either Zsa Zsa Gabor or Elizabeth Anscombe next door, and false in every other world. Among the worlds in the context set, that proposition is the one that is true at \( w_1 \) and \( w_2 \), and false at \( w_3 \). More generally, in this case we want the proposition that is evaluated for truth or falsity at a world \( w_i \) relative to the denotation of ‘that’ at \( w_i \). If we index worlds in their dual roles as both determiners of the proposition expressed, and the places at which the propositions thus expressed are either true or false, like we have above, then this proposition will be the main diagonal of the resulting matrix. The dagger operator \( \dagger \) takes the main diagonal of matrix and projects it into the horizontal axis, thus bringing the resulting matrix into conformity with the third constraint.\(^5\)

Now, we might notice that we can get the same effects as Barker’s semantics—both descriptive and metalinguistic—by taking the diagonal as we see in Figure (4.3): Again, if the common ground (the context set) starts with the threshold for ‘tall’ being 179cm, but Feynman’s height is not part of the common ground, then we start with \( W = \{ w_1, w_2, w_3 \} \) and end with \( W' = \{ w_2, w_3 \} \), thus we have pure descriptive entailment. If it is part of the

\(^5\)Note that in this case the end result—removing \( w_3 \) from the context set—is the same whether we shrink the context set to bring the utterance into conformity with the third constraint, or apply the \( \dagger \) operator and then remove \( w_3 \) upon accepting the speaker’s assertion. But the mechanism is different: the former involves the receiver thinking that the context was defective, the latter does not. More generally, the shrinking or expanding of the context set is not guaranteed to have the same effect as applying the dagger operator and then shrinking the context set.
context set that Feynman is exactly 180 cm tall, but the context set does not determine just what counts as ‘tall’, we start with $W = \{w_2, w_5\}$ and end with $W' = \{w_2\}$. And if we know only that the standard for tallness is between 179 cm and 181 cm, and Feynman’s height is between 178 cm and 182 cm, then we start with $W = \{w_1, w_2, w_3, w_4, w_5, w_6\}$ and end with $W' = \{w_2, w_3, w_6\}$. We thus see that the dynamic semantics for vagueness that Barker proposes is equivalent to applying the $\dag$ operator.

### 4.2.2 From Stalnaker to Montague

We might also note we can also define the $\dag$-operator in the system of intensional logic found in Montague (1973), as the $\wedge \vee$-operator. Consider if our sentence ‘Feynman is tall’ expresses different propositions in different possible worlds, perhaps depending on the threshold for tallness in those worlds. For the sake of brevity, we let $w_1, w_2, w_3, w_4$ in Equation (4.3) be worlds $w_1, w_2, w_4, w_5$ in Figure (4.3), and let ‘Feynman is tall’ = $P$. Then following Montague we define $Type$ as the smallest set $X$ such that:

1. $e, t \in X$

2. if $a, b \in X$ then $\langle a, b \rangle \in X$
3. if \( a \in X \) then \( \langle s, a \rangle \in X \)

where \( e, t, s \) are distinct, fixed objects and not ordered pairs or triples. Where \( A \) is a domain of entities and \( W \) a set of possible worlds, we then recursively define \( D_{a,A,W} \), the set of possible denotations of type \( a \), as follows:

1. \( D_{e,A,W} = A \)

2. \( D_{t,A,W} = \{0, 1\} \)

3. \( D_{\langle a,b \rangle,A,W} = D_{b,A,W}D_{a,A,W} \)

4. \( D_{\langle s,a \rangle,A,W} = D_{a,A,W}W \)

We can then define a model \( \mathfrak{A} = \langle A, W, F \rangle \), where \( F \) is a function that assigns to every \( \alpha \in \text{Con}_a \), where \( \text{Con}_a \) is the set of constants of type \( a \), for every \( a \in \text{Type} \), a member of \( D_{w,A,W} \). Finally, we can define the intension of \( P \) relative to a model \( \mathfrak{A} \) and assignment function \( g \), \( [P]_{e}^{g} \), as that function \( f \) with domain \( W \) such that for \( w \in W \), \( f(w) = [P]_{w}^{A,g} \).\(^6\)

Then, suppose:

\(^6\)Montague actually included a set of times \( T \) and a simple ordering \( \leq \) over \( T \) in his models \( \mathfrak{A} = \langle A, W, T, \leq, F \rangle \), by relativizing interpretations in \( D_{a,A,W,T} \) (instead of \( D_{a,A,W} \)) to models, worlds, times, and assignment functions, and defined the range of \( F \) as a \( D_{a,A,W,T,W \times T} \). Note that we also depart from Montague’s original notation in using double brackets.
Montague specifies that if $\alpha \in D_{(s,a)}$ then $\llbracket \land \alpha \rrbracket^\mathfrak{a},w,g = \llbracket \alpha \rrbracket^\mathfrak{a},w,g(w)$. And, if $\alpha \in D_a$, then we can say that $\llbracket \land \alpha \rrbracket^\mathfrak{a},w,g$ is that function $f$ from the set of worlds $W$ such that for all $w' \in W$, $f(w') = \llbracket \alpha \rrbracket^\mathfrak{a},w',g$. Then for any $w \in W$, $\llbracket \land \land \land P \rrbracket^\mathfrak{a},w,g(w_1)$ is $f(w_1)$ which is $\llbracket \land \lor \rrbracket^\mathfrak{a},w_1,g$ which is $\llbracket \land \lor \rrbracket^\mathfrak{a},w,g$.

\begin{align*}
\llbracket P \rrbracket^\mathfrak{a},g = & \begin{bmatrix}
\alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 \\
0 & 1 & 0 & 1 \\
0 & 1 & 0 & 1 \\
0 & 1 & 0 & 1 \\
0 & 1 & 0 & 1 \\
\end{bmatrix} \\
\end{align*}

\[ (4.3) \]

Montague originally claimed that $\llbracket \land \alpha \rrbracket^\mathfrak{a},w,t,g = \llbracket \alpha \rrbracket^\mathfrak{a},w,g(w)$. As Montague and Thomason (1974) pointed out, Montague’s intention here is clear, but the definition is not quite correct, since then $\llbracket \alpha \rrbracket^\mathfrak{a},g$ is undefined when $\alpha$ is not a constant, which $\alpha$
\[ [P]^{\mathfrak{a},w_{1},g}(w_{1}) \] which is 0. Similarly, \([^{\wedge \vee}P]^{\mathfrak{a},w_{2},g}(w_{2})\] is \([P]^{\mathfrak{a},w_{2},g}(w_{2})\] which is 1, and similarly for \(w_{3}\) and \(w_{4}\). So for any \(w \in W\) we have:

\[ [^{\wedge \vee}P]^{\mathfrak{a},w,g} = \begin{pmatrix}
    w_{1} & \rightarrow & 0 \\
    w_{2} & \rightarrow & 1 \\
    w_{3} & \rightarrow & 0 \\
    w_{4} & \rightarrow & 0
\end{pmatrix} \quad (4.4) \]

Note that we see that \([^{\wedge \vee}P]^{\mathfrak{a},w,g}\] is not the same as any of the propositions \(P\) expresses in any given world. And, we can see that we are again tracing the diagonal as we determine \([^{\wedge \vee}P]^{\mathfrak{a},w,g}\].

### 4.2.3 From Stalnaker/Kaplan to Bayes-Grice

To see how we can get from Stalnaker’s diagonal proposition and \(\dagger\)-operator to the kind of Bayesian development of Gricean pragmatics seen in Chapter 2, consider that we can transform the propositional concept matrix into a \(\theta,h\) matrix by taking the \(\theta\)-equivalence class of worlds, ordering the classes by \(\theta\), ordering the worlds in a class by \(h\), and then the truth-value at \(\theta_{m},h_{n}\) is the truth-value of the \(h_{n}\) world in the \(\theta_{m}\)-class. Or equivalently, as we have the matrix displayed above, take the first row of the \(\theta,h\) matrix to be the first row of the sub-matrix of the propositional concept matrix such that that sub-matrix contains all may not be. Thomason thus defines when \(\alpha \in Con_{a}, [\alpha]^{\mathfrak{a},w,t,g} = F(\alpha)(\langle w,t \rangle)\), and then defines \([^{\wedge \alpha}]^{\mathfrak{a},w,t,g}\) as we have above \((modulo the relativization to world-time pairs, instead of simply worlds.) This seems to have become standard, see Dowty, Wall, and Peters (1981).
and only world-world pairs where both worlds have the first lowest $\theta$ value—in Figure (4.3) this sub-matrix will be the top left quadrant; take the second row of the $\theta, h$ matrix to be the second row of the sub-matrix of the propositional concept matrix such that that sub-matrix contains all world and only world-world pairs where both worlds have the second lowest $\theta$ value—in Figure (4.3) this will be the bottom right quadrant, and so on for propositional concepts containing more world-world pairs. (It does not matter if either procedure is done before or after applying $\dagger$. ) And of course we can reverse the operation to go from a $\theta, h$ matrix back to the propositional concept; to do so, we replicate sub-matrices horizontally to get back to the propositional concept prior to application of $\dagger$, and replicate sub-matrices vertically to get back to the propositional concept after application of $\dagger$. ) The resulting $\theta, h$ matrix is shown in Figure (4.4).

Now, we might note that diagonalization works just as well on worlds outside the context set as worlds inside the context set; thus, if we expanded Figure (4.3) to show more worlds, and then applied the same procedure as we used to make Figure (4.4), the result would be Figure (4.5). Notice, however, that diagonalization by itself does not give us Stalnaker or Barker’s picture. Diagonalization does not account for the existing (one might even say prior) context set: Barker thinks of the listener as having certain prior conceptions of what the range of heights is, and what the range of thresholds could be. Thus we might think of the listener as thinking of the range of heights as being between 178 cm and 182 cm, and the

<table>
<thead>
<tr>
<th></th>
<th>178</th>
<th>180</th>
<th>182</th>
</tr>
</thead>
<tbody>
<tr>
<td>179</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>181</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
</tbody>
</table>

Figure 4.4
range of thresholds as being between 179 cm and 181 cm. Then we would have Figure (4.6), where the relevant ranges of height and threshold values are highlighted in blue; applying conjunction to the corresponding cells in Figures (4.5) and (4.6) would yield the matrix in Figure (4.7).

Now, if in Figure (4.6) we replace true and false with 0 and 1, then we might think of the rows between 179 and 181 inclusive as having the value 1, and all other rows as having the value 0; similarly we might think of the columns between 178 and 182 inclusive as having the value 1, and all other columns as having the value 0. Then the values in the cells in

<table>
<thead>
<tr>
<th>175</th>
<th>176</th>
<th>177</th>
<th>178</th>
<th>179</th>
<th>180</th>
<th>181</th>
<th>182</th>
<th>183</th>
<th>184</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

Figure 4.5

<table>
<thead>
<tr>
<th>175</th>
<th>176</th>
<th>177</th>
<th>178</th>
<th>179</th>
<th>180</th>
<th>181</th>
<th>182</th>
<th>183</th>
<th>184</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

Figure 4.6
Figure (4.6) will be the product of the two. And again if in Figure (4.5) we replace true and false with 0 and 1, then we can see the denotation of a sentence such as ‘Feynman is tall’ on the Barker/Stalnaker picture as involving a product of three terms:

1. the value of the diagonal given that $h$ and $\theta$ combination,

2. the prior value of $\theta$, and

3. the prior value of $h$.

This should look roughly like the kind of Bayesian development of Gricean pragmatics that was discussed in Chapter 2, except that where the Bayesian has probabilities, this picture has true and false (or 1 and 0): on either the Barker or Stalnaker picture, each world represents a combination of $h$ and $\theta$ values, and we are in fact multiplying, for any given $h$ and $\theta$ combination, the truth-value assigned to that combination along the diagonal, by the prior truth-value assigned to $h$, and the prior truth-value assigned to $\theta$, in order to assign an updated truth-value to any $h$, $\theta$ combination: every $h$, $\theta$ combination that makes the utterance false, is assigned the value false. Consider then the kind of Bayes-Grice model
from Chapter 2, reprinted here in Equation (4.5). Here, instead of prior truth-values, we have prior probabilities, assigned by \( \phi(h) \) to each height \( h \) and by \( p(\theta) \) to each threshold \( \theta \). Instead of the diagonal, we have \( \sigma_{n-1}(u|h, \theta) \), which tells us the probability that the sender utters \( u \), given that Feynman is \( h \) tall, and the threshold for tallness is \( \theta \). And, just as we can consider the result of updating according to the Barker/Stalnaker model as the product of its three terms, so also we defined updating according to the Bayes-Grice picture, as the product of its three terms. Thus, mathematically, the Barker/Stalnaker model can be seen as a special case of:  

\[
\rho_n(h, \theta|u) = \frac{\sigma_{n-1}(u|h, \theta) \cdot p(\theta) \cdot \phi(h)}{\sigma_{n-1}(u)} \quad (4.5)
\]

Now, we could think of the values above the diagonal line in the \( \theta, h \) matrix in Figure (4.5) as answering the question: would the sender have uttered what they uttered given this \( h \) and \( \theta \) value? Or, we could think of the values above the diagonal line as just, would the utterance have been true if \( h \) and \( \theta \) were as they are in this world? These result in the same matrix in the Barker/Stalnaker picture, but perhaps different matrices in the Bayes-Grice picture: The former question becomes: what is the probability that the sender uttered what they uttered, conditional on this \( h \) and \( \theta \) value? The latter question becomes: what is the probability that the utterance is true, conditional on this \( h \) and \( \theta \) value? The latter will result in \( \sigma_{n-1}(u|h, \theta) \) being 1 for every \( h, \theta \) combination such that \( h \geq \theta \), and 0 otherwise, (thus capturing the  

\^8\text{This is clearest in the case when } ALT = \{Feynman is tall, Feynman is not tall, } \emptyset \text{ and thus we have Equation (22) from Chapter 2, but we can see the same point if we marginalize out } \theta_2 \text{ from the right side of the equation in Equation (11) before conditionalization on } h \text{ and } \theta_1 \text{ in Equation (11) from Chapter 2.}
truth-conditions of the utterance). The former question may take on non-trivial probabilities depending on the model of speaker behavior, as we see in Chapter 2. Either way, however, the Barker/Stalnaker model can be seen as a special case of a Bayes-Grice model.³⁹,¹⁰

4.2.4 Requirements on a Model of Interpretation

4.2.4.1 What we mean by interpretation

Now, I think that conformity with Equation (4.5) is a necessary condition on any Bayesian model of how, given utterances with vague predicates, listeners draw the pre-theoretically expected statistical inference patterns from beliefs concerning the truth-conditions of such utterances. To be clear, these inference patterns are also, roughly, ones that a speaker expects a listener to make on the basis of the speaker’s utterance: when I tell you ‘John is tall,’ and you nod and act as you do when you believe what I say, I expect you to increase

³⁹In Lassiter and Goodman (2017) the authors relate Stalnakerian update to the literal ρ₀ listener defined in Chapter 2; they claim that update is equivalent to set intersection on the worlds in the common ground and worlds in which the utterance is true; conditionalization is then set intersection followed by renormalization by the measure of the resulting set. That is, they observe the parallel between update of the common ground by the proposition expressed in the horizontal lines in Stalnaker’s picture, and how the literal listener in their picture updates their probabilities. In fact there is also a parallel between diagonalization and the Bayes-Grice models such as the one that Lassiter and Goodman (2017) propose, as we see above. See also Cumming (2007) for the relation between Bayes-Grice models and the diagonal.

¹⁰I will for now continue to think of φ(ḥ) and p(θ) as the probability of h and θ. Another possibility, given that the question is, would the sender have uttered what they uttered?, is to think of φ(ḥ) and p(θ) as the probability that the sender thinks the height is h and the threshold is θ, respectively. I postpone discussion of this possibility to future work.
the probability you would assign to John being 6'0", relative to what I think you would have assigned to John being 6'0" before I told you ‘John is tall’. And of course, you expect me to expect that you do this, as that is the point of your acting as you do when you believe me.

It is clear that we are here approaching a probabilistic analogue of the notion of common knowledge discussed in D. Lewis (1969), and defining this analogue requires a probabilistic analogue of the iterated belief hierarchy. I will thus offer below a few remarks concerning how a probabilistic analogue of the iterated belief hierarchy might be defined, where such a notion might be useful, and how such a notion might be approximated, in order to at least justify the assumption here that we can take \( \phi(h) \) to be (roughly) what, for the purposes of the discourse, the speaker is disposed to act as if he believes the receiver is disposed to act as if he believes the probability to be, that the object under discussion is \( h \) degrees of deviation above the average height of the relevant comparison class. First, however, I want to more carefully relate the inference patterns under discussion here to those discussed in Chapter 2, and isolate them from other inferences failure to explain which ought not count against the Bayesian models captured by Equation (4.5).

It might seem as if these inference patterns are distinct from the ones that the Bayes-Grice models discussed in Chapter 2 were attempting to explain, since in that chapter we did not explicitly require that the statistical inference patterns that we were attempting to get the various models to generate also be ones that, roughly, the speaker expects a listener to make on the basis of their utterance. However, I think it is clear that speakers do in fact expect listeners to make those inference patterns that we were attempting to get the models to generate: thus, for example, speakers (roughly) expect that upon hearing ‘Feynman is
tall’, listeners change their probability distribution for Feynman’s height in accordance with Figure (2.1) in Chapter 2. And, I think we should take it to be no objection to those models if in some cases listeners actually infer on the basis of the speaker’s utterance something additional to what the speaker expects them to infer on the basis of their utterance. For example, if a listener were to infer that the speaker is not from around these parts on account of their pronouncing of a local place name in a manner at odds with regional standards, it should be considered no failure of such a model that it failed to predict such inferences; and this because only to the degree that an inference pattern is distinctively linguistic, should we require that the Bayes-Grice models discussed in Chapter 2 predict that pattern to occur. I am thus assuming that there is a distinctively linguistic manner of gaining information about the world, and it is this capacity that requires explanation; I take Equation (4.5) to govern any Bayesian model that might give us such an explanation. Such a model may appeal to cognitive faculties and processes that are not uniquely linguistic, but it should nonetheless be able to characterize the differences between, for example, how we gain information about Feynman’s height upon being told ‘Feynman is tall’, and how we gain information about Feynman’s height upon seeing him, or learning that he played center for his high school basketball team. And of course, part of what is distinctive about how information is transmitted via language is that speakers expect listeners to make those inferences that they do in fact make on the basis of the speaker’s utterance, and failure to predict inferences which are not thus expected ought not count against such a model.¹¹

¹¹This characterization of (part of) what is distinctive about how information is transmitted via language may seem to be at odds with the claims made in Chapter 3; however, I claim there only that under the listed conditions, the encoding and decoding behavior characteristic of vague predicates emerges, and this
Admittedly, it is possible to take this too far: if experimental results demonstrate that upon hearing someone called ‘tall’, listeners systematically over-estimate that person’s height relative to what speakers expect listeners to do, we certainly could not reject such evidence out of hand on the grounds that that is not precisely what speakers expect listeners to infer, and is hence unable to count as evidence against a theory that predicts that such over-estimation does not occur. Still, it seems \textit{prima facie} reasonable to start from the position that there are patterns of statistical inference that a listener draws on the basis of a speaker’s utterance, which the speaker expects a listener to draw on the basis of their utterance; we then take $\rho_n(h|u)$ (after marginalizing out $\theta$ from $\rho_n(h,\theta|u)$) as what, for the purposes of the discourse, the receiver is disposed to act as if he believes the probability to be, given the speaker’s utterance, that the object under discussion is $h$ degrees of deviation above the average height of the relevant comparison class. Likewise, we take $\phi(h)$ as what, for the purposes of the discourse, the receiver is disposed to act as if he believes the probability to be, that the object under discussion is $h$ degrees of deviation above the average height of the relevant comparison class.$^{12}$

is consistent with the possibility that such behavior is preserved under the conditions which lead speakers to expect listeners to make those inferences that they do make on the basis of the speaker’s utterance. Furthermore, even under assumptions of bounded rationality in which an $n$ level sender probabilistically chooses a signal that tends to maximize utility—understood as the informativity to a $n - 1$ level receiver whose decoding behavior is dictated by the transposition of the sender’s own $n - 1$ level encoding weight matrix, minus the cost of the utterance—and an $n$ level receiver probabilistically Bayesian-ly decodes a signal by accounting for the encoding probabilities of an $n - 1$ level sender whose encoding behavior is dictated by the transposition of the receiver’s own lower-order decoding weight matrix, simulations (not yet reported) indicate that such vague signaling behavior persists even when $n = 3$.

$^{12}$I phrase this notion in terms what the receiver is disposed to act as if he believes the probability to be,
Since we assume that a speaker expects a listener to engage in precisely these statistical inference patterns, we also take $\phi(h)$ to be (roughly) what, for the purposes of the discourse, the speaker is disposed to act as if he believes the receiver is disposed to act as if he believes the probability to be, that the object under discussion is $h$ degrees of deviation above the average height of the relevant comparison class. How can we make this roughly more precise, and thus define a probabilistic analogue of the iterated belief hierarchy? If a 0th-level receiver is defined by what, for the purposes of the discourse, the receiver is disposed to act as if he believes the probability $x$ to be, that the object under discussion is $h$ degrees of deviation above the average height of the relevant comparison class, then we define a 1st-level sender by what, for the purposes of the discourse, the sender is disposed to act as if he believes the probability $y$ to be, that the receiver is disposed to act as if he believes the probability $x$ to be, that the object under discussion is $h$ degrees of deviation above the average height of the relevant comparison class. Similarly, we would define a 1st-level receiver by what, for

as opposed to what the receiver is disposed to act as if the probability is, in order to highlight that this is a psychologically loaded notion of action: in a less psychologically loaded sense of action, I am disposed to act as my body is subject to the laws of special relativity, just because like any physical object my body is in fact subject to the laws of special relativity. But I am not, in this sense of action, thereby disposed to act as if I believe that my body is subject to the laws of special relativity. This notion is intended to be a probabilistic analogue of Stalnaker’s notion of what is added to the common ground of the conversation. (See Stalnaker, ‘Assertion’.) Note also that the disposition to act seems not required to be de se attitude: to say that John is disposed to act as if that guy over there (where John does not think that that guy over there is him, but is in fact him) believes Feynman is 6’3” tall, is the same as to say that John is disposed to act as if he himself believes Feynman is 6’3” tall; which is just to say that among other things if you were to ask him how tall Feynman is, he would say that Feynman is 6’3” tall.
the purposes of the discourse, the receiver is disposed to act as if he believes the probability
$y$ to be, that the sender is disposed to act as if he believes the probability $x$ to be, that the
object under discussion is $h$ degrees of deviation above the average height of the relevant
comparison class.

To see how this probabilistic analogue of the iterated belief hierarchy might be useful, con-
sider that one plausible manner for generating numbers for the latter notion would be to
define $\rho_n(\sigma_{n-1}(h') = x|h)$: given that Feynman is $h$ degrees of deviation above the average
height of the American adult male, what probability $y$ would the receiver be disposed to act
for the purposes of the discourse as if he would assign to the sender assigning probability $x$ to
height $h'$? On this view, the receiver thinks that no matter what height $h$ Feynman might ac-
tually be, the sender assigns some probability to Feynman being $h'$. However, just how much
might depend on whether the receiver thinks of the sender as an eyewitness as to Feynman’s
height, or as just passing along how he has heard others describe Feynman. Thus, if Feynman
is $6'3''$ tall, what is the probability that the sender assigns a high probability to him being
$5'9''$ tall? If the receiver thinks of the sender as an eyewitness, $\rho_n(\sigma(h' = 5'9'') = x|h = 6'3'')$
will be quite low when $x$ is high; if the receiver thinks of the sender as just passing along how
he has heard others describe Feynman, $\rho_n(\sigma(h' = 5'9'') = x|h = 6'3''')$ will be comparatively
greater for the same value of $x$. We can then think of the receiver as taking his posterior
height distribution given the speaker’s utterance—$\rho_n(h|u)$—and using that to calculate a
posterior probability $p(x)$ that the sender assigns probability $x$ to Feynman being $h'$ tall,
after marginalizing out $h$.

All of this is quite complicated, however, and I think that for the present purposes we can
define the *expected probability* \( z \) as \( \int_0^\infty xp(x)dx \), and then say that \( z \) is what the sender is disposed to act as if he believes the receiver is disposed to act as if he believes the expected probability to be, that the object under discussion is \( h' \) degrees of deviation above the average height of the comparison class. Then, when we say that we take \( \phi(h) \) to be (roughly) what, for the purposes of the discourse, the receiver is disposed to act as if he believes the speaker is disposed to act as if he believes the probability to be, that the object under discussion is \( h \) degrees of deviation above the average height of the relevant comparison class, we are really saying that \( \phi(h) \approx z \). Similarly for when we say that we take \( \phi(h) \) to be (roughly) what, for the purposes of the discourse, the sender is disposed to act as if he believes the receiver is disposed to act as if he believes the probability to be, that the object under discussion is \( h \) degrees of deviation above the average height of the relevant comparison class.

### 4.2.4.2 Why this requirement

As for why conformity with Equation (4.5) is a necessary condition on any Bayesian model of how, given utterances containing vague predicates, listeners draw the pre-theoretically expected statistical inference patterns from beliefs concerning the truth-conditions of such utterances, consider first the contrast between the truth-conditions of utterances containing vague predicates and the statistical inference patterns listeners draw from them: the truth-conditions of such utterances must be defined relative to a point on the degree scale such that if the subject of the utterance has a degree of the property encoded by the predicate greater than (or greater than or equal to) that threshold, the utterance is true; else, false. Take ‘John is tall’: for any threshold relative to which the utterance is defined to be true, the utterance
is equally true for all heights greater than (or greater than or equal to) that threshold; in contrast, given an utterance of ‘John is tall’, listeners seem to smoothly and monotonically increase the probability they assign to John’s height from 0 feet up to approximately 6’0”, and then smoothly and monotonically decrease the probability they assign to John’s height above approximately 6’0”. Even if ‘John is tall’ is equally true if John is 6’6” or 7’6”, listeners still assign a lower probability to John being 7’6” than to him being 6’6”, given an utterance of ‘John is tall’. The obvious reason is that the prior probability they assign to John being 7’6” is lower than the prior probability they assign to John being 6’6”. Thus, given an utterance containing a vague predicate, listeners’ statistical inference patterns seem to account for a prior probability distribution for the property encoded by that predicate.

Note that this prior probability distribution does not have to be receiver’s own prior probability distribution for John’s height, since $h$ represents only what, for the purposes of the discourse, the receiver is disposed to act as if he believes the probability to be, that the object under discussion is $h$ degrees of deviation above the average height of the relevant comparison class, and what, for the purposes of the discourse, the sender is disposed to act as if he believes the receiver is disposed to act as if he believes the probability to be, that the object under discussion is $h$ degrees of deviation above the average height of the relevant comparison class. Thus, the receiver might be disposed for the purposes of the discourse to act as if he believes the probability distribution for Feynman’s height to be in accordance with the background distribution for American adult males, and the sender might be disposed for the purposes of the discourse to act as if he believes that the receiver is disposed to act as if he believes the probability distribution for Feynman’s height to be in
accordance with the background distribution for American adult males, even if it is neither the case that the receiver actually believes the probability distribution for Feynman’s height to be in accordance with the background distribution for American adult males, nor that the sender actually believes that the receiver believes the probability distribution for Feynman’s height to be in accordance with the background distribution for American adult males, if for example the receiver is interviewing the sender in front of a studio audience: ‘So, tell me about Richard Feynman—like, what did he look like?’ the receiver might say to the sender, the sender then responding, ‘Well, he was tall...’

Second, consider ‘John is tall’ again: given a speaker’s utterance of this sentence, listeners typically will have a probability distribution for John’s height—the result of the pre-theoretically expected statistical inference patterns—that is distinct from some prior probability distribution for John’s height. That is, listeners have a posterior probability distribution for John’s height given a speaker’s utterance of ‘John is tall’; this is the statistical inference pattern we are trying to generate. Assuming that interpretation happens via a Bayesian reasoning process, the only way to generate a posterior probability distribution for John’s height given a speaker’s utterance of ‘John is tall’, from a prior probability distribution for John’s height, is via a conditional probability function for the speaker’s utterance given a height for John. Thus our model must at the least conform to the following equation:

\[
\rho_n(h|u) = \frac{\sigma_{n-1}(u|h) \times \phi(h)}{\sigma_{n-1}(u)} \tag{4.6}
\]

where \(u = \text{‘John is tall’}\). Of course, this equation fails to account for the thresholds relative
to which the truth-conditions for utterances containing vague predicates must be defined; nonetheless, it seems that the clear way to do so within the Bayesian framework we have adopted so far is exactly parallel to how we accounted for listeners’ prior distributions for heights: we define a conditional probability function for the speakers utterance given both a height and a threshold, and multiply by both a prior probability distribution for the threshold and a prior probability distribution for the subject’s height. Indeed, within a Bayesian framework for inference, it seems that accounting for a variable just is multiplying a conditional probability by a prior probability distribution. We thus arrive at Equation (4.5), as long as we assume the statistical independence of $\phi(h)$ and $\theta$.

### 4.3 Against Truth-Conditional Statistical Inference

#### 4.3.1 A General Argument For Non-Truth Conditional Inference

In this section I want to offer three arguments against any Bayesian model of how, given utterances containing vague predicates, listeners draw the pre-theoretically expected statistical inference patterns from beliefs concerning the truth-conditions of such utterances; these patterns also being those that a speaker expects a listener to engage in. The first argument concerns purely descriptive uses of vague predicates like ‘tall’, where no information is added to the probabilistic common ground about the degree of height required in the context of utterance to warrant usage of that predicate.\(^{13}\) For example, suppose you ask me, ‘Describe Richard Feynman to me. What did he look like? Was he at all like how he is commonly

\(^{13}\)Barker himself defines purely descriptive uses in terms of the (non-probabilistic) common ground.
portrayed in the media? Suppose further that it is common knowledge (or rather, roughly, our probabilistic analogue of common knowledge) that you know no more about Richard Feynman than that he was an American adult male. If I say, ‘Well, he was tall, and certainly he could be charming at times’, intuitively, you seem to gain no information about how I use ‘tall’, or about the standards for tallness around these parts, and I believe that you gain no such information. Thus, it seems that in these cases,

\[ \rho_n(\theta|u) = \int_0^\infty \rho_n(h, \theta|u) \, dh = p(\theta). \] (4.7)

Note that here and in what follows the integral limits in this case are set to 0 and \( \infty \) because we are talking about ‘tall’, and at least intuitively, for any comparison class there is a finite number of standard deviations from the measure of central tendency down to the lower bound on heights—nothing that is the kind of thing that can have heights, can have a height that is less than or equal to 0—but no upper bound on the number of standard deviations above that measure of central tendency. Next, it seems that in purely descriptive uses of vague predicates, the probability distributions for heights and thresholds are statistically independent in the posterior: this means that upon my saying that Feynman is tall, the ratio of the probability that you would assign to Feynman being 5′6″ and the threshold for tallness being 5′0″ to the probability that you would assign to Feynman being 5′6″ and the threshold for tallness being 6′0″, is the same as the ratio of the probability that you would assign to Feynman being 6′6″ and the threshold for tallness being 5′0″ to the probability that you would assign to Feynman being 6′6″ and the threshold for tallness being 6′0″. That is, the ratio of probability of the threshold being 5′0″ vs 6′0″, is the same for all heights, in
the posterior distribution. Then

$$
\rho_n(h, \theta|u) = \rho_n(h|u) \ast \rho_n(\theta|u).
$$

(4.8)

It follows that

$$
\rho_n(\theta|u) \ast \rho_n(h|u) = \sigma_{n-1}(u|h, \theta) \ast p(\theta) \ast \phi(h)
$$

(4.9)

and then

$$
\sigma_{n-1}(u|h, \theta) = \frac{\rho(h|u)}{\phi(h)}
$$

(4.10)

It then follows that in purely descriptive uses of ‘Feynman is tall’, for any given height, the sender’s behavior is independent of the threshold, and can be determined by the prior and posterior probability assigned to that height. In other words, in purely descriptive uses, the probability that the sender will say ‘Feynman is tall’ is determined solely by Feynman’s height, and for any given height is the same for any threshold value, whether the threshold is greater or less than that height. If listeners use Bayesian reasoning to derive the pre-theoretically expected changes in probability assigned to any given height, it then seems they do so by considering the probability that the speaker would say ‘Feynman is tall’ given that Feynman is that height, independent of any threshold for usage, and multiplying that
by the prior probability that Feynman is that height, as in Equation (4.6). Of course, the truth-conditions for vague predicates like ‘tall’ are only defined relative to thresholds, and thus it seems that the truth-conditions play no role in the how receivers derive the pre-theoretically expected statistical inference patterns in cases of purely descriptive uses of utterances containing vague predicates. We thus have the following argument:

**Argument for Non-Truth Conditional Inference**

1. In purely descriptive uses of ‘tall’, the posterior threshold distribution \( \rho_n(\theta|u) \) is identical to the prior threshold distribution \( p(\theta) \).

2. In purely descriptive uses of ‘tall’, the posterior threshold distribution \( \rho_n(\theta|u) \) is statistically independent with respect to the posterior height distribution \( \rho_n(h|u) \).

3. Hence, in purely descriptive uses of ‘tall’, the probability of the utterance \( \sigma_{n-1}(u|h, \theta) \) is independent of the threshold variable.

4. If in purely descriptive uses of ‘tall’, the probability of the utterance \( \sigma_{n-1}(u|h, \theta) \) is independent of the threshold variable, then any listener using Bayesian reasoning to draw the pre-theoretically expected statistical inference patterns based on the probability of the utterance, does not do so by accounting for a threshold for ‘tall’.

5. If any listener using Bayesian reasoning to draw the pre-theoretically expected statistical inference patterns based on the probability of the utterance, does not do so by accounting for a threshold for ‘tall’, and if truth-conditions for ‘tall’ are defined only relative to a threshold for ‘tall’, then the truth-conditions for ‘tall’ do not feature in any Bayesian explanation of how listeners draw the pre-theoretically expected statistical
6. Truth-conditions for ‘tall’ are defined only relative to thresholds.

7. Therefore, the truth-conditions for ‘tall’ do not feature in any Bayesian explanation of how listeners draw the pre-theoretically expected statistical inference patterns.

Here, premise 3 follows from premise 1 and 2 as a matter of mathematical fact. Premise 4 is a matter of Bayesian reasoning: if the probability of the utterance is independent of the threshold, then for any two threshold values, the sending probabilities are the same for a fixed height, and hence the posterior probability assigned to that height after conditionalizing on the utterance will be unchanged between the two threshold values. Since the posterior probabilities are unchanged between the two threshold values, any two prior probability distribution over thresholds will result in the same posterior distribution over heights, and hence we cannot explain the appropriate statistical inferences even partly in terms of a Bayesian reasoner accounting for the threshold for ‘tall’, since the distribution for that threshold makes no difference for the statistical inferences, (at least, in purely descriptive uses.)

The argument for premise 5 is simply that since the truth-conditions are defined relative to thresholds, without an assignment of probabilities to threshold values, and thus by extension without a threshold value at all, any listener using Bayesian reasoning to draw the pre-theoretically expected statistical inference patterns cannot use the truth-conditions in a model of speaker behavior to derive the appropriate statistical inferences. For example, consider the model of Lassiter and Goodman (2017) discussed in Chapter 2: The probability that the speaker will say ‘Feynman is tall’ is proportional to the utility of the utterance, which
is defined partly in terms of the informativity of the utterance to a literal listener, which is itself defined partly in terms of the truth-conditions of the utterance. This then requires a value to be assigned to the threshold, without which the truth-value of the expression is undefined. Similarly, for the modified version of Lassiter and Goodman (2017) presented in §8 of Chapter 2: the distance-weighted definition of informativity to a literal listener again requires a value to be assigned to the threshold, in order for the probability-weighted distance of the literal posterior to be defined. Premise 6 is a widely-held semantic claim. This leaves premises 1 and 2 as the most substantive, and it is to them that I want to turn next.

Given how we have described what a purely descriptive use of a vague predicate is, premise 1 seems true by definition. What seems more important to point out, then, is that the vast majority of uses of ‘tall’ are purely descriptive uses: among competent users of English, it is part the probabilistic common ground that all parties to a discourse have acquired the word ‘tall’; and thus from any given usage of ‘tall’ among such users, no information about the distribution for the standard for tallness is added to the probabilistic common ground, when the probabilistic common ground contains no more information about the height of the object other than that it is a member of relevant comparison class. Thus, knowing no more about Feynman’s height than that he is an American adult male, and then hearing me call him ‘tall’ you learn something about his height, but nothing about what counts as tall. Note that if we consider the weaker notion of just what you believe about the probability distributions for Feynman’s height and the thresholds for tallness and what I (the speaker) believe that you believe about the probability distributions for Feynman’s height and the thresholds for
tallness, the point remains: if I believe that you believe no more about Feynman’s height than that he is an American adult male, and I believe that you believe no more or less about the probability distribution for the thresholds for tallness than what any competent user of English believes, then after I call Feynman tall, I still believe that you believe no more or less about the probability distribution for the thresholds for tallness than what any competent user of English believes. But I do believe that you know more about Feynman’s height than you did before. And typically, you do believe no more or less about the probability distribution for the thresholds for tallness than what any competent user of English believes, which is what you believed before; and, you do know more about Feynman’s height than you did before.14

Regarding premise 2, since we are discussing the probabilistic common ground, it is surely relevant that after hearing the utterance the listener does not think the speaker thinks that even though the probability of John being 5’6” is low, the probability of the threshold being 5’0” and John being 5’6” versus the probability of the threshold being 6’0” and John being 5’6”, is any different than the the probability of the threshold being 5’0” and John being 6’6” versus the probability of the threshold being 6’0” and John being 6’6”. The easier way to phrase this is to ask if the speaker thinks that the probability that John is 5’6” and the threshold is 5’0” is any different than the probability that John is 6’6” and the threshold is 5’0”: the correct answer is yes, but the intuitive answer is no. The intuitive answer is not the correct answer because we read the conjunction as a conditional probability. At any

14Note that we simplify away from what for the purposes of the discourse you and I are disposed to act as if these probabilities might be, and assuming that there are no objections from you, the listener, to my assertion that Feynman is tall, but the point remains, I think.
rate, I think that for the speaker, John’s height makes no difference to the probability of the threshold, and the listener believes correctly it makes no difference for the speaker. Will it be the case that Feynman’s height makes a difference for the threshold probability distribution for the listener, but the listener thinks that Feynman’s height makes no difference for the threshold probability for the speaker? That would seem to predict that upon finding out that Feynman is 5’6”, the listener will think that the sending probabilities are different than what the listener thinks the speaker will think the sending probabilities are like, upon similarly finding out the Feynman is 5’6”. And that seems wrong: if Feynman’s height makes a difference for the listener’s threshold probability distribution it would seem to do so by being lower for lower heights, so as to make the speaker’s utterance come out true; but this would entail that upon finding out that Feynman is 5’6”, the listener will think that the sending probabilities are skewed lower, but will think that the sender thinks the threshold probabilities are unchanged. So if the listener thinks that Feynman’s height makes no difference for the threshold probabilities for the speaker, Feynman’s height should make no difference for the threshold probabilities for the listener too.

One objection is that perhaps the listener thinks the speaker would assign no probability to Feynman being 5’6” and the threshold for tallness being 6’6”, because in that case the speaker would not call Feynman tall. But this is false: the speaker might assign some small probability to Feynman being 5’6”, and some probability to the threshold being 6’6”, and still call Feynman tall. Of course, the listener thinks that there’s no probability that both the speaker is certain that Feynman is 5’6”, and the speaker does not call ‘tall’ anyone he is certain is less than 6’6” tall. But that is different from the listener thinking that the speaker
might assign some low but still non-zero probability to Feynman being 5'6'', and some low but non-zero probability to someone shorter than 6'6'' being called ‘tall’ by competent speakers of English. Even assuming the listener only attributes consistent mental states to the speaker, there is nothing inconsistent about such a probability assignment.

On the other side of the hierarchy, I think that a speaker does not think that after hearing the utterance, a listener thinks that even though the probability of John being 5'6'' is low, the probability of the threshold being 5'0'' and John being 5'6'' versus the probability of the threshold being 6'0'' and John being 5'6'', is any different than the the probability of the threshold being 5'0'' and John being 6'0'' versus the probability of the threshold being 6'0'' and John being 6'0''. Surely in such cases I the speaker think that you the listener think that whether John is 5'6'' or 6'6'', the probability that someone 6'0'' would be called tall is the same, even after hearing my utterance. Consider that prior to my utterance I the speaker think that you the listener think that given that Feynman is 5'6'', the probability that someone 5'1'', 5'2'', ... and so on will be called ‘tall’ takes on values x, y,... and so on. I think prior to my utterance, I the speaker think that you the listener think that given that Feynman is 6'6'', these probabilities are the same. In the kind of prototypical uses of ‘tall’ that we have been talking about, do speakers think that uttering ‘Feynman is tall’ has the effect of making listeners shift the probabilities that various heights will count as ‘tall’, differently at different heights, even while keeping the overall sending probabilities over heights the same? I think clearly not: Suppose I have it on good authority that Feynman is tall. If I said to you ‘Feynman is tall’, given that this is a prototypical use of ‘tall’ would I expect you to think it more likely that someone 5'6'' counts as tall, if my source
was wrong and you were to find out later that Feynman is 5'9" tall, than if my source was right and you were to find out later that Feynman is 6'3" tall? Clearly not. If you were to report back to me that you had discovered that Feynman was only 5'6" tall, I would not expect you to think that what I had meant all along, by ‘tall’, involved some much lower standards than we normally associate with ‘tall’. You would think I had spoken wrongly, not that I had been misunderstood, and I would expect you to think I had spoken wrongly. And of course, if you would think I had spoken wrongly, then it seems like your posterior probability distribution for the threshold for ‘tall’ given that Feynman is 5'6" remains the same as your prior probability distribution for the threshold, and thus also for any height besides 5'6", so that your posterior threshold distribution is independent of your posterior height distribution.

Now, one might object to this argument on the grounds that we cannot assume that we can probe these joint posterior probabilities by asking, ‘What would someone think of the threshold probabilities after hearing ‘John is tall’ and then actually observing John’s height?’, or ‘What would someone think of the threshold probabilities after hearing ‘John is tall’ and then finding out John’s actual height?’, because we do not always monotonically narrow down a space of probabilities. What someone would think of the threshold probabilities, might be affected by what they learn from finding out his actual height from more trusted sources, or from their first-hand observations; perhaps in the posterior from hearing ‘John is tall’ the threshold distribution really is not independent from the height distribution, (and at any given height, the threshold distribution is in fact effectively ‘redistributed’ to be below that height) but this is masked by the fact that upon finding out John’s actual height, the
receiver in this scenario rejects the assertion ‘John is tall’, including any information about the threshold contained in that assertion. Having rejected the assertion, the receiver returns to the prior threshold distribution. When we think about what someone would think of the threshold probabilities after finding out John is only 5'6" tall, we come upon these prior probabilities to which the listener has reverted.

In response to this objection one might note that in some cases we do seem to monotonically narrow down the space of probabilities: for instance, when some some readers of Genesis reinterpret ‘day’ to mean a vastly longer time period than a literal day. Here, we have a case where something—the old age of the universe—that seems improbable, perhaps even completely ruled out, given the truth of the sentence, is later judged to be highly probable, and so we change what is allowed to count as a day in order to maintain the truth of the sentence, or perhaps, to maintain the presumption that however much time it took to create the universe, that is how long the Author said it took. Ignore for now that ‘day’ is a nominal and that it seems in principle absolutely precise, 24 hours and no more nor less—the latter stipulation perhaps more in keeping with the original Ancient Near Eastern context. If we do seem to monotonically narrow down the space of probabilities in this case, so the response might go, then why do we not do so with ‘tall’?

However, I do not think this is a good response to the objection, insofar as we in this case seem to shift the threshold down, at least the threshold for the author’s use of the word, contra the parallel claim that I made about ‘tall’. Having not rejected the assertion and thus not reverted to the prior threshold distribution, we seem to be seeing that the posterior threshold distribution really is shifted down. For someone who wants to reject premise 2, the
difference between the two cases might then be claimed to be the amount of trust given to the utterer; where the utterer is human and thus not so wholly trusted, the assertion is rejected along with the lower threshold distribution; where the utterer is divine and thus wholly trusted, the assertion remains accepted along with the lower threshold distribution. The statistical interpretation in both cases, however, might be taken to include a non-uniform threshold posterior.

Now, I think it is clear that speakers do sometimes use ‘tall’ to convey information about how ‘tall’ is used, and that listeners do sometimes gain information about how ‘tall’ is used from usages of ‘tall’, and that speakers expect them to gain such information on the basis of the speaker’s usage of ‘tall’. Thus, any complete model of how listeners make inferences from speaker’s utterances will have to have a component with a threshold variable in the posterior distribution, if that inferential process is Bayesian; since what we infer about the thresholds for ‘tall’ in such cases is sensitive to what we think the height of the object is, there must be a height variable in the prior, and thus in the posterior. Within this component, we can permit that the resulting joint posterior the for heights and thresholds may not be statistically independent.

Having thus granted that in non-prototypical uses of ‘tall’ there may be a posterior threshold distribution that is not statistically independent from the posterior distribution over heights, I think a better response to the objection is to point out what I think is an instructively contrasting case where we do conditionalize on something that, given what we have been told, we consider to be unlikely. Suppose I say to you, ‘Feynman is fat.’ I think you would think it quite unlikely that an American adult male called ‘fat’ weighed 150 lbs, but if you
were to find out that Feynman did indeed weigh 150 lbs, you would infer that he must be rather short. Note that like ‘tall’, ‘fat’ is a relative gradable adjective that appeals to at least two dimensions: weight and height. Thus, we do sometimes change our probability distribution over a property—height—relevant to the degree to which some object ‘counts as’ some predicate, given information about some other property—weight—which is also relevant to the degree to which that object ‘counts as’ that predicate. This is in contrast to flat out rejecting the assertion that the person is ‘fat’, and along with it the lower distribution for heights upon finding out that their weight is only 150 lbs. (Assume even that I am considered just as trustworthy in both cases, so that we get the exact parallel.) This then seems to imply that thresholds are, in a sense, different than other degrees over which relative gradable adjectives are distributed: unlike height and weight, the degree is more resistant to redistribution. The upshot is that we get to keep a distinction between semantics and pragmatics, in a sense: the thresholds are distinct from the other properties encoded in the word. But, at any rate, if we do sometimes monotonically narrow down the space of probabilities as in the case of ‘fat’, why do we not do so in the case of ‘tall’? Absent a motivation independent of the need to preserve thresholds (and thus truth-conditions) in the interpretation of vague predicates, it seems the best explanation is that the threshold distribution is not really part of the posterior, and hence not part of the prior, (that is, the threshold distribution is independent of the height distribution in the posterior, and is identical to the prior threshold distribution.)

Finally, even if we grant that asking ‘What would someone think of the threshold probabilities after hearing ‘John is tall’ and then actually observing John’s height?’ or ‘What would
someone think of the threshold probabilities after hearing ‘John is tall’ and then finding out
John’s actual height?’ are not good ways of probing the posterior probability distribution
for \( h \) and \( \theta \) from hearing ‘John is tall’, I think there are other ways of probing the posterior
distribution, and which open the door for experimental confirmation of premise 2 (and 1).
For example, if a speaker were to call someone ‘tall’ we might ask what odds we would have
to set in order for the listener to bet $1 that the person is shorter than 6’0”, 5’11”, ..., 5’6”
and so on. The difference in odds is the distribution for heights. If the odds are \( x : y \) then
the probability is \( y/(x + y) \). Here \( x \) is what the house puts up and \( y = \$1; \) the question is
then what the minimum \( \$x \) that the house must put up for the receiver to take the bet. It
might seem confusing when \( x < y \), (and thus the receiver thinks there’s even money that
John is over or under that given height), but we can just rescale and set \( x = \$1 \) and ask, how
much would the receiver bet that John is less than, say, 6’6” tall, to win \( \$1? \) Then we could
also ask the same question about John’s height and various threshold values: how much
would the receiver bet that John is 5’6” and the threshold is 5’0”, versus how much would
the receiver bet that John is 5’6” and the threshold is 6’0”? We might then compare that
ratio to the ratio of how much would the receiver bet that John is 6’6” and the threshold
is 5’0”, versus how much would the receiver bet that John is 6’6” and the threshold is 6’0”.
The prediction from premise 2 is that these ratios should be the same. We might similar
confirmation of premise 1; if the results accord with my intuitions, it would provide strong
evidence that in prototypical, purely descriptive uses of vague words like ‘tall’, if listeners
use Bayesian reasoning to draw the pre-theoretically expected statistical inference patterns
based on the probability of the utterance, they do not do so by accounting for a threshold
for ‘tall’; by extension, it would follow that the truth-conditions for ‘tall’ do not feature
in any Bayesian explanation of how listeners draw the pre-theoretically expected statistical inference patterns.

4.3.2 A Disjunctive Argument against Truth-Conditional Inference

Given that conformity with Equation (4.5) is a requirement for any plausible model of generating the pre-theoretically expected probabilistic listener inference patterns from utterances of sentences containing vague predicates, from beliefs concerning the truth-conditions of such utterances, we might next categorize the manners in which a model might put ‘numerical meat’ on the abstract bones of Equation (4.5).

4.3.2.1 A Realistic Threshold Distributions

First, we might ask what a plausible prior threshold distribution might be. I think it is clear that if listeners have a prior threshold distribution in any Bayesian calculation leading to the pre-theoretically expected statistical inferences, it must be genuinely informative, and thus not a uniform distribution as in Lassiter and Goodman (2017).\footnote{Note that the sense of ‘informative’ here is merely that the prior contains some information, not that that information is accurate.} As Qing and Franke (2014) points out, the model in Lassiter and Goodman (2017) has the counter-intuitive result that prior to hearing Feynman called ‘tall’, the listener does not know what ‘tall’ means when applied to an adult male, but afterwards the listener has gained knowledge about both Feynman’s height, and what ‘tall’ means; it seems more intuitive that prior to
hearing Feynman called ‘tall’, the listener has prior information about what ‘tall’ means when applied to an adult male, and the listener uses that information to infer Feynman’s height upon hearing him called ‘tall’. As noted earlier, Lassiter and Goodman (2017) justify uniform prior distributions for thresholds so as to not make ‘tall skyscraper’ mean something too short, as it would if ‘tall man’ is to come out as meaning what it should, but this problem can be avoided if we think of the distribution as a distribution not over measures of physical dimensions, but over measures over degrees of deviation from a measure of central tendency.

Next, I want to claim that if listeners make use of a genuinely informative prior threshold distribution in any Bayesian calculation leading to the pre-theoretically expected statistical inferences, then that distribution must also be accurate. To see what this means, consider ‘tall’: for ‘tall’ and heights we have an in principle lower bound, since nothing can be of negative height, but no obvious in principle upper bound on the degrees of deviation an object might be above some measure of central tendency for the comparison class, since there seems in principle no upper bound on the height of an object. For concreteness, suppose the comparison class is adult males: since the heights of adult males are roughly normally distributed with a mean of 5'9" and a standard deviation of roughly 3"., we may take the prior on heights to be a standard normal distribution $\mathcal{N}(0, 1)$ and take the lower bound on thresholds and heights to be $-23$, since it is roughly 23 standard deviations from the average height of 5'9" down to 0". When we shift the prior distribution on heights to a standard normal distribution $\mathcal{N}(0, 1)$, a plausible prior distribution for thresholds would

---

16 Is a 1000 foot tall man still a man? My intuition is that as the height reaches absurd proportions, we are less inclined to call it a man, but I think this perspective can only be adopted by those willing to accept degreeed nominals.
be approximately \( \mathcal{N}(1, .5) \), corresponding to the intuition that one would almost certainly be called ‘tall’ if one were 6’3”, and almost certainly not be called tall if one were 5’9”, with perhaps equal chances if one were 6’0”. We have thus taken the threshold distribution average to be 6’0”, with a standard deviation of half the standard deviation of the height, so that by the time one’s height was two threshold standard deviations greater than the threshold average—that is, 6’3”—one would almost certainly be called ‘tall’.

Why assume that if the listener makes use of prior information about the threshold, he makes use of accurate information about the threshold? Suppose he thinks of the threshold distribution as much lower than it is, (either due to a genuine mistake, or as some kind of presupposition or pretense of conversation); then ‘John is tall, he is 5’10”,’ should be acceptable. Or suppose he thinks of the threshold distribution as being much higher than it is; then ‘John is not tall, he is 6’1”,’ should also be acceptable. But neither is. Maybe we can say that the former is unacceptable because the posterior in heights given ‘tall’ is incompatible with him being 5’10”, and the latter because the posterior is incompatible with being 6’1”. However, cases like

\begin{enumerate}
\item[(20)] a. Bill: John is tall.
\item b. Jill: How tall is John?
\item c. Bill: He’s 5’10”.
\item d. Jill: Huh?
\item e. Bill: Anyone 5’10” is tall.
\end{enumerate}

seem to show us that an artificially low threshold prior cannot be adopted as a presupposition
or pretense for interpretation, for then Bill’s last utterance would not be surprising, but it is. Of course the fact that it is surprising also seems to tell against the claim that the oddness of the last utterance is due to its asserting something that is already presupposed. Similarly for an artificially high threshold prior:

(21) a. Bill: John is not tall.
   b. Jill: How tall is John?
   c. Bill: He’s 6’1”.
   d. Jill: Huh?
   e. Bill: No one 6’1” is tall.

Now, perhaps we might object to these examples on the grounds that the pretense or presupposition of artificially low or high threshold distributions does not project out to the latter statements from the earlier statements. On the other hand, discourses about fiction seem to provide examples where the pretense or presupposition does project across sentences and speakers; for example:

(22) a. Bill: Sherlock Holmes lived at 221b Baker Street.
   b. Jill: In what city?
   c. Bill: He lived in London.

Here the pretense that Sherlock Holmes existed remains in the last sentence.

Finally, thinking of the receiver as presupposing, or adopting as a pretense, some artificially high or artificially low prior threshold distribution seems to not fit the usual Gricean
paradigms: in the case of sarcasm or metaphor or other non-literal interpretation, the literal meaning of the words results in some obvious violation of the maxims that (perhaps) the listener thinks the speaker thinks the listener can see, so the listener reinterprets the usage in some new non-literal way in order for the speaker not to be in violation of the conversational maxims. But here, we are trying to explain what the literal interpretation is in the first place, not attempting to explain non-literal interpretation. Instead of explaining non-literal interpretation by literal interpretation flouting the maxims, we are explaining literal interpretation, and furthermore, doing this by assumption of something all parties to the conversation believe to be false. Or consider the case of presupposition or pretense: here we already have a literal interpretation and we have to assume additional information to prevent some maxim from being violated. In contrast, the model here is attempting to explain how the literal interpretation comes to be, and does not start with a fixed literal interpretation and then generate additional inferences.

4.3.2.2 A Realistic Prior Height Distribution

Next, we might ask what a plausible prior height distribution $\phi(h)$ might be. Remember, we are attempting to model with $\rho_n(h|u)$ what, for the purposes of the discourse, the receiver is disposed to act as if he believes the probability to be, given the speaker’s utterance, that the object under discussion is $h$ degrees of deviation above the average height of the comparison class, and what, roughly, for the purposes of the discourses, the speaker is disposed to act as if he believes the receiver is disposed to act as if he believes the probability to be, given the speaker’s utterance, that the object under discussion is $h$ degrees of deviation
above the average height of the relevant comparison class. Similarly, for \( \phi(h) \) we remove the stipulation ‘given the speaker’s utterance’ and specify that we are attempting to model with \( \phi(h) \) what, for the purposes of the discourse, the receiver is disposed to act as if he believes the probability to be, that the object under discussion is \( h \) degrees of deviation above the average height of the comparison class, and what, roughly, for the purposes of the discourse, the speaker is disposed to act as if he believes the receiver is disposed to act as if he believes the probability to be, that the object under discussion is \( h \) degrees of deviation above the average height of the relevant comparison class.

Now, what for the purposes of the discourse the receiver is disposed to act as if he believes the probability to be, that the object under discussion is \( h \) degrees of deviation above the average height of the comparison class, will vary from situation to situation: if we are both looking straight at Feynman, measuring 6'9" next to the height strip at the local 7-Eleven, I would normally be disposed for the purposes of the conversation to narrowly distribute Feynman’s height around 6'9" tall, and (if you were the speaker) you would normally (roughly) be disposed for the purposes of the conversation to act as if you believed that I narrowly distribute Feynman’s height around 6'9" tall. On the other hand, suppose as before in our prototypical uses of ‘tall’ that it is common knowledge (or rather, roughly, our probabilistic analogue of common knowledge) that I know no more about Richard Feynman than that he was an American adult male, when I ask you ‘Describe Richard Feynman to me. What did he look like? Was he at all like how he is commonly portrayed in the media?’ Then I would normally be disposed for the purposes of the conversation to distribute Feynman’s height around 5'9" tall, with a standard deviation of 3", and (if you were the speaker) you would
normally (roughly) be disposed for the purposes of the conversation to act is if you believed that I distribute Feynman’s height around 5’9” tall, with a standard deviation of 3”.\textsuperscript{17}

Finally, for similar reasons as with for $\theta$, I think it is clear that we cannot think of the listener as, due to some presupposition or pretense, being disposed for the purposes of the conversation to act as if he believes that Feynman’s height is anything other than in accordance with the background distribution for American adult males: an utterance of ‘Feynman is tall’ is not clearly false in this case, and thus cannot trigger some unrealistic height distribution. Furthermore, if it did trigger such a presupposition or pretense, we would be able to say things such as

(23)  
\begin{itemize}
  \item a. Bill: John is tall.
  \item b. Jill: How tall is John?
  \item c. Bill: He’s 5’10”.
  \item d. Jill: Huh?
  \item e. Bill: 5’10” is tall. The vast majority of people aren’t even 5’8”.
\end{itemize}

when clearly we cannot.

\textsuperscript{17}This situation was claimed earlier to be one in which the posterior on thresholds is identical to the prior on thresholds, but I do not want to claim that here; I only want to identify this as the kind of prototypical use of ‘tall’ that was discussed earlier.
4.3.2.3 Substantive Sending Probabilities

If in prototypical uses of ‘tall’ we thus must think of the receiver as having a realistic prior height distribution, and a realistic prior threshold distribution, we might next ask what a plausible conditional sending probability $\sigma_{n-1}(u|h, \theta)$ could be. I see two options for a plausible sending function:

First, for a given $\theta$, the sending probability function could increase gradually as $h$ increases. The center of the increase (or the inflection point in the case of a sigmoid function) might be some fixed distance from $\theta$, for all $\theta$ values; or, the center of the increase or the inflection point could become increasingly close to $\theta$ as $\theta$ decreases towards some lower bound, thus the sending curve steepens as the threshold decreases. Either way, we take the conditional sending probabilities to be generated by some more substantive process, such that for any given threshold, the probability that the speaker makes utterance $u$ increases monotonically as the height increases, but does not instantaneously transition from 0 to 1 at the threshold.

Of course, what seems implausible is that the sending probability at some given threshold is such that $\sigma_{n-1}(u|h_1, \theta) > \sigma_{n-1}(u|h_2, \theta)$ where $h_1 < h_2$, or that $\sigma_{n-1}(u|h, \theta)$ is uniform for all $h$ at any given $\theta$.\(^{18}\)

\(^{18}\)This may not hold in case of words like ‘big’, where there is always ‘huge’ which might put downward pressure on the sending probabilities of ‘big’ as the size of an object reaches the extremes of the distribution of the relevant comparison class. Thus, a shrimp 1 foot long might be more likely to be called ‘huge’ than ‘big’. One question is whether the sending probabilities for ‘big’ are static when the QUD is just the null signal and ‘big’, which might lend some explanation for the intuition that ‘big’ is just as true for a 1 foot shrimp as for a 4 inch shrimp.
Second, the sending probability function could be 0 when \( h < \theta \) and 1 when \( h \geq \theta \). Two views could lead to this kind of function: we could take the conditional sending probabilities to be either the probability that the speaker’s utterance is true given that the threshold is \( \theta \) and the height is \( h \), or the probability that the speaker would make utterance \( u \), given that the speaker believed the height was \( h \), believed the threshold was \( \theta \), and the speaker was attempting to speak truthfully.\(^{19}\) Either way, the sender model is such that for any given height and threshold, the function takes the value 0 if that height is less than that threshold, and 1 if that height is greater than or equal to that threshold.

However, I think there are serious problems for both kinds of sending function. Consider the first kind of sending function: suppose that one is 6’3” tall. Intuitively, the probability that one will be called ‘tall’ should be the probability that the threshold to be called ‘tall’ among American adult males, is less than or equal to 6’3”. Thus, it should be that \( \sigma_{n-1}(u|h) \), which in Bayesian terms is defined as:

\[
\int_0^\infty \sigma_{n-1}(u|h, \theta) \times p(\theta) \, d\theta \tag{4.11}
\]

should be equal to:

\[
\int_0^h p(\theta) \, d\theta \tag{4.12}
\]

Now, it is true that for certain \( p(\theta) \), \( \sigma_{n-1}(u|h, \theta) \) can fail to be the truth-conditions for

\(^{19}\)This formulation results in a posterior not on heights, but on the speaker’s believing any given height.

We would then need some further theory as to how that changes the receivers posterior on heights.
‘tall’, and this equality will still hold. For example, suppose $p(\theta)$ is distributed over the non-negative real line, and is some constant $c$ for $\theta < k$ and $\int_0^k p(\theta) \, d\theta = .5$, (so $c = \frac{.5}{k}$), and $p(\theta)$ asymptotically approaches 0 for $\theta \geq k$. Then if we let

$$
\sigma_{n-1}(u|h, \theta) = \begin{cases} 
1 & \text{if } h \geq \theta \\
0 & \text{else} \\
\frac{h}{k} & \text{else}
\end{cases}
$$

(4.13)

then the equality will hold. In effect, we can gerrymander a $\sigma_{n-1}(u|h, \theta)$ to make it the case.

However, the only way to define $\sigma_{n-1}(u|h, \theta)$ so as to guarantee that, for any distribution $p(\theta)$, the sending probability $\sigma_{n-1}(u|h)$ as a function of $h$ is $\int_0^h p(\theta) \, d\theta$, when $\sigma_{n-1}(u|h)$ is defined as $\int_0^\infty \sigma_{n-1}(u|h, \theta) \ast p(\theta) \, d\theta$ is to define $\sigma_{n-1}(u|h, \theta)$ as the truth-conditions.

Similarly, suppose the threshold to be called ‘tall’ among American adult males is $6'3''$. Then, the probability that a randomly selected American adult male will be called ‘tall’ should be the proportion of the population that is taller than $6'3''$ tall. Thus, for a given $\theta$, it should be that $\sigma_{n-1}(u|\theta) = 1 - \Phi(\theta)$, where $\Phi(h)$ is the cumulative height distribution. Again, we can for a given $\phi(h)$ gerrymander a $\sigma_{n-1}(u|h, \theta)$ that is not the truth-conditions but which will guarantee that $\sigma_{n-1}(u|\theta)$ will be equal to the probability that a randomly selected member of the relevant comparison class will be taller than $\theta$: that is, $1 - \Phi(\theta)$. But, the only way to
define $\sigma_{n-1}(u|h, \theta)$ so as to guarantee that, for any distribution $\phi(h)$, the sending probability $\sigma_{n-1}(u|\theta)$ as a function of $\theta$ is $1 - \Phi(\theta)$, when $\sigma_{n-1}(u|\theta)$ is defined as $\int_0^\infty \sigma_{n-1}(u|h, \theta) \ast \phi(h) \, dh$

is to define $\sigma_{n-1}(u|h, \theta)$ as the truth-conditions.

To see that this is so: w.l.o.g., assume there is some $\theta_0$ s.t. there is an $h_0 > \theta_0$ s.t. $\sigma_{n-1}(u|h_0, \theta_0) < 1$, and that for all other $h > \theta_0$, $\sigma_{n-1}(u|h, \theta_0) = 1$. Assume $\phi(h_0) > 0$. Then if $\int_0^\infty \sigma_{n-1}(u|h, \theta_0) \ast \phi(h) \, dh = 1 - \Phi(\theta_0)$ there must be some $h_1 < \theta_0$ s.t.

$\sigma_{n-1}(u|h_1, \theta_0) \ast \phi(h_1) = \phi(h_0) - \sigma_{n-1}(u|h_0, \theta_0) \ast \phi(h_0)$. (We need to make up for the amount of $\phi(h_0)$ that $\sigma_{n-1}(u|h_0, \theta_0)$ loses.) Now either $\sigma_{n-1}(u|h_1, \theta_0) = 1$ or $0 < \sigma_{n-1}(u|h_1, \theta_0) < 1$. If the latter then let $\phi^*(h)$ be just like $\phi(h)$ except $\phi^*(h_1) = 0$ and for some $h_2 > \theta_0$, $\neq h_0$, $\phi^*(h_2) = \phi(h_2) + \phi(h_1)$. Then $\int_0^\infty \sigma_{n-1}(u|h, \theta_0) \ast \phi^*(h) \, dh > \int_0^\infty \sigma_{n-1}(u|h, \theta_0) \ast \phi(h) \, dh$, and hence $\int_0^\infty \sigma_{n-1}(u|h, \theta_0) \ast \phi^*(h) \, dh > 1 - \Phi(\theta)$. If the former then let $\phi^*(h)$ be just like $\phi(h)$ except $\phi^*(h_0) = \phi(h_0) + \phi(h_1)$. Then $\int_0^\infty \sigma_{n-1}(u|h, \theta_0) \ast \phi^*(h) \, dh < \int_0^\infty \sigma_{n-1}(u|h, \theta_0) \ast \phi(h) \, dh$, and hence $\int_0^\infty \sigma_{n-1}(u|h, \theta_0) \ast \phi^*(h) \, dh < 1 - \Phi(\theta)$. Basically, if to make up what is missing you have all of something else—the former case—then the new distribution will only give you part of that thing. Else if to make up what is missing you have part of something else—the latter case—then the new distribution will only give you all of that thing. The proof that the only way to define $\sigma_{n-1}(u|h, \theta)$ so that $\int_0^\infty \sigma_{n-1}(u|h, \theta) \ast p(\theta) \, d\theta = \int_0^h p(\theta) \, d\theta$, is similar.

4.3.2.4 Sending Probabilities as Truth-(ful utterance) Conditions

To see the problem with the second kind of sending function, according to which $\sigma_{n-1}(u|h, \theta)$ is 0 if $h < \theta$, and 1 if $h \geq \theta$, consider that in the case of heights of American adult males the average is approximately 5’9”, and the standard deviation is approximately 3”, and we
assume that one would almost certainly be called ‘tall’ if one were 6′3″, and almost certainly not be called tall if one were 5′9″, with perhaps equal chances if one were 6′0″, so we are taking the threshold distribution average to be 6′0″, with a standard deviation of half the standard deviation of the height (so that by the time one’s height was two threshold standard deviations greater than the threshold average—that is, 6′3″—one would almost certainly be called ‘tall’.) We thus take the distribution of heights of American adult males to be \( N(0.,.1.) \), and the distribution of thresholds for ‘tall’ will be \( N(1.,.5) \). In this case we see the results in Figure (4.8). Although the height posterior is plausible, we immediately see a problem: the threshold posterior is implausible; a lower threshold distribution is not added to the common ground in prototypical uses of ‘tall’.

Now, it must be admitted that we do see that when it is part of the common ground that Feynman’s height is quite specifically known, the metalinguistic effect highlighted by Barker (2002) is present, as in Figure (4.9). Even better, the metalinguistic effect is weaker as the prior estimate of Feynman’s height increases: as we see in Figures (4.10) and (4.11), by the time it is part of the common ground that Feynman is two standard deviations above the
background height for American adult males—about 6'3''—virtually no new information is added to the common ground about the sending probabilities. The model also has some other nice features: suppose the common ground includes that Feynman is either 5'6'' or 6'3''; with hearer thinking he is short and the speaker thinking he is tall. If we think of the common ground as being an average of the two, then we could see how Feynman’s height comes to be agreed upon, as in Figure (4.12). Also, as we see in Figure (4.13), when it is part of the common ground that the threshold is some very specific value, then from an utterance of ‘Feynman is tall’ we learn almost nothing about the threshold value distribution, and very much about Feynman’s height.

However, all this leaves unresolved the fact that in normal, prototypical uses of ‘tall’, the model predicts that listeners infer that the distribution of thresholds is shifted distinctly downwards, when in fact listeners seem to do no such thing. And in fact, the striking success of the model in metalinguistic cases seems to indicate that this model is operative in metalinguistic cases, but some other regime is at work in prototypical uses.
Figure 4.10: middle specific height prior uncertain threshold prior

Figure 4.11: high specific height prior uncertain threshold prior

Figure 4.12: bimodal height prior plausible threshold prior
4.3.3 Negation Again

The final objection against the kind of Bayes-Grice models captured in Equation (4.5) concerns negation again, and how the listener arrives at knowledge of the threshold distribution. This problem applies to both kinds of models detailed in §4.3.2, and will remain even if we can manage to solve either of the problems associated with those kinds of models.

Recall that in the kind of models explored in Chapter 2, there is a substantive sender model generating the probability of the speaker making the utterance, and the threshold prior is an implausible uniform bounded distribution. This resulted in problems with negation, since the same processes that give rise to plausible threshold posteriors will apply in the case of negation. More precisely: the sending probability has to be (relative to pragmatic factors) a function of the truth-conditions of the utterance: given two \( h \) and \( \theta \) combinations that are equally true, the sending probability should, \textit{modulo} other pragmatic considerations like comparative informativity (which is controlled in the Lassiter & Goodman model by the QUD), cost, and speaker rational choice parameter, be the same. Then the sending proba-
bilities for ‘tall’ and ‘not tall’ will be the same for \((h, \theta)\) and \((-h, \theta)\), respectively, assuming other pragmatic factors like the cost, comparative informativity, and speaker rational choice parameter \((\lambda)\) are the same, since ‘tall’ for \((h, \theta)\) is equally true as ‘not tall’ for \((-h, \theta)\).

(We are assuming that the only thing that matters for the truth-conditions are the distance between \(h\) and \(\theta\). We are also assuming a suitably symmetric QUD that includes ‘tall’, ‘not tall’, and the null phrase, or perhaps ‘tall’, ‘short’, ‘not tall’, ‘not short’, and the null phrase; not a gerry-mandered QUD such as ‘tall’, ‘not tall’, and ‘of miniscule stature’, where the higher cost of the last phrase might push the sending probabilities for ‘not tall’ to include not just heights we would consider short, but more middling heights too.)

We then got incorrect symmetrical interpretations for ‘tall’ and ‘not tall’, since the prior \(\theta\) distribution is uniform over \(\mu \pm k \times \sigma\), and the ‘not tall’ and ‘tall’ sending probabilities are mirror-images of each other with respect to the prior \(h\) distribution, which is, in the case of ‘tall’ when applied to adult American males, approximately a lower bounded normal distribution where when converted to a standard normal distribution the lower bound is \(-23\sigma\) below 0—which seems too low for the boundedness to plausibly introduce the required asymmetry of interpretation. The additional cost of ‘not tall’ relative to ‘tall’ only pushes the sending probabilities for ‘not tall’ farther away from the central tendency of the \(h\) prior, and thus we end up with ‘not tall’ meaning something in fact stronger than even ‘short’, not to mention ‘tall’. Of course, intuitively we know that ‘not tall’ is not as informative as ‘tall’; the trouble is finding room within this category of model for this intuition. If the speaker’s receiver model relative to which comparative informativity is defined, is the literal receiver who simply conditionalizes on the truth-conditions of the utterance, then there is no room
to introduce such an asymmetry at the level of the literal receiver, and thus we must look for some other pragmatic factor that will generate the required asymmetry of interpretation between ‘tall’ and ‘not tall’ out of an asymmetry of sending probability between ‘tall’ and ‘not tall’. Note that this argument against this category of model seems even to apply to degree-of-truth theories; we could take the degree of truth to be a function of the threshold and the height, and then the sender model is either the degree of truth itself, or incorporates the degree of truth into some substantive sender model with various pragmatic variables.

Now, we might suppose we can avoid the problem of unwanted metalinguistic inferences with the second kind of sending function by simply marginalizing out \( \theta \) from the sending probabilities, before conditionalizing on \( \phi(h) \); if we have a model for generating the requisite threshold distributions, we might suppose our work to be done, since there will then be no posterior \( \theta \) distribution: this would in essence derive our model in Equation (4.6) from knowledge of the thresholds. Thus, Qing and Franke (2014) present a Bayesian model of statistical inference from vague predicates like ‘tall’ that has a sending probability conditional on only the height of the object and a particular distribution for the threshold (but not the threshold itself), and a prior distribution for the height of the object, but not for the threshold.

\[
\rho_n(h|u_1) = \sigma_{n-1}^{p(\theta)}(u_1|h) * \phi(h) \tag{4.14}
\]

The sending probabilities \( \sigma_{n-1}^{p(\theta)}(u_1|h) \) are defined in terms of a threshold distribution \( p(\theta) \); the message is sent if and only if the height is greater than or equal to the threshold, so that
the probability that a certain height will be encoded as ‘tall’ is equal to the probability that the threshold is lower than that height:

\[ \sigma_{n-1}^{p(\theta)}(u_1|h) = \int_{-\infty}^{h} p(\theta) \, d\theta \]  

(4.15)

In turn, the distribution \( p(\theta) \) is defined by the level 2 receiver’s conception of a level 1 sender of finite rationality who approximates a threshold of maximal utility:

\[ p(\theta) = e^{\lambda^* U(\theta)} \]  

(4.16)

The utility of a threshold is defined in terms of its expected success \( ES(\theta) \) for a level 0 literal receiver, minus the expected cost of the utterance, (note that when \( h < \theta \) the message is not sent and thus no cost is incurred):

\[ U(\theta) = ES(\theta) - \int_{\theta}^{\infty} \phi(h) * c \, d\theta \]  

(4.17)

Finally, the expected success of a threshold \( ES(\theta) \) for a level 0 literal receiver is defined as the sum of the expected true negative decoding rate (the probability that the level 0 receiver decodes the null signal as height \( h \), times the probability that it was in fact height \( h \) that was encoded by the null signal), and the expected true positive decoding rate (the probability that the level 0 receiver decodes ‘tall’ as height \( h \), times the probability that it was height \( h \) that was encoded by ‘tall’):
Here, as with Lassiter and Goodman (2017), the level 0 literal listener decodes the null signal as the background distribution, and ‘tall’ by renormalizing above the threshold:

\[
ρ_0(h|u_0, θ) = φ(h)
\]

(4.19)

\[
ρ_0(h|u_1, θ) = \begin{cases} 
φ(h) / \int_0^∞ φ(h) \, dh & \text{if } h ≥ θ \\
0 & \text{otherwise}
\end{cases}
\]

(4.20)

Qing and Franke (2014) demonstrate plausible posterior height distributions \(ρ_2(h|u_1)\) for \(λ = 4\) and \(C(u_1) = 2\).

However, just as with Lassiter and Goodman (2017)—and I think this is illustrative of a larger phenomenon—this model again runs into problems with negation: ‘Feynman is not tall’ will be evaluated either relative to a threshold \(θ_{\text{tall}}\) or, if we prefer partially defined predicates in the style of Soames et al. (1999), a threshold \(θ_{\text{not tall}}\). In either case, the same cognitive processes outlined in Equations (4.14)-(4.19) will apply, when we define the level 0 literal receiver’s interpretation of ‘not tall’ as:

\[
ρ_0(h|¬u_1, θ) = \begin{cases} 
φ(h) / \int_0^θ φ(h) \, dh & \text{if } h < θ \\
0 & \text{otherwise}
\end{cases}
\]

(4.21)
The resulting threshold distribution will be the opposite of that resulting from ‘Feynman is tall’, but more extreme for the same cost function and rationality constant \( \lambda \) because of the extra cost of ‘not’, and then the sending curve and posterior height distribution \( \rho_{n-1}(h|\text{not tall}) \) will also be opposite of ‘tall’, but only more extreme, and thus ‘not tall’ will be interpreted as akin to ‘very short’. (Furthermore, in the former case, we will have two threshold distributions for ‘tall’, one resulting from ‘Feynman is tall’, and another from ‘Feynman is not tall’.)

Finally, I think it is important to note that the foregoing problem with negation still applies even if we find a way to solve the problem of unexpected threshold posteriors with the second kind of model, and even if we can find a way to solve the problem of ‘non-threshold’ behavior with the first kind of model. It occurred with both the original version of Lassiter and Goodman (2017) and with our modified version in Chapter 2. We might even define a distance-weighted version of Qing and Franke (2014) such that:

\[
ES(\theta) = \int_{-\infty}^{\theta} \phi(h) \ast \left( \int_{-\infty}^{\infty} f_h(x) * \rho_0(h|u_0, \theta) \, dx \right) \, dh + \\
\int_{\theta}^{\infty} \phi(h) \ast \left( \int_{\theta}^{\infty} f_h(x) * \rho_0(h|u_1, \theta) \, dx \right) \, dh \tag{4.22}
\]

and as before:

\[
f_h(x) = \frac{1}{\sqrt{2\pi\sigma_w^2}} \times e^{-(x-h)^2 / 2\sigma_w^2} \tag{4.23}
\]
where $\sigma_w$ is a weighting constant that tells us ‘how close’ we care about being to $h$. Note that the lower limit on the inner integral of the right summand is $\theta$ since $\rho_0(h|u_1, \theta) = 0$ for $h < \theta$. Again, we can expect the same problem with negation to occur, since if this generates plausible threshold distributions for ‘Feynman is tall’ it will generate opposite but equal threshold distributions for ‘Feynman is not tall’.

This kind of problem with negation thus seems generalizable: deriving plausible posteriors on heights from ‘Feynman is tall’ in a Bayesian framework requires that $\sigma_{n-1}(u|h)$ be roughly sigmoid, with the inflection point about 1 standard deviation to the right of the average of the prior height distribution. Lassiter and Goodman (2017) accomplish this via a substantive sender model $\sigma_{n-1}(u|h, \theta)$ such that when we conditionalize on a bounded uniform prior $p(\theta)$ and then marginalize $\theta$ out, we get the desired sender behavior. Here the key notion is the informativity of ‘tall’, compared to other words, about a height $h$, relative to a given threshold $\theta$. Qing and Franke (2014) accomplish this by thinking of the sender as calling the object ‘tall’ when and only when the threshold is less than the height, and then generating a distribution for the threshold by assuming the sender approximates an optimal threshold. Either way, we see that if we start with the threshold as a free variable relative to which the sender’s behavior must be defined and then marginalize out that threshold, we run into the problem that the same cognitive processes invoked to define that behavior—whether that process acts on the sending probabilities $\sigma_{n-1}(u|h, \theta)$ or $p(\theta)$—will result in equal but opposite behavior in the case of the negated form. We could even combine a substantive sender model with a non-uniform prior for $\theta$, and this would still apply, since the cognitive processes would be equal and opposite for both $\sigma_{n-1}(u|h, \theta)$ and $p(\theta)$. The problem, it
seems, is that we start with $\theta$ as a free variable. Why can’t we think of $p(\theta)$ as being fixed for all contexts, so that in any context even ‘not tall’ gets interpreted relative to the proper threshold distribution for ‘tall’? Once a threshold distribution is fixed relative to any context, it becomes part of the meaning of the term, just as when we specify that the denotation of ‘Feynman’ is invariant relative to context.

4.4 A Bayes-Grice Model Without Truth-Conditions

4.4.1 Squaring the Circle: Descriptive Usages

Let’s review the evidence: We have seen that unless we allow the threshold distribution to remain fixed invariant of context, we end up with implausible statistical inferences from negations of vague predicates once we invoke some cognitive process to generate the requisite threshold distribution. This suggests that we should indeed let $p(\theta)$ be fixed from context to context, if we think of the degree scale as a scale of degrees of deviation from a measure of central tendency for the relevant comparison class. One way for the threshold distribution to be fixed invariant of context is for it to be part of the character of ‘tall’, since the character is fixed invariant of context. Now, character is a function from worlds to propositions (or intensions more generally); so analogously if the threshold distribution is straightforwardly identified with the character we would expect the character to be a function from thresholds to probabilities of thresholds. But of course that gets the subject matter of vague predicates wrong: they are supposed to tell us about heights, not thresholds. Perhaps thinking of the character as a rule for determining the proposition expressed in any given context will shed
light on the matter, and that rule should again be a function; we will then want to think of
the threshold distribution as part of the definition of that function (that is, the character)
expressed by ‘tall’. Intuitively, we want that function to tell us how tall something is; the
clear way a threshold distribution can do that, is to fix the probability that something will
be called ‘tall’ as its height varies. This suggests that the other component is what we have
identified thus far as the truth-conditions of ‘Feynman is tall’: \( \sigma_{n-1}(u|h, \theta) \).

However, as we have seen, in purely descriptive uses of vague predicates, the thresholds ap-
pear to be unnecessary to any Bayesian model of how information about the height of the
object under discussion is added to the probabilistic common ground, and, if we assume real-
istic prior distributions for the threshold and height in prototypical uses of vague predicates
like ‘tall’, we end up with unrealistic posterior threshold distributions in any Bayesian model
in which the sender component \( \sigma(u|h, \theta) \) reflects the truth-conditions. This suggests that
neither the threshold distribution, nor the truth-conditions, are part of any Bayesian model
of how listeners make the expected statistical inferences. How can we square this circle? I
want to suggest that we can do so by inverting the order of explanation of the relationship
between our knowledge of the threshold distribution and our knowledge of how the sending
probabilities depend on the height of the object. Along with this inversion, I want suggest
that the meaning of ‘tall’ is fundamentally probabilistic. Thus, I suggest that we take the
model in Equation (4.6)

\[
\rho_n(h|u) = \frac{\sigma_{n-1}(u|h) \times \phi(h)}{\sigma_{n-1}(u)},
\] (4.6)
to capture how we infer how tall something is, when we hear it called ‘tall’. Since we want that inferential process to depend on the meaning of ‘tall’, I propose that we take the meaning of ‘Feynman is tall’ to be a function from heights to encoding probabilities. This suggests that the denotation is something like:

$$[\text{Feynman is tall}] = \sigma_{n-1}(\text{Feynman is tall} \mid h(\text{Feynman}))$$

(4.24)

where \(h(x)\) is a function from entities to degrees of deviation from some measure of central tendency for the height scale, (and not simply heights.) Note that in light of the divorce of the degree scales from actual physical quantities, necessitated in order to account for why ‘completely fast’ and ‘perfectly cold’ are unacceptable even though velocity is upper bounded by the speed of light and temperature is lower bounded at 0 Kelvin, we ought to maintain some theoretical distance between the degree scale for ‘tall’, and actual physical distances. Nonetheless, even if we should not think of the height scale as actual physical distance, it seems we ought to think of it as being quite close to the physical distance, as in this case the relation between the degree scale and the physical quantity seems close enough to afford the statistical inferences regarding physical height that we have thus far been discussing. In general, it seems that the interloper between our words and the actual physical quantities themselves, would be concepts: thus our concept of speed or temperature when our words evolved lacked an upper, or lower, bound, and thus we see the patterns of adverbial modification that we do.

As for the threshold distribution, we can let it be defined in terms of these encoding proba-
Thus, instead of coming to know how the encoding probabilities vary with (the degree of deviation from some measure of central tendency of) height as a result of our knowledge of the threshold distribution, I suggest that we know what the threshold distribution is like, based on our knowledge of how the encoding probabilities vary with (the degree of deviation from some measure of central tendency of) height.

Now, Equation (4.24) defines the output of $J \cdot K$ in terms of the input to that function, and this is unacceptable in a definition of the static part of the model: the denotation is to be applied to an existing probability distribution to yield an updated probability distribution; the effect is to be dynamic, but the denotation is to be static. Note that this is not a problem facing Barker (2002), since he gives a single denotation for purely descriptive, purely metalinguistic, and mixed descriptive and metalinguistic usages; but it is a problem here, where we are attempting to give a static denotation to ‘tall’, to reflect the fact that no metalinguistic inferences are made in purely descriptive uses of the word.

Thus, instead of Equation (4.24), we might take the meaning of ‘tall’ to be, roughly:
\[ \text{[Feynman is tall]} = \text{a monotonically increasing probability function on heights with a derivative that is a roughly normal distribution } \mathcal{N}(\mu^*, \sigma^*), \text{ where } \\
\mu^* = \mu + \sigma \text{ and } \sigma^* = .5\sigma, \text{ where } \mathcal{N}(\mu, \sigma) \]

are the average and standard deviation on the height scale of the members of the relevant comparison class.

One possible objection to this proposed denotation for ‘Feynman is tall’ is that it may not result in the intuitively correct posterior distribution, given a prior distribution that is not roughly normally shaped. For example, if the prior distribution over the height scale is a bounded uniform distribution, or a bounded distribution with significant probability mass towards the lower bound such as in a Beta(9, 1) distribution, or a bounded distribution with both lower and upper bounds as in a Beta(.5, .5) distribution, the resulting posterior distribution may fail to match any intuitively plausible posterior distribution over the degree scale for height, which (intuitively, at least) should be roughly normally distributed.

More generally, if we can find a relative gradable adjective such that for some relevant comparison class the posterior distribution over the degree scale is not remotely close to the result of applying the probability function specified in the foregoing denotation to the prior distribution over that degree scale, that might be a counter-example to the proposed semantics. Thus, for example, if the prior distribution is a uniform distribution over (0, 5), and the intuitively plausible posterior distribution is a uniform distribution over (1, 6), then
since the latter distribution cannot result from the application of the denotation specified in Equation (4.26) to the former distribution, as required in Equation (4.6), it would constitute a possible counter-example to the view we are considering. In response, I want to note first that it seems a virtue of the approach that it generates falsifying predictions. Second, although this does not demonstrate that no such cases exist, I have trouble finding them myself: words like small, big, short, cheap, and expensive all seem to have the requisite relation between prior and posterior.

### 4.4.2 Squaring the Circle: Metalinguistic Usages

Of course, this cannot be the whole story, because this model allows for no metalinguistic inferences to be drawn from an utterance of ‘Feynman is tall.’, since there are no threshold variables in the prior or posterior, and clearly we do sometimes gain information about the threshold distribution (and thus the sending probabilities) from such an utterance. Nonetheless, I think the solution is clear: We assume the independence of $h$ and $\theta$, and then ask ourselves how we can define a function $\sigma_{n-1}(u|h, \theta)$ that will allow us to learn about the threshold distribution from an utterance of ‘tall’; our function must conditionalize on $h$, since clearly it is knowledge of Feynman’s height that allows us to gain knowledge of the threshold, and it must conditionalize on $\theta$, in order for $\theta$ to appear in the posterior. If we assume $\theta$ acts like a true threshold and thus define $\sigma_{n-1}(u|h, \theta)$ as:
\[ \sigma_{n-1}(u|h, \theta) = \begin{cases} 
1 & \text{if } h \geq \theta \\
0 & \text{else} 
\end{cases} \quad (4.27) \]

and further assume, as before, Equation (4.5), reprinted here:

\[ \rho_n(h, \theta|u) = \frac{\sigma_{n-1}(u|h, \theta) \ast p(\theta) \ast \phi(h)}{\sigma_{n-1}(u)} \quad (4.5) \]

then we can say that

\[
\rho_n(h|u) = \int_{0}^{\infty} \rho_n(h, \theta|u) \, d\theta \\
= \int_{0}^{\infty} \sigma_{n-1}(u|h, \theta) \ast p(\theta) \ast \phi(h) \, d\theta \\
= \phi(h) \ast \int_{0}^{h} \sigma_{n-1}(u|h, \theta) \ast p(\theta) \, d\theta \\
= \phi(h) \ast \int_{0}^{h} p(\theta) \, d\theta \\
= \phi(h) \ast \sigma_{n-1}(u|h) 
\]

thus getting back our original equation. We thus have derived from our simpler equation (4.6) without thresholds an equation (4.5) with thresholds, which is equivalent to our original model in its statistical inferences regarding height, but which allows us to gain new information about the threshold distribution, once we marginalize out the height distribution from the posterior:
\[ \rho_n(\theta|u) = \int_0^{\infty} \rho_n(h, \theta|u) \, dh \]
\[ = \int_{\theta}^{\infty} \sigma_{n-1}(u|h, \theta) \ast p(\theta) \ast \phi(h) \, dh \]

To be clear: the view I am advocating here is that there are at bottom no thresholds; even though it seems we have given back what we had earlier taken away, our threshold distribution is derived from the receiver’s view of the sending probabilities, and the receiver requires no additional information that she did not have before in order to define her probability distribution for thresholds: she simply takes the derivative of her view of the sending probabilities.

A second model, similar to this first model, is that we marginalize out the heights in metalinguistic usages prior to conditionalization on our existing threshold distribution:

\[ \rho_n(\theta|u) = \sigma_{n-1}(u|\theta) \ast p(\theta) \]
\[ = p(\theta) \ast \int_{\theta}^{\infty} \sigma_{n-1}(u|h, \theta) \ast \phi(h) \, dh \]

This helps address a concern with the first model that we see in Figure (4.9): typically, when we use someone’s utterance of ‘Feynman is tall’ to infer how tall one must be to count as tall, we do not change our distribution for Feynman’s height at all; instead, we hold it fixed and change our threshold distribution. The slight shift in posterior height distribution we see in Figure (4.9) can thus be eliminated, and not merely written off as so small as to be undetectable, in this second model, since there is no height prior, and hence no height posterior, at all. (And thus, any objection to the argument in §4.3.2.4 that the unexpected
threshold posterior in purely descriptive usages that we see in Figure (4.8) might also be simply written off, can be rejected without falling prey to a *tu quoque* response.)

One apparent advantage of this model is that what we consider as the truth-conditions of the utterance ‘Feynman is tall’ naturally fall out of the model, as a device for learning about just how tall someone has to be, to count as ‘tall’. Note also that a similar definition of ‘short’ will also be available for \( \sigma_{n-1}(\text{short}|h, \theta) \) just as for \( \sigma_{n-1}(\text{tall}|h, \theta) \):

\[
\sigma_{n-1}(\text{short}|h, \theta) = \begin{cases} 
1 & \text{if } h \leq \theta \\
0 & \text{else} 
\end{cases}
\]  

(Note however that in the corresponding proof of equivalence of the simple Bayesian model and the complex Bayesian model with regards to statistical inferences regarding heights, we will have to integrate from \( h \) up to \( \infty \), instead of from the lower bound (or \( -\infty \)) up to \( h \).)

Although it might look rather abstruse for the receiver to define these conditional sending probabilities in these different manners, we might not think it so: the receiver already has a notion of how the sending probabilities change with height, so in a sense defining the conditional sending probabilities in this manner is what one knows is required.

Finally, we can see that thinking of the denotation of ‘tall’ as a function from degrees of deviation from some measure of central tendency of height, to a probability of encoding as ‘tall’ is that we get a plausible solution to our problem with negation, without changing our intuitive notion of negation. If we define:
then we can see that our sending function will be the mirror image, flipped vertically, of the purple line in Figure (4.8), and the resulting posterior on $h$ from our simple Bayesian model will have the upper reaches of $\phi(h)$ reduced, while smoothly redistributed over the middle and lower portions of the distribution. Thus, ‘not tall’ will not mean ‘short’, but just ‘not tall’.

4.4.3 A More Formal Semantics

Now, the fact that what we traditionally consider as the truth-conditions of ‘Feynman is tall’ falls out of the model of metalinguistic interpretation above might make us suspect that we can integrate the underlying probabilistic picture here, with the existing truth-theoretic framework. I think this can be done, and I would like to sketch in this section how to do so. Consider that we suggested earlier taking the meaning of ‘$x$ is tall’ to be a function from degrees for $x$ on the scale for tallness, to the probability that a speaker would utter, ‘$x$ is tall’ given that $x$ is that tall:

$$[x \text{ is tall}] = \sigma_n(‘x \text{ is tall’} | \text{tall}(x))$$

(4.33)

We also noticed that what we might think of as the truth-conditions for ‘Feynman is tall’, reprinted here, fall out of the model when we engage in metalinguistic interpretation:
However, we might also notice this seems very close to a degree-theoretic denotation for the adjective ‘tall’, which can combine with degree morphemes such as ‘-er’ and ‘-est’ to generate the appropriate meanings for ‘taller’ and ‘tallest’:

\[
\llbracket \text{AP tall} \rrbracket = \lambda \theta \lambda x. \text{tall}(x) \geq \theta
\]

(4.34)

where \( \text{tall}(x) \) is a function of type \( D_{(e,d)} \) that maps any object \( x \) to its maximal degree of height \( h \). This is slightly different from the denotation defined in Chapter 1, as we have replaced a measure function with a relation between individuals and degrees of tallness; nonetheless, as we can see below, it will still generate the required meanings.

Now, the parallel between Equations (4.27) and (4.34) suggests that in place of truth-values, we will have probabilities of utterance. Thus, instead of a domain of truth-values \( D_t = \{0, 1\} \), we will have a domain of probability densities \( D_p = [0, 1] \). And this suggests that we can define

\[
\llbracket \text{AP tall} \rrbracket = \lambda x \lambda h_d. \text{TALL}(x)(h)
\]

(4.35)

where \( \text{TALL} \) is a function that specifies the probability that an object \( x \) of degree \( h \) on the height scale will be called ‘tall’. Then, we might use a type-shifter as in Jacobson (1999) to
reverse the order of the arguments:

\[
[rev] = \lambda g_{(e,dp)} \lambda h_d \lambda x_e \cdot g(h)(x) \tag{4.36a}
\]

\[
[rev \text{ tall}] = \lambda h_d \lambda x_e. \text{TALL}(x)(h) \tag{4.36b}
\]

Then, if we think of the height \( h' \) that \( x \) might be as itself a feature of context, and taking inspiration from Montague (1973), we can define for any given height \( h' \):

\[
[\lor \ rev \text{ tall}]^{h'} = \lambda x_e. \text{TALL}_{h'}(x) \tag{4.37}
\]

Note that since the effect of \( \lor \) is to evaluate \([rev \text{ tall}]\) at \( h' \), this has the effect of creating an \( h' \)-specific function that outputs the probability that an object \( x \) that is \( h' \) tall, will be called ‘tall’.

Now, in §4.2 I claimed that in metalinguistic interpretations, listeners introduce a threshold variable to learn about how senders use ‘tall’. In semantic terms, I propose that similar to \([\lor \ rev \text{ tall}]^{h'}\), there is for every adjective like ‘tall’ a word ‘tall*’ such that:

\[
[\text{tall*}]^\theta = \lambda x_e. \text{TALL}_\theta^*(x) \tag{4.38}
\]

where \( \lambda x_e. \text{TALL}_\theta^* \in D_{(e,p)} \). Intuitively, we can think of \( \lambda x_e. \text{TALL}_\theta^*(x) \) as telling us the probability that an object \( x \) will be called ‘tall’ if it is taller than the threshold \( \theta \) to be called
‘tall’. More formally, we can define $\theta$ and its distribution $p(\theta)$ as:

$$p(\theta) = \frac{d}{dh} \text{TALL}(x)(h) \quad (4.39)$$

Finally, we can define

$$\lambda\theta[\text{tall}^*]^\theta = \lambda\theta\lambda x.e.\text{TALL}^*(\theta)(x) \quad (4.40)$$

where

$$\lambda\theta\lambda x.e.\text{TALL}^* = \begin{cases} 1 & \text{if height of } x \geq \theta \\ 0 & \text{else} \end{cases} \quad (4.41)$$

Note that $\lambda\theta[\text{tall}^*]^\theta$ stands to $[\text{tall}^*]^\theta$ as $[\text{rev tall}]$ stands to $[\lor \text{rev tall}]^h'$. Finally, we can use $\lambda\theta[\text{tall}^*]^\theta$ to generate the appropriate meanings for ‘taller’. If we say:

$$[\ [\text{deg -er } \ ]] = \lambda g_{(d.ep)} \lambda y.e. \lambda x.e. \exists \theta [g(\theta)(x) \& \neg g(\theta)(y)] \quad (4.42)$$

then
Thus, we have derived in intuitively correct semantics for ‘x is taller than y’ from our probabilistic picture. Similar treatments ought to be available for degree constructions such as ‘Feynman is six feet tall’, ‘Feynman is the tallest man in the room’, and ‘Feynman is as tall as Einstein.’

While a full treatment of the syntactic issues involved lies beyond the scope of this dissertation, we might note that the foregoing approach does not fall prey to a common objection to the degree-theoretic treatment of vagueness: that the semantic complexity of the degree-theoretic treatment of the unmarked, positive form in both predicative (‘is tall’) and attributive (‘is a tall door’) positions belies its apparent morphological simplicity. Since we do not have to posit an unpronounced positive null morpheme ‘pos-’, and we likewise derive the adjectival form required for composition with degree morphemes from an underlying probabilistic form, we can have morphological simplicity track with semantic simplicity.

### 4.4.4 Defining the Transition

Finally, this view needs to specify under what conditions listeners change their mode of interpretation from the simple Bayesian model in Equation (4.6) to the complex Bayesian model
in Equation (4.5). It does not seem to be a sufficient condition for metalinguistic interpretation merely that the listener have some fairly narrow distribution for Feynman’s height, or even that the listener be disposed for the purposes of the discourse to act as if he has some such distribution for Feynman’s height. Suppose that for the purposes of the discourse I am disposed to act as if I believe that Feynman’s height is narrowly distributed around 5′9″ tall. Even so, if for the purposes of the discourse I am disposed to act as if I believe that you believe that I believe that Feynman’s height is in accordance with the background distribution for American adult males, I might not engage in the metalinguistic interpretation—for example, if I believe that you believe that I don’t know how tall Feynman is, and have only background information about his height, I will only engage in the descriptive interpretation and reason according to the simpler Bayesian model. This case seems to indicate that at the least, in order for the listener to engage in the metalinguistic interpretation, it must roughly be the case that the listener is disposed for the purposes of the discourse to act as if he believes that the sender believes that the listener believes that Feynman’s height is narrowly distributed around some specific height. And it seems it must be at least a fairly narrow height distribution, since otherwise the listener will think the sender is trying to tell him about Feynman’s height, not the threshold distribution. And it might be required not just that the listener believes that the sender believes that the listener believes that Feynman’s height is narrowly distributed around some specific height; it might be required that the listener believes that the sender believes that some fairly narrow distribution for Feynman’s height is part of the common ground.

In order for me to interpret you in a metalinguistic manner, is it necessary that for the
purposes of the discourse I am disposed to act as if I believe that Feynman’s height is narrowly distributed around some fairly specific height? I don’t think so—suppose you and I are at the local 7-Eleven and we see our friend Feynman walking around. Suppose I am disposed to act for the purposes of the discourse as if I believe that you believe it is part of the common ground that Feynman is 6'6". I see you see him standing next to the height strip by the door, and I see you see me see him standing there, and I see you see me see you see him standing there. Unbeknownst to you, I think I see him wearing stilts, but of a height that is unknown to me, and I don’t think you caught those stilts; given the chance, I’d have told you about them to let you know that we really can’t tell how tall Feynman is just from where his head lines up with the height strip. Thus, I am disposed for the purposes of the discourse to act as if I have only background knowledge of the Feynman’s height. Nonetheless, it seems in this case at least that so long I am disposed for the purposes of the discourse to act as if I believe that you believe it to be part of the common ground that I do not know what counts as ‘tall’ around here—suppose I have just asked you, ‘What counts as tall around here?’—I will interpret your utterance metalinguistically.

This suggests that the individually necessary and jointly sufficient conditions for me to (be disposed for the purposes of the discourse to act as if I) engage in the metalinguistic interpretation of your utterance are that:

1. I am disposed for the purposes of the conversation to act as if I believe that you believe that some fairly narrow distribution for Feynman’s height is part of the common ground.

2. I am disposed for the purposes of the conversation to act as if I believe that you believe
that it is part of the common ground that I do not know what counts as ‘tall’ around here.

Admittedly, we have not proved that these constitute necessary and sufficient conditions for metalinguistic interpretation, but they are at least initially plausible.
Chapter 5

Conclusion

5.1 How we got here

I argued in Chapter 2 that one prominent model from Lassiter and Goodman (2017) of how we get from our knowledge of the meaning of ‘Feynman is tall’ to the kind of statistical inferences that we typically engage in upon hearing that sentence, when coupled with the kind of truth-conditional semantics for vague predicates discussed in Chapter 1, leads to the wrong predictions about the kinds of statistical inferences we typically engage in upon hearing the negations of that sentence. After searching for solutions within the model, in various extensions of the model, or in various modifications of the model, I argued that the intuitively correct statistical inferences from ‘Feynman is not tall’ can be predicted by the model if we assume that listeners have realistic, informative prior distributions for the thresholds relative to which ‘Feynman is tall’ is true or false. I in turn argued that this suggests a simpler model according to which listeners use their knowledge of how the probability of the speaker uttering that sentence varies with Feynman’s height, in combination with a prior probability distribution for Feynman’s height, to derive a posterior probability distribution
for Feynman’s height. Absent a variable for thresholds in the model, relative to which truth-conditions for sentences containing vague predicates must be defined, and assuming the requisite statistical inferences are driven by knowledge of the meaning of the utterance, it appears that this is evidence for a non-truth conditional view of meaning.

Now, one might respond that our knowledge of the truth-conditions, including their dependence on the threshold for tallness, allows us to know how the encoding probabilities vary as the height varies: once we conditionalize on the threshold distribution and marginalize out the thresholds, before conditionalizing on the height distribution, we arrive at a function from heights to encoding probabilities. After conditionalizing on our prior height distribution, we arrive at a posterior height distribution. However, this still requires a prior threshold distribution that is fixed from context to context as part of the meaning of ‘tall’, unless we posit a pragmatic process as in Qing and Franke (2014) that generates a threshold distribution based on what the receiver takes to be common knowledge of the truth-conditions of our words, along with a characterization of the speaker’s relevant psychological features, such as their degree of rationality, and their subjective cost of talking. However, that option again makes the wrong predictions about the statistical inferences we make upon hearing ‘Feynman is not tall’: with the threshold for tallness again determined by a cognitive process that accounts for the truth conditions of the utterance and the relevant features of the speaker’s psychology, the receiver will think that ‘not tall’ means something akin to ‘very short’.

Of course, it would seem to be a small circle if we infer a threshold distribution for ‘tall’ from our observations of how the encoding behavior of our co-linguals varies with the heights of the things they are describing, and the explanation of how the encoding behavior of
our co-linguals varies with the heights of the things they are describing must be generated from their knowledge of the truth-conditions of utterance encoding those heights. Thus, I argued in Chapter 3 that there is a plausible account, originating in a game-theoretic framework, for why natural languages have the kind of probabilistic encoding behavior that seems characteristic of vagueness. Even though such encoding behavior is not part of a strict Nash equilibrium, nor is it an evolutionarily stable strategy on either the replicator dynamics interpretation of evolutionary game theory, or the best response dynamics interpretation of evolutionary game theory, it can be predicted to arise as a result of a learning strategy known as stimulus generalization: here, after a given iteration of the game, senders in a signaling game reinforce not just on the state that the signal encoded but also on states nearby that state; similarly, receivers reinforce not just on the state that the signal was decoded as, but also on states nearby that state.

Finally, in Chapter 4 I attempted to define a larger class of models, one member of which was discussed in Chapter 2, for how receivers move from their knowledge of the truth-conditions of utterances containing vague predicates, to the kinds of statistical inferences that we typically make when we hear a speaker say, for example, ‘Feynman is tall.’ In order to define this class of models, I demonstrated that the dynamic semantics for vague predicates from Barker (2002) is equivalent to context updating via diagonalization from R. C. Stalnaker (1978), which is again equivalent to the \(^v\)-operator from Montague (1973), all of which are special cases of the following equation capturing a Bayesian update procedure inspired by Gricean considerations:
\[
\rho_n(h, \theta|u) = \frac{\sigma_{n-1}(u|h, \theta) \ast Pr(\theta) \ast \phi(h)}{\sigma_{n-1}(u)}
\]

I take this equation to characterize the general class of models that I am concerned to reject, and I argued that conformation to this equation is in fact a requirement for any plausible Bayesian model of how, given utterances containing vague predicates, listeners draw the pre-theoretically expected statistical inference patterns from beliefs concerning the truth-conditions of such utterances. After presenting three arguments against any model conforming to this equation, I proposed an alternative model of the manner in which utterances of sentences with vague predicates give rise to the expected statistical inference patterns, of both metalinguistic and non-metalinguistic kinds: In non-metalinguistic usages of vague predicates, receivers generate a posterior probability distribution for heights by conditionalizing on a prior height distribution, and considering how the probability of the utterance varies with the height of the object:

\[
\rho_n(h|u) = \frac{\sigma_{n-1}(u|h) \times \phi(h)}{\sigma_{n-1}(u)}
\]

On the other hand, in metalinguistic usages, receivers define an encoding probability conditional on both the height and the threshold, which is thus recognizable as the truth-conditions for ‘Feynman is tall’, and then define a distribution for the threshold as the derivative of how the encoding probabilities vary with Feynman’s height. By then conditionalizing on a prior height distribution, perhaps marginalizing over heights, and then conditionalizing on this threshold distribution, listeners can calculate a posterior threshold distribution, and thus
learn how tall someone must be, to count as ‘tall’. Finally, I argued that on the assumption that knowledge of the meaning of ‘Feynman is tall’ is required for generating the expected statistical inference patterns in non-metalinguistic usages, we might think of the fundamental meaning of ‘Feynman is tall’, as probabilistic in nature: its meaning, on this view, is an encoding probability function over various heights for Feynman. Analogous to the distinction between content and character in a traditional truth-theoretic semantics, we might then also define another, derivative meaning: this being an encoding probability function over various heights for Feynman and thresholds for tallness. Finally, given the similarity of this derivative meaning to an existing degree-theoretic, truth-conditional semantics, we saw how we can generate the insights of such a semantics from a fundamentally probabilistic semantics.

5.2 Areas for Further Exploration

If the foregoing conception of the semantics for vague predicates is correct, it suggests a number of areas for further exploration. I would like to first briefly discuss a few such areas, and then end by discussing two areas at greater length: the distinction between absolute and relative gradable adjectives, and of course, the Sorites.

5.2.1 Various Extensions

First, we might attempt to determine if the view escapes the problem of higher-order vagueness: in addition to the vagueness of ‘tall’, there is also the vagueness of ‘definitely tall’: even
though someone 5'11" might be tall, it seems clear they are not definitely tall. Someone 7'0"
seems definitely tall. Between these again lie borderline cases, and a probabilistic semantics
for ‘definitely’ and ‘tall’ should, one would hope, resolve this as well as it resolves first-order
vagueness.

Second, vagueness is not restricted to adjectives: there are vague nominals, such as ‘heap’,
and perhaps even cases of vague identity, such as the ship of Theseus. We might hope that
the probabilistic view advocated here could be extended to these cases too.

Third, in contrast to the truth-conditional semantics introduced in Chapter 1, we have left
unaddressed so far the effect of comparison classes on our probabilistic semantics: to be told
‘Feynman is tall’, ‘Feynman is a tall man’, and ‘Feynman is tall for an American’ all drive
different patterns of statistical inference about Feynman’s height, and a good theory of the
semantics of vague predicates should explain how that happens. This is especially pressing
insofar as I have defined the degrees of deviation, as degrees of deviation from some measure
of central tendency, for the relevant comparison class.

Fourth, and related to the third area for further exploration, is providing a syntax for vague
predicates: whether comparison classes are arguments to the function denoted by ‘tall’, or
if they are domain restrictors, may affect what kind of syntactic structures relative gradable
adjectives can stand in.

Fifth, our semantics seems in some respects a probabilistic generalization of two-dimensional
semantics. However, the precise relation between the two frameworks is not entirely clear.
For example, certain operations which can be easily defined in the probabilistic framework,
such as marginalization before conditionalization, are not as clearly definable in the tra-
ditional two-dimensional framework. While we might view this as a point in favor of the probabilistic view proposed here, insofar as the availability of the operation allows us to respect the intuition that one can learn about the language if one knows the features of the world, or one can learn about the features of the world if one knows the language, it remains to be seen what other operations definable in one framework, will or will not find an analogue in the other framework. Furthermore, constraints on what kinds of operators are available in the two-dimensional framework—specifically, what Kaplan (1989) called ‘monsters’—may or may not be definable, or even well-motivated, in a probabilistic generalization.

Sixth, and related to the fifth, is finding a probabilistic analogue of Stalnaker’s constraints on assertion. This is clearly related to the project of defining the transition between descriptive and metalinguistic interpretations: since the projection of the diagonal proposition into the horizontal is motivated by those constraints, we should also be able find analogous probabilistic constraints on interpretation that motivate the shift from our probabilistic descriptive interpretation to probabilistic metalinguistic interpretation.

Seventh, it should be noted that we have passed over an option for a non-Bayesian, non-truth conditional view of the meaning of vague predicates: We could take the meaning of ‘Feynman is tall’ to simply be the function from a prior height distribution for the relevant comparison class, to the height distribution for the relevant comparison class that is generated by our signaling game. Since the receiver in these games is simply probabilistically decoding the utterance as various heights for Feynman in proportion to the rewards garnered over previous iterations of the game, they are clearly not engaged in a Bayesian reasoning process. Now, one problem with such as view is that it is unclear how we can think of the negation
of such a meaning; in contrast, thinking of the negation of a conditional encoding probability as one minus that probability, is simply a generalization of the traditional notion of negation. We could claim perhaps that the positive form is interpreted by the non-Bayesian process resulting from the signaling game, and the negation is interpreted by the Bayesian process involving the speaker’s encoding probabilities. However, this seems ad hoc. On the other hand, we might notice that the same underlying structure that generates the decoding probabilities, also generates the encoding probabilities: decoding of the un-negated form is just normalization across states for a fixed signal, and decoding the negated form as 1 - the encoding probabilities is at least not an ascension of the sender-receiver hierarchy: it is just normalization across utterances, followed by normalization across states for the given utterance, followed by 1 - the encoding probabilities for that utterance. Both methods of decoding make use of the same underlying matrix, and we thus might try to develop a probabilistic semantics in terms of these underlying matrices of weights that determine the encoding and decoding behavior.

5.2.2 Scale structure

Recall that in Chapter 1 we introduced the distinction between absolute and relative gradable adjectives, with only the latter giving rise to the phenomena characteristic of vagueness. The difference between absolute and relative gradable adjectives can also be observed in the patterns of entailments they allow. Consider:

(24) Sample A is more impure than sample B
(25) Sample A is purer than sample B

Then (24) entails that sample A has more than the absolute minimum amount of impurity, and is hence impure, and (25) entails that B cannot have the absolute minimum amount of impurity and is hence not pure. Thus we see that (26) and (27) follow from (24) and (25), respectively:

(26) Sample A is impure
(27) Sample B is not pure

In contrast, we see that (28) nor (29) entail (30) nor (31), respectively.

(28) Aaron is taller than Baron
(29) Aaron is shorter than Baron
(30) Aaron is tall
(31) Baron is not short.

This distinction between relative and absolute gradable adjectives is initially unexpected, however, on a truth-conditional, degree-theoretic framework for gradable adjectives: on all views within such a framework, just as ‘tall’ means, roughly, having a degree greater on the height scale than a degree standard for the relevant comparison class, so also ‘pure’ will mean having a degree greater on the purity scale than a degree standard for the relevant comparison class. What then requires that the degree standard for ‘tall’ for the relevant comparison class be contextually sensitive, whereas that for ‘pure’ for the relevant comparison class always be the maximum value on the degree scale, invariant of context, so that the former, but not the
latter, admits of no natural precisifications, invariant of context, and thus always gives rise to borderline cases? Likewise, what requires that the degree standard for ‘pure’ be always the maximum value on the degree scale such that (25), which entails that sample B is not maximally pure, thus entails (27)? And what requires that the degree standard for ‘impure’ be always, in effect, having non-zero degree of impurity, so that (24) always entails (26), no matter how small the difference in purity between sample A and sample B?

On Kennedy’s view, the degree standard of a positive form gradable adjective is chosen so as to ensure the object ‘stands out’ in the context of utterance. (The requirement that the standard be chosen so as to ensure the object stand out seems to make sense; if the standard were not chosen to do so, gradable adjectives would not allow language users to communicate distinctions among objects in a given comparison class.) Borrowing a term from Williamson (2002), Kennedy claims that degree scales with neither maximum nor minimum values, however, have no ‘natural transitions’ relative to which an object might stand out and so where the degree standard might be placed, so adjectives that encode such scales can only be relative gradable adjectives with contextually variant degree standards. However, this cannot be the complete explanation for the distinction between relative and absolute gradable adjectives, since even though degree scales with minimum or maximum values do have natural transitions (from minimum to non-minimum values or maximum to non-maximum values), so far nothing requires that the degree standard be placed at the minimum or maximum values. Kennedy thus appeals to a principle of interpretive economy, according to which participants in a discourse are required to compute the truth conditions of a sentence in a way that maximizes the contributions of the conventional meanings—including the
topological properties of any encoded degree scales—of its constituents.

There are problems with this explanation, however: First, there are questions as to how it fits into a broader theory of economization. Speakers generally seem to face pressure to balance informativeness, brevity, and ease of listener processing. What guarantees that maximization of the contribution of the conventional meanings of the constituents of a sentence to the computation of its truth conditions will not be in tension with these broader pressures? Furthermore, how does the principle of interpretive economy serve these broader communicative aims, and is it derivable from them? Potts (2008) attempts to derive the principle as a result of the cognitive prominence of endpoints, which allows them to ease coordination in signaling games; however, it is unclear how to extend his analysis to applications of interpretive economy beyond gradable adjectives.

Second, it seems there are exceptions to these generalizations: Consider

(32) Snowboarding isn’t dangerous, but it carries some degree of risk.

This does not seem contradictory; however, the degree scale for the *safe/*dangerous antonym pair is lower open and upper closed, as evidenced by the distribution of adverbial modifiers in example (10) in Chapter 1.

Interpretive economy would then require that ‘dangerous’ require only non-maximal degree of safety, and so predicts it should be contradictory. (Potts’ analysis makes the same prediction, and so faces the same problem.) We might claim that pragmatic slack giveth where interpretive economy taketh away, so that even though dangerousness strictly requires only non-maximal safety and so that many more things are dangerous than are actually called
‘dangerous’, an assumption of speaker informativity allows us to speak as if less than merely non-maximal safety is required. However, this explanation should generalize to other absolute gradable adjectives that take maximal degree modifiers, but it does not, as we see if we consider some examples from §1.2.2 of Chapter 1:

(33) * The assay is not impure, but there are some undesired surfactants in it.
(34) * The outcome is not uncertain, but there is some doubt about it.
(35) * That’s not inaccurate, but it is wrong in some respects.

However, it does not. Now, perhaps there are competing pressures that override pragmatic slack for most cases of absolute gradable adjectives, thus allowing the truth conditions generated by interpretive economy to surface; however, in the absence of such, we might hope that the probabilistic picture advanced here might shed light on the relative versus absolute gradable adjective. Recent work by Lassiter and Goodman (2013) and Qing and Franke (2014) has attempted characterize the distinction in terms of the distribution of thresholds that their respective probabilistic models of interpretation generate, and then to explain the distinction in terms of listener’s prior beliefs about what he takes to be the commonly known truth-conditions of the utterances, and a model of the speaker’s cognitive processes, along with assumptions about the prior distribution of the objects in the relevant comparison class on the relevant degree scale. In contrast, the view proposed here would attempt to characterize the distinction between absolute and relative gradable adjectives in terms of the fundamentally probabilistic encoding and decoding behavior generated by the appropriately constrained signaling game; ideally, we might hope to discover under what conditions
senders and receivers in signaling games engage in signaling behavior such that words like ‘perfectly’ combine with words that also can combine with comparative, equative, and measure phrase morphology, and also under what conditions senders and receivers in signaling games engage in signaling behavior such that ‘perfectly’ and ‘slightly’ do not combine with words that also combine with such degree morphology. This would seem to also require that we can also evolve signaling games in which senders and receivers engage in compositional signaling behavior, so that the probabilities of encoding and decoding a complex signal can be determined by the probabilities of encoding and decoding its constituent parts.

5.2.3 The Sorites

Perhaps most obviously, from a philosophical standpoint, is an application of the current framework to the various Sorites arguments. For example, consider:

(36)  a. Feynman is tall.
    b. If Feynman is tall, then Einstein is tall.
    c. If Einstein is tall, then Dirac is tall.
    d. ...
    e. Therefore, Avogadro is tall.

First, given our distinction between the primary, descriptive meaning and the secondary, metalinguistic meaning of ‘x is tall’ we might notice that this Sorites argument appear to be in some sense metalinguistic: we gain no information about the height of the objects mentioned from either the initial premise or from the conditionals. If we did not know the
height of Dirac, then when we hear ‘If Einstein is tall, then Dirac is tall’, we would infer that Dirac is at least as tall as Einstein. The effect of the Sorites requires that we already know the heights of all of the objects. This seems to indicate that each occurrence of ‘$x$ is tall’ in the Sorites sequence is metalinguistic, since its height is already in the common ground: that is, we conditionalize on the height to gain information about the thresholds. This then allows a measure of the probability of the conditional premises: we can define this as the proportion of the area under the posterior threshold distribution for the antecedent that is also under the posterior threshold distribution for the consequent. Since the initial premise is chosen such that Feynman is clearly tall, we measure its probability as 1, or else as the proportion of the area under the threshold distribution prior to the initial premise, that is also under the threshold distribution posterior to the first premise. We can then take the probability of the conclusion to be the product of the probabilities of the premises; then, if the differences in heights are small, each premise has high probability, but the conclusion has low probability.

An alternative is to define the probability of the initial premise as the area under the sending curve below the height of the person who is called tall in the initial premise. The probability of each conditional premise is then defined as the area below sending curve between the height of the person in the antecedent and the height of the person in the consequent. The probability of the conclusion is again product of the probabilities of the premises. Here again, if the differences in height are small, each premise may have high probability, but the conclusion will have a low probability. We might also note that the area under the sending curve below the height of the person called ‘tall’ in the conclusion is a measure of the prior
probability of the conclusion; and we might consider various desirable constraints on the relationship between the prior probability of the conclusion, and the probability assigned to the conclusion by the product of the probabilities of the premises. We might so something similar under the approach in the previous paragraph, but with the prior probability of the conclusion understand as the proportion of the area under the threshold distribution prior to any of the premises, that is also under the the threshold distribution posterior to the conclusion: how much, that is, is the claim ‘Avogadro is tall’ asking you to change the threshold distribution?

Understanding each occurrence of ‘tall’ in this argument as metalinguistic also makes it possible to understand each conditional premise as non-assertoric: just as we say, ‘If you are going to the store then please remember the milk,’ we might also think of each conditional premise as a request, or a proposal, as to where to put the threshold, conditional on its already being somewhere else: ‘If Coulomb counts as tall, then please let’s count Bernoulli as tall too.’ After all, even accounting for the fact that we seem compelled to interpret each occurrence of ‘tall’ in the argument metalinguistically, it seems we are not gaining information about the thresholds from argument: it is not as if . Then, again we can define a measure of the resistance we will encounter to each of the conditional premises as the proportion of the area under the posterior threshold distribution for the antecedent that is also under the posterior threshold distribution for the consequent. Again, we might also define the measure of the resistance we are likely to encounter to the conclusion, as the product of the resistance to each of the premises.

Finally, similar to the assertoric option, we might also define the measure of the resistance
we will encounter to the initial premise as the area under the sending curve above the height of the person who is called ‘tall’ in the initial premise. The resistance to each conditional premises is then defined as the area under the sending curve between the height of the person in the antecedent and the height of the person in the consequent. The resistance to the conclusion is again the area under the sending curve below the height of the person called ‘tall’ in the conclusion. Here again, if the differences in height are small, each premise may have low resistance, but the conclusion will have a high resistance.

There are, of course, other versions of the Sorites, and a full treatment on the current proposal of all of the various forms, along with a comparison of its costs and benefits relative to other approaches, lies beyond the scope of the present work. Nonetheless, I think it is clear that the probabilistic, dynamic approach proposed here is promising.
Chapter 6

Bibliography


