THE PRAGMATICS OF FIGURATIVE LANGUAGE UNDERSTANDING

Dissertation Proposal

Justine T. Kao
November 2015

Abstract

Human communication is rife with figurative language, ranging from hyperbole, to irony, to metaphor. How do people go beyond the literal meaning of an utterance to infer the speaker's intended meaning? In my dissertation, I will model people's interpretations of nonliteral language using an extended version of Rational Speech Act (RSA) models, a family of computational models that formalize language understanding as recursive reasoning between speaker and listener. Using a series of behavioral experiments, I will show that an RSA model extended to incorporate uncertainty about the question under discussion (QUD) is able to predict people's interpretations of hyperbole, irony, and metaphor. I will argue that despite apparent differences among these subtypes of figurative language, the same computational framework can flexibly produce fine-grained interpretations for a range of nonliteral uses. I use this as evidence suggesting that the rich and often affectively-laden meanings expressed by figurative language can be explained by basic principles of communication.

Contents

List of Tables

Chapter 1

Introduction

From "Juliet is the sun" to "That woman is a bombshell," nonliteral language is, quite literally, everywhere. Metaphor, hyperbole, and sarcasm are ubiquitous in human communication, often creating poetic or humorous effects that add rich dimensions to language (Glucksberg, 2001; Pilkington, 2000; Lakoff & Turner, 2009; Roberts & Kreuz, 1994). While figurative statements are often false under their literal meanings, people are highly adept at inferring relevant and true information from these utterances. How do our linguistic, cognitive, and social faculties work together to allow us to fluently and accurately understand figurative language?

An ocean of ink has been spilled on this topic across many areas, including psychology, linguistics, philosophy, computer science, and literary theory (Glucksberg, 2001; Papafragou, 1996; Li & Sporleder, 2010; Kreuz & Roberts, 1993). Some researchers focus on the cognitive mechanisms that enable particular types of figurative language, such as the process of aligning shared properties and analogous relations to understand metaphor (Gentner & Wolff, 1997; Bowdle & Gentner, 2005; Glucksberg, 2003). Others focus on the communicative principles that guide interpretation, such as using conversational maxims to select the appropriate meaning of an utterance (Grice, 1975; Searle, 1979; Sperber & Wilson, 2008; Ortony, 1993; Tendahl & Gibbs, 2008). In my dissertation, I will adopt an approach closer to the latter, with the goal of proposing a general pragmatic framework that explains the basis of figurative communication.

1.1 Background

In this section, I will first review some classic ideas in pragmatics. I will then describe two main bodies of work that specifically examine the pragmatics of figurative language understanding. Finally, I will suggest that a complete theory of figurative communication should more carefully account for certain important factors that affect general language understanding.

1.1.1 Language in general

One of the most important insights in pragmatics is that listeners tend to assume speakers to be rational and cooperative agents who aim to be informative, known as the Cooperative Principle (Grice, 1975; Clark, 1996; Levinson, 2000). When interpreting an utterance, a listener uses these assumptions of rationality and informativeness to reason about what meaning a speaker could want to convey that would lead him to choose a particular utterance. This reasoning between listener and speaker is responsible for many phenomena in pragmatics and language understanding, such as various types of conversational implicatures (Horn, 2006; Levinson, 2000).

When speakers and listeners reason about each other to communicate, they also consider and make use of their shared background knowledge (Clark, 1996). Suppose Liz asks, "Is Bob an honest person?" and Sam replies, "He's a politician." Although Sam did not directly answer Liz's question, it is likely he means "no." Liz is able to successfully interpret Sam's utterance, and Sam is able to successfully use this utterance, because they both have access to the relevant *encyclopedic meaning* of "politician"—the network of background knowledge shared among people in a community, which includes the stereotype that politicians may be dishonest or corrupt (Taylor, 2003; Langacker, 1987). Naturally, Liz's interpretation is sensitive to the contents of the background knowledge they share. If Liz and Sam belong to a community where all politicians are believed to be honest, then Liz would interpret Sam's reply to mean that yes, Bob is an honest person.

Besides background knowledge, Liz's interpretation also relies on her prior beliefs about specific aspects of the state of the world, in this case Bob's profession. Suppose prior to her exchange with Sam, Liz did not know what Bob does for a living. She will have learned two facts about Bob from Sam's utterance: Bob is a politician, and

Bob is not an honest person. Suppose Liz knew *a priori* that Bob is a politician. She will not have learned anything new about Bob's profession from Sam; however, even though she already knows that politicians in general are believed to be dishonest, Sam's utterance makes her more certain that Bob *in particular* is dishonest, because that is the most informative and relevant meaning given her initial question about Bob's honesty. Finally, suppose Liz knows that Bob is a concert pianist and not professionally involved in politics at all. How will Liz interpret Sam's utterance? Instead of updating her beliefs about Bob using the dictionary meaning of "politician," she will rely on the encyclopedic meaning to conclude that Bob is dishonest (but still a concert pianist). In other words, she will interpret the utterance non-literally. These examples show that interpretation of the same utterance in the same local context can vary in a rich and graded manner based on the speaker and listener's prior beliefs.

Listener's interpretations are also often sensitive to the local context. Suppose instead of inquiring about Bob's honesty, Liz had asked, "Is Bob a persuasive speaker?" Sam's utterance may now be interpreted as a compliment: Bob is indeed a persuasive speaker. The local context—whether constructed by an explicit question, a situation, or a salient goal—sets up a "question under discussion" (QUD) to which the information conveyed by the speaker is intended to be maximally relevant (Roberts, 1996).

1.1.2 Figurative language in particular

Two main bodies of work examine the pragmatic principles governing figurative language understanding. The first is the standard pragmatic view, which analyzes figurative utterances using standard Gricean Maxims (Grice, 1975; Searle, 1979). This view proposes a three-step process to understand figurative utterances: (1) determine the literal meaning of the utterance (2) determine whether the literal meaning violates the quality maxim (3) reanalyze the utterance to identify implied or metaphorical meanings that would allow the utterance to adhere to the conversational maxims. While the standard pragmatic view is appealing in that it fits naturally within Grice's general theory of communication, it has met with several criticisms. For one, many figurative statements do not violate the quality maxim because their literal meanings are also true. "No man is an island," for example, is a literally true statement in addition to a figurative one (Gibbs Jr, 1992). By relying on the violation of the quality

maxim to arrive at a figurative interpretation, the standard pragmatic view does not provide a satisfying explanation for these types of utterances. Another common criticism of the standard pragmatic view is that it requires the listener to first access the literal meaning of the utterance, verify that it is contextually inappropriate, compute potential figurative interpretations, and then select the interpretation that best satisfies conversational maxims. Given the number of extra steps involved, a figurative utterance should take longer to understand than a literal utterance. However, many experiments have shown that the figurative meanings of utterances are often accessed as quickly or even more quickly than their literal meanings given certain supporting contexts (Glucksberg, 2003; Gildea & Glucksberg, 1983; Gibbs Jr, 1992). These empirical results suggest that literal meanings do not have to be explicitly rejected to make way for figurative interpretations.

Instead of appealing to the maxim of quality as the guiding principle for interpreting figurative utterances, relevance theory proposes that the principle of relevance is key to explaining a range of phenomena in communication, including figurative language (Sperber et al., 1986). Relevance theorists view figurative language as employing the same comprehension processes as literal language—to maximize the relevance of the interpretation to the speaker's communicative goal (Sperber & Wilson, 2008). Suppose Liz is a chemist who needs to make a solution using water that is exactly 100 degrees celsius. She asks, "How's the temperature?" and Sam replies, "The water is boiling." Liz will interpret this to mean that the water is at boiling point. Now suppose Liz is about to give her baby a bath and asks her husband Sam to test the temperature. Sam replies, "The water is boiling." Liz will interpret this to mean that the water is much too hot, but quite unlikely that it is 100 degrees celsius. Now consider the utterance: "Bob is boiling." If what is relevant is Bob's temperature, this utterance means that Bob has a fever. If the topic is Bob's emotional state, this means that Bob is angry. A relevance theory account of metaphor is that all meanings, not just figurative ones, are selected based on relevance to the question under discussion.

While both the standard pragmatic and relevance theoretic view provide useful frameworks for understanding the pragmatics of figurative communication, certain factors that are important in general communication remain overlooked. For example, neither the standard pragmatic nor relevance theoretic view explicitly take into

account the speaker's rationality or desire to be informative. There is also little consideration of the shared encyclopedic knowledge associated with different utterances, or representation of the listener's prior beliefs. While relevance theory considers which interpretations are maximally relevant to the question under discussion, they do not provide a clear operationalization of relevance. We believe that an adequate model of figurative language understanding should flexibly integrate these components to determine an appropriate interpretation. In addition, figurative language is often used to express subjective experiences and emotional attitudes (Riloff et al., 2005; Roberts & Kreuz, 1994). Since these emotional subtexts contribute greatly to the appeal of figurative language, it is important for a theory of figurative language understanding to include analyses of these effects.

1.2 A proposal

In my dissertation, I will propose a formal modeling framework for figurative language understanding that includes the following components: reasoning about the speaker's choice of utterance, with assumptions of rationality and informativeness; the literal meaning of utterances; shared background knowledge between speaker and listener; specific prior beliefs; local contextual information; affective subtext. While many researchers have suggested that the construction of meaning involves an interplay of these components (Coulson & Oakley, 2005; Gibbs, 1984; Clark, 1996; Stalnaker, 2002), to our knowledge there is no formal model that explicitly describes the relationships among these components and integrates them to produce concrete, fine-grained predictions that can be evaluated against empirical data. In the next chapter, I will describe such a formal model and argue that it may illuminate our understanding figurative communication.

Chapter 2

RSA models with QUD inference

In the last chapter, I proposed that in order to successfully interpret a figurative utterance, a listener needs to (1) reason about a rational and informative speaker (2) consider the literal meaning of an utterance (3) consider relevant background knowledge associated with utterance (4) consider relevant prior beliefs (5) consider the local context and question under discussion (6) understand the subtext or attitudes expressed by the utterance. Here I will review the basic RSA framework and introduce an extension that incorporates these components.

2.1 Basic RSA

Rational Speech Act (RSA) models are a family of probabilistic models that formalize the Cooperative Principle to model how people arrive at pragmatically enriched meanings of utterances (Goodman & Lassiter, 2014; Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013). Under the RSA framework, speaker and listener recursively reason about each other to communicate. A speaker reasons about how to get a particular meaning across to a naive listener; a more sophisticated listener then reasons about the speaker and uses Bayesian inference to recover the intended meaning.

Suppose a speaker S_1 intends to communicate meaning m. He reasons about a naive listener L_0 and considers how likely it is that she will infer m from u. We assume that S_1 incurs a "cost" by uttering u, which is proportional to the physical or cognitive effort required to produce u. In accordance with many models of decision-making, we compute the probability that S_1 will choose utterance u given meaning

m and cost c using the Luce choice rule (Luce, 2005):

$$S_1(u|m) \propto L_0(m|u) \cdot e^{-Cost(u)} \tag{2.1}$$

However, a real listener does not naively interpret utterances in a social vacuum; she considers *why* the speaker chose utterance u to communicate meaning m. This more sophisticated listener, whom we shall call L_1, uses Bayes' Rule to infer m based on her model of how S_1 chooses his utterances. This is captured by the following equation:

$$L_1(m|u) \propto P(m)S_1(u|m) \tag{2.2}$$

where $P(m)$ is the prior probability of the meaning m. This allows L_1 to incorporate background knowledge of m in her interpretation. In principle, the speaker and listener can recursively reason about each other to an arbitrary depth. However, rich pragmatic effects can emerge from depths 1 and 2, which is reason to believe that this framework may be psychologically plausible for modeling pragmatic language understanding.

RSA has proven to be an extremely flexible and productive framework for modeling language understanding. It has been used to explain Horn implicatures (Bergen et al., 2012), vagueness and context-sensitivity in gradable adjectives (Lassiter & Goodman, 2014), the pragmatic use and interpretation of prosody (Bergen & Goodman), effects in syllogistic reasoning (Tessler & Goodman), and more (Goodman & Lassiter, 2014). However, in most of these cases, the pragmatically strengthened meanings do not stray very far from their literal meanings, and thus do not count intuitively as "figurative" interpretations.

Since the RSA framework operates under the assumption that speakers optimize informativeness, it predicts that a rational speaker would never choose an utterance whose literal meaning directly contradicts his intended meaning. For example, suppose S_1 wants to communicate that Bob is a fierce person. According to the basic RSA model, he reasons about the literal listener L_0 to choose the utterance that will most likely convey this information. Since L_0 is a literal listener, given the utterance "Bob is a wolf," she will believe that Bob is literally a wolf. She will thus *not* interpret that Bob is a fierce person, or a person at all. Since the speaker has no reason to say "Bob is a wolf" to communicate that Bob is a fierce person (because the literal

listener would not receive the intended meaning), a pragmatic listener who reasons about the speaker will not interpret "Bob is a wolf" to mean that Bob is a fierce person. This example suggests that while the RSA framework requires significant and theoretically important extensions to explain nonliteral communication.

2.2 Reasoning about QUD

We extend the RSA framework to address the ways in literal meaning, background knowledge, and contextual information shape language understanding through reasoning about relevance to the question under discussion, or QUD (Roberts, 1996). A QUD picks out an immediate topic under discussion, which the speaker is likely to address in order to maintain discourse coherence and achieve the goals of the conversation. Note that the QUD does not have to be established by an explicit question, although it can be; instead, a question under discussion is more generally a contextual element that "proffers a set of relevant alternatives which the interlocutors commit themselves to addressing" (Roberts, 1996). In other words, a QUD helps determine the speaker's communicate goal by constraining the set of things that the speaker may want to address.

The basic RSA models already naturally incorporate certain aspects of background knowledge and prior beliefs. For example, to compute the probability that Bob is a wolf given the utterance "Bob is a wolf," the pragmatic listener must consider the prior probability that Bob is a wolf. However, believing that Bob is a wolf is more than believing that Bob is a large wild animal that often hunts in groups. Once you believe that Bob is a wolf, you are more likely to believe that Bob is furry, fierce, loyal, fast, hungry, etc. These beliefs are graded; you may have a strong belief that any given wolf is fierce, but only a weak belief that any given wolf is loyal. This network of background knowledge forms a rich multi-dimensional representation of what it means to be a wolf. Note that while these other dimensions of meaning may not be part of the core "literal" meaning of the word "wolf," they are associative and easily accessible. As a result, we assume that the basic-level literal listener has access to these dimensions of meaning. Given a core literal meaning c, associated encyclopedic meanings $\vec{a}$, and an utterance u, the literal listener's interpretation of u is now given by the following:

$$L_0(c, \vec{a}|u) = \begin{cases} P(\vec{a}|c) & \text{if } c = u \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

This provides a formal way to enrich literal encoded meaning with associated background knowledge. However, introducing multiple dimensions of meaning alone is insufficient for explaining figurative language understanding. While the literal listener has access to the associated meanings, she still interprets utterances literally. Given the utterance "Bob is a wolf," the literal listener will believe that Bob is a fierce, furry, and loyal wolf with some probability ($P(\vec{a}|c)$); however, she does *not* believe that Bob is a fierce person or any kind of person at all, because she believes that he is a wolf with probability 1. For figurative meaning to arise, the speaker and pragmatic listener must reason about which dimension of meaning is relevant to the question under discussion (QUD).

We formalize relevance to the QUD by introducing a function Q, which projects the meaning that a literal listener derives from an utterance onto only the dimension that is under discussion. This leads to the following utility function for speaker S_1:

$$U(u|c, \vec{a}, Q) = \log \sum_{c,\vec{a}} \delta_{Q(c,\vec{a})=Q(c',\vec{a}')} L_0(c', \vec{a}'|u) \tag{2.4}$$

Given this utility function, the speaker's choice of utterance is the following:

$$S_1(u|c, \vec{a}, Q) \propto e^{\lambda U(u|c,\vec{a},Q)}, \tag{2.5}$$

where λ is a speaker rationality parameter (Luce, 2005). The pragmatic listener L_1 then performs Bayesian inference to guess the intended meaning given prior knowledge and her internal model of the speaker. Since she is uncertain about the precise question under discussion, she marginalizes over the possible QUDs under consideration:

$$L_1(c, \vec{a}|u) \propto P(c)P(\vec{a}|c) \sum_{Q} P(Q)S_1(u|c, \vec{a}, Q)$$

This equation now includes multiple dimensions of meaning, the QUD, a model of the speaker's choice given he wants to be relevant and informative, and the listener's prior beliefs. Something quite magical happens when all of these elements are combined. Since the literal listener is likely to believe that Bob is fierce if she believes that Bob is

a wolf, the speaker is motivated to say "Bob is a wolf" to get her to believe that Bob is a wolf and thus fierce. Furthermore, since the speaker only cares to communicate Bob's fierceness and not which species Bob belongs to, he does not mind that the literal listener will believe that Bob is actually a wolf. The pragmatic listener knows this about the speaker and also knows that Bob is very unlikely to actually be a wolf. Combining these pieces of information, the pragmatic listener ultimately believes that Bob is a fierce person, which is the intuitive interpretation of the utterance "Bob is a wolf."

This simple example suggests that by incorporating QUD inference, the RSA model is able to produce nonliteral interpretations of utterances that match our intuitions. In my dissertation, I will describe three domains in which we empirically tested the extended RSA model—termed qRSA—and show that they predict people's interpretation with high accuracy. In particular, I will show that the model captures several desired effects in the interpretation of hyperbole, irony, and metaphor: (1) nonliteral interpretation (2) sensitivity to background knowledge (3) sensitivity to utterance cost (4) inferences about affective subtext (5) sensitivity to local context (6) sensitivity to alternative utterances.

Chapter 3

Modeling figurative language

This chapter describes behavioral experiments and specific implementations of the qRSA model that demonstrate interesting pragmatic effects in three types of figurative language: hyperbole, irony, and metaphor. The work on hyperbole has been published (Kao et al., 2014) and will only be briefly summarized; the experiments and model for irony will be described in detail; and the section on metaphor is ongoing and will be presented as a series of proposed studies.

3.1 Hyperbole

A hyperbole is an exaggerated statement that purposefully presents its subject as more extreme than it actually is (Roberts & Kreuz, 1994; McCarthy & Carter, 2004). In a modern analysis of a corpus of spoken English, McCarthy & Carter (2004) found that hyperbole occurs frequently in everyday conversations and is often used to express emotions and provoke humorous responses. To examine how people arrive at appropriate interpretations and affective subtexts of these types of literally false utterances, Kao et al. (2014) focused on utterances that involve number words. We chose numbers because the literal meanings of numbers are very easy to formalize: for example, the literal meaning of the utterance "a thousand" is simply 1000.

We conducted experiments using the prices of three everyday items—electric kettles, watches, and laptops. To examine the effect of prior knowledge on hyperbole interpretation, we first asked participants to rate the probabilities of various prices as well as the probability of someone thinking that a certain price is too expensive.

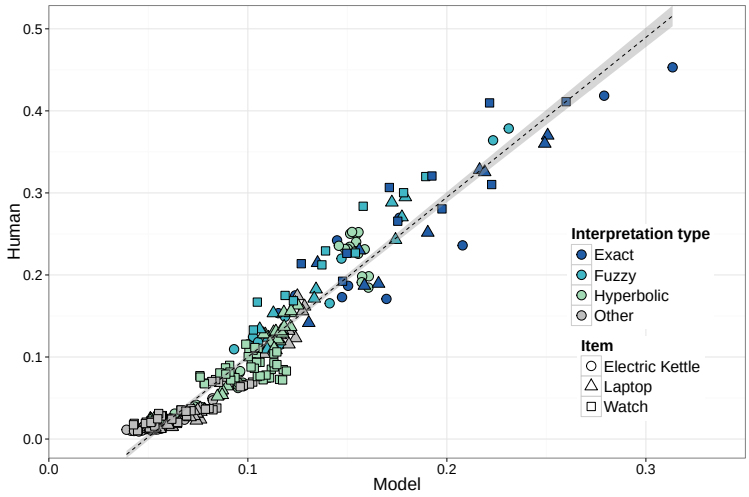

Figure 3.1: Model predictions v.s. average human interpretations. Each point represents an utterance and price state pair (u, s). The x-coordinate of each point is the probability of the model interpreting utterance u as meaning price state s; the y-coordinate is the empirical probability. Correlation between model and human interpretations is 0.968.

Using these empirically measured priors, we obtained the meaning distributions predicted by the qRSA model for each utterance. The model reasons about different types of communicative goals that the speaker may have, including the goal to communicate affect about the price of the item. By reasoning about relevance to these communicate goals, the model captures a basic feature of hyperbole: utterances whose literal meanings are less likely given the price prior are more likely to be interpreted hyperbolically. For example, "The watch cost 1000 dollars" is more likely to be interpreted hyperbolically than "The laptop cost 1000 dollars."

To quantitatively evaluate the model's predictions, we asked participants to interpret potentially hyperbolic utterances. For example, given that Sam said: "The watch cost 1000 dollars," how likely is it that the watch cost x dollars? For all utterances, we then compared the model's and participants' interpretations. The model predictions are highly correlated with people's interpretations ($r = 0.968, p < 0.0001$) (Figure 3.1), suggesting that the qRSA model is able to combine linguistic information, background knowledge, and reasoning about the speaker's goals to interpret hyperbolic utterances.

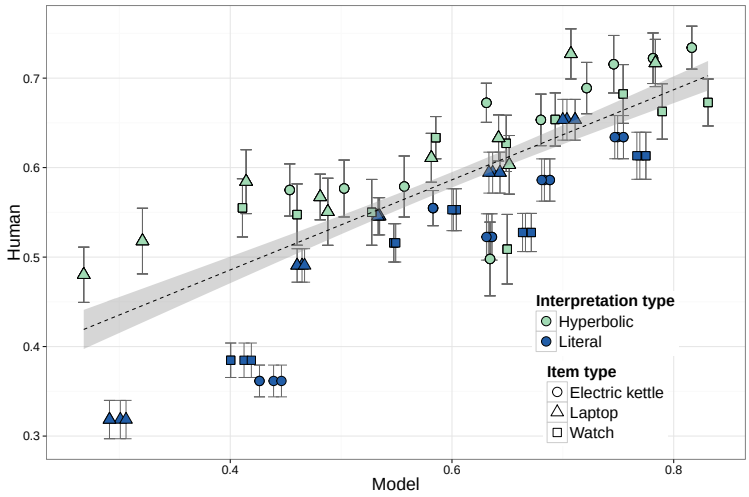

Figure 3.2: Model predictions of affect v.s. human ratings. Each point represents an utterance and price state pair (u, s). For pairs where $u = s$, the utterance is literal; for $u > s$, the utterance is hyperbolic. The x-coordinate of each point is the model's prediction of the probability that the utterance/price state pair conveys affect; the y-coordinate is participants' affect ratings (error bars are standard error). Correlation between model and humans is 0.775.

In addition to producing the appropriate corrective response to hyperbolic utterances, the model also captures the affective subtext of hyperbole. We conducted a separate experiment to examine peoples' interpretation of affect in hyperbolic versus literal utterances. Participants read scenarios in which a speaker bought an item that cost s dollars and says it cost u dollars, where $u \geq s$. They then rated how likely it is that the buyer thinks the item was too expensive. Results showed that utterances u where $u > s$ (hyperbolic utterances) are rated as significantly more likely to convey affect than utterances where $u{=}s$ ($F(1, 25) = 12.57, p < 0.005$). Moreover, if a watch actually cost 100 dollars and Sam says something hyperbolic such as "The watch cost 1000 dollars," people are more likely to believe that Sam thinks the watch is too expensive than if the watch actually cost 1000 dollars and Sam says "The watch cost 1000 dollars." This suggests that listeners infer affect from hyperbolic utterances above and beyond the affect associated with a given price state. Quantitatively, we compared model and human interpretations of affect for each of the 45 utterance and price state pairs (u, s) where $u \geq s$. While there is a significant amount of noise in the human judgments (average split-half correlation is 0.833), the model predicts

human interpretations of the utterances' affective subtext significantly better than chance ($r = 0.775, p < 0.00001$), capturing most of the reliable variation in these data (Figure 3.2).

Results from Kao et al. (2014) suggest that by incorporating inferences about the speaker's communicative goals, the qRSA model successfully interprets hyperbolic utterances and appropriately recovers the affective subtext. However, in this initial exploration of applying the qRSA model to figurative language, we only considered a simplified space of affect, namely the presence or absence of negative feeling. In the next section, we explore how expanding the space of affect to include emotions with positive/negative valence and high/low arousal accounts for people's interpretations of ironic utterances.

3.2 Irony

An ironic statement describes something as contrary to what it actually is (Roberts & Kreuz, 1994; Gibbs & Colston, 1999). For example, a speaker who says "Such lovely weather we are having" in the middle of a storm means that the weather is *not* lovely and expresses a negative attitude towards it. How do people appropriately interpret these superficially positive or negative utterances? Can our model use QUD inference to interpret an utterance when its literal meaning is not just an exaggerated version of the intended meaning, but rather its opposite? In this section, we will examine the consequence of expanding the set of emotions we consider to an empirically derived affect space. We show that this minimal change enables the qRSA model to capture many of the rich inferences resulting from verbal irony.

In what follows, we will examine interpretations of potentially ironic utterances in an innocuous domain—the weather. We chose the weather as the victim of irony for several reasons. First, people are quite familiar with talking (and complaining) about the weather. Second, we can visually represent the weather to participants with minimal linguistic description in order to obtain measures of nonlinguistic contextual knowledge. Finally, given the critical role that context plays in understanding irony, we can vary the weather states to observe how the same utterance is interpreted differently given different contextual knowledge. This offers to our knowledge the first fine-grained manipulation and quantitative measure of context in studies of irony.

We first explore how an enriched space of affect impacts the qRSA model and find that it produces ironic interpretations in some simple simulations. We then present two behavioral experiments that examine people's interpretations of utterances given different weather contexts. We show that by accounting for two types of affective dimensions, valence and arousal, our model produces interpretations that closely match humans'.

3.2.1 Model

Following the qRSA model described in (Kao et al., 2014), a speaker chooses an utterance that most effectively communicates information regarding the question under discussion (QUD) to a literal listener. We consider a meaning space that consists of the variables s, A, where s is the state of the world, and A represents the speaker's (potentially multidimensional) affect towards the state. We formalize a QUD as a projection from the full meaning space to the subset of interest to the speaker, which could be s or any of the dimensions of A. We specify the speaker's utility as information gained by the listener about the topic of interest—the negative surprisal of the true state under the listener's distribution given an utterance, u, along the QUD dimension, q. This leads to the following utility function:

$$U(u|s, A, q) = \log \sum_{s', A'} \delta_{q(s,A)=q(s',A')} L_0(s', A'|u) \tag{3.1}$$

where L_0 describes the literal listener, who updates her prior beliefs about s, A by assuming the utterance to be true of s. The speaker S chooses an utterance according to a softmax decision rule Sutton & Barto (1998): $S_1(u|s, A, q) \propto e^{\lambda U(u|s,A,q)}$, where λ is the rationality parameter. A pragmatic listener L_1 then takes into account prior knowledge and his internal model of the speaker to determine the state of the world as well as the speaker's affect. Because L_1 is uncertain about the QUD, he marginalizes over the possible QUDs under consideration:

$$L_1(s, A|u) \propto P(s)P(A|s) \sum_q P(q)S(u|s, A, q)$$

The resulting distribution over world states and speaker affects is an interpretation of the utterance.

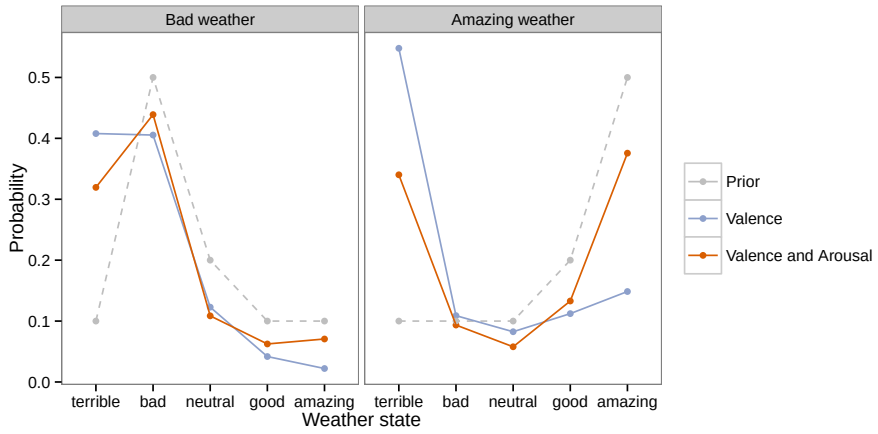

Figure 3.3: Model interpretations of "The weather is terrible" given different prior beliefs about the weather state and affect dimensions. Gray dotted lines indicate prior beliefs about weather states given a weather context; blue lines indicate interpretations when reasoning only about the speaker's valence; orange lines indicate interpretations when reasoning about both valence and arousal.

We performed the following simulations to examine the model's behavior using affect spaces, A, that differ in complexity and structure. We assume that s has five possible ordered values: `terrible`, `bad`, `neutral`, `good`, and `amazing`. We consider two different weather contexts: apparently bad weather and apparently amazing weather, which are each specified by a prior distribution over these states (see gray dotted lines in Figure 3.3). We then examine how the model interprets the sentence "The weather is terrible" in the two weather contexts, given different affect spaces.

We first consider a one-dimensional affect space, where the dimension is emotional valence, and the values are whether the speaker feels negative or positive valence towards the state. The blue lines in Figure 3.3 show the model's interpretation of "The weather is terrible" using this one-dimensional affect space. The model is capable of non-literal interpretation: it produces a hyperbolic interpretation (that the weather is merely `bad`) given "The weather is terrible" in the bad weather situation. However, it produces a literal interpretation (that the weather is `terrible`) in the amazing weather situation. Since a pragmatic listener that only considers emotional valence does not believe that the speaker has any reason to choose a negative utterance to express positive affect (because the utterance communicates no true information), a

model that only considers emotional valence is unlikely to infer a positive world state from a negative utterance (and vice versa), thus failing to evidence verbal irony.

What true information *could* a speaker communicate about a positive world state using a negative utterance? Affective science identifies two dimensions, termed valence and arousal, that underly the slew of emotions people experience (Russell, 1980). For example, *anger* is a negative valence and high arousal emotion, while *contentment* is a positive valence and low arousal emotion. Could speakers leverage the arousal dimension to convey high arousal and positive affect (e.g. excitement) using utterances whose literal meanings are associated with high arousal but negative affect (e.g. "The weather is terrible!")? We test the consequences of incorporating the arousal dimension. The orange lines in Figure 3.3 show simulations of the qRSA model with a two-dimensional affect space: whether the speaker feels negative/positive valence and low/high arousal towards the weather state. Given strong prior belief that the weather state is `bad`, the model interprets "The weather is terrible" to mean that the weather is likely to be `bad`, again producing a hyperbolic interpretation. However, given strong prior belief that the weather is `amazing`, the model now places much greater probability on the ironical interpretation of "The weather is terrible," meaning that the weather is likely `amazing`. This is because, with the enriched two-dimensional affect space, the pragmatic listener realizes that the speaker may be using "terrible" to communicate high emotional arousal. Note that this result is not simply due to the model falling back on the prior: given the same priors, the model interprets the neutral utterance "The weather is ok" as the weather state being `neutral` and not `amazing`. These simulations suggest that a psychologically realistic, two-dimensional affect space enables the qRSA model to interpret ironic utterances in addition to hyperbolic ones.

3.2.2 Experiments

To quantitatively test whether the qRSA model with expanded affect space can capture a range of ironic interpretations, we need appropriate prior distributions as well as data for human interpretations. We conducted Experiment 1 to measure prior beliefs over weather states ($P(s)$) for a range of weather contexts as well as the likelihood of various emotions towards each weather state. The latter allows us to empirically derive the affective space and priors, $P(A|s)$, for this domain. In Experiment 2, we

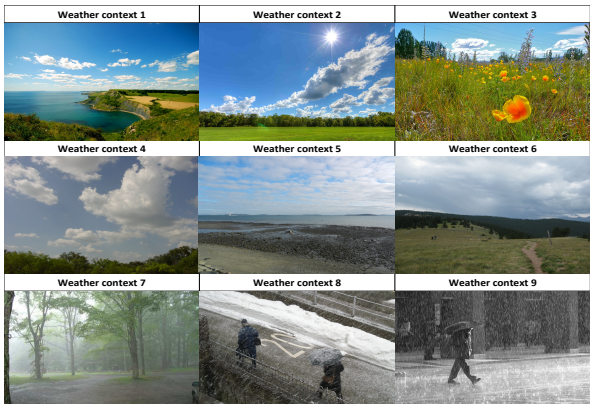

Figure 3.4: Weather images shown to participants in Experiments 1 and 2.

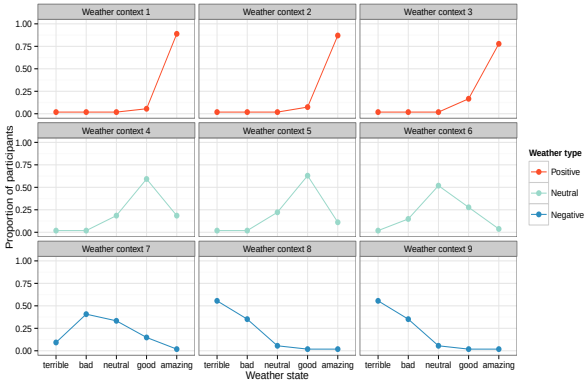

Figure 3.5: Proportion of participants who rated each weather context as each weather state.

collected people's ratings of how a speaker perceives and feels about the weather given what she says in a weather context (e.g. "The weather is terrible!" when the context clearly depicts sunny weather).

Experiment 1: Prior elicitation

Materials and methods

We selected nine images from Google Images that depict the weather. To cover a range of weather states, three of the images were of sunny weather, three of cloudy weather, and three of rainy or snowy weather. We refer to these images as weather contexts. Figure 3.4 shows these nine images. 49 native English speakers with IP addresses in the United States were recruited on Amazon's Mechanical Turk. Each participant saw all nine images in random order. In each trial, participants were told that a person (e.g. Ann) looks out the window and sees the view depicted by the image. They then indicated how Ann would rate the weather using a labeled 5-point Likert scale, ranging from `terrible`, `bad`, `neutral`, `good`, to `amazing`. Finally, participants used slider bars (end points labeled "Impossible" and "Absolutely certain") to rate how likely Ann is to feel each of the following seven emotions about the weather: *excited*, *happy*, *content*, *neutral*, *sad*, *disgusted*, and *angry*, which are

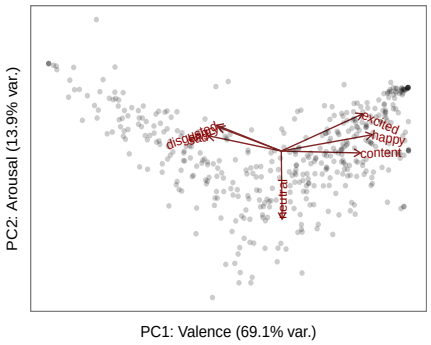

Figure 3.6: Biplot of the first two principle components of the seven emotion ratings, which roughly correspond to valence and arousal.

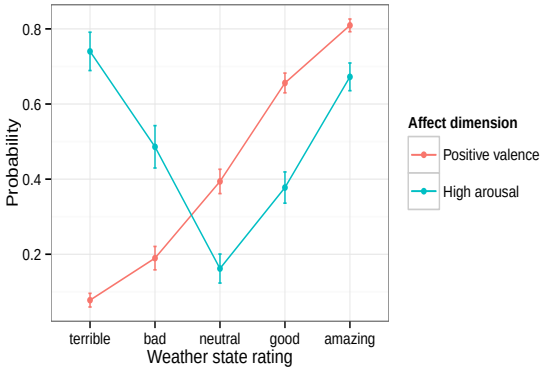

Figure 3.7: Average probabilities of positive valence and high arousal associated with each weather state; error bars are 95% confidence intervals.

common emotion categories (Ekman, 1992)[1]. The order of the emotions was randomized for each participant but remained consistent across trials [2].

Results

For each of the nine weather contexts, we obtained the number of participants who gave each of the weather state ratings. By performing add-one Laplace smoothing on the counts, we computed a smoothed prior distribution over weather states given each context, namely $P(s)$ (Figure 3.5). To examine participants' ratings of the affect associated with each context, we first performed Principal Component Analysis (PCA) on the seven emotion category ratings. This allowed us to compress the ratings onto a lower-dimensional space and reveal the main affective dimensions that are important in this domain, as is often done in affective science Russell (1980). We found that the first two principal components corresponded to the dimensions of emotional valence and emotional arousal, accounting for 69.14% and 13.86% of the variance in the data,

[1]From the most frequently cited set of six basic emotions, we removed *fear* and *surprise* and added *content* and *excited* to have a balanced set of positive and negative emotions. We also added *neutral* to span a wider range of emotional arousal.

[2]Link to Experiment 1: `http://stanford.edu/~justinek/irony_exp/priors/priors.html`

respectively. The PCA represents emotion ratings for each trial as real values between negative and positive infinity on each of the dimensions. To map these values onto probability space, we first standardized the scores on each dimension to have zero mean and unit variance. We then used the cumulative distribution function to convert the standardized scores into values between 0 and 1. This gives us the probabilities of Ann feeling positive (vs. negative) valence and high (vs. low) arousal for each trial, which is a two-dimensional probabilistic representation of her affect. By calculating the average probabilities of positive valence and high arousal given each weather state rating, we obtain the probability of positive valence and high arousal associated with each weather state, namely $P(A|s)$ (Figure 3.7).

Experiment 2: Irony understanding

Results from Experiment 1 give us the components to generate interpretations of utterances from our model. Here we describe an experiment that elicits people's interpretations of utterances, which we then use to evaluate model predictions.

Materials and methods

59 native English speakers with IP addresses in the United States were recruited on Amazon's Mechanical Turk. Each participant saw all nine images from Figure 3.4 in random order. In each trial, participants were told that a person (e.g. Ann) and her friend are in a room looking out the window together and see the view depicted by the image. Ann says, "The weather is _____!" where the adjective is randomly selected at each trial from the following set: "terrible," "bad," "ok," "good," and "amazing." Participants first rated how likely it is that Ann's statement is ironic using a slider with end points labeled "Definitely NOT ironic" and "Definitely ironic." They then indicated how Ann would actually rate the weather using a labeled 5-point Likert scale, ranging from `terrible`, `bad`, `neutral`, `good`, to `amazing`. Finally, participants used sliders to rate how likely it is that Ann feels each of seven emotions about the weather [3].

Results

[3]Link to Experiment 2: `http://stanford.edu/~justinek/irony_exp/interpretation/interpretation_askIrony.html`

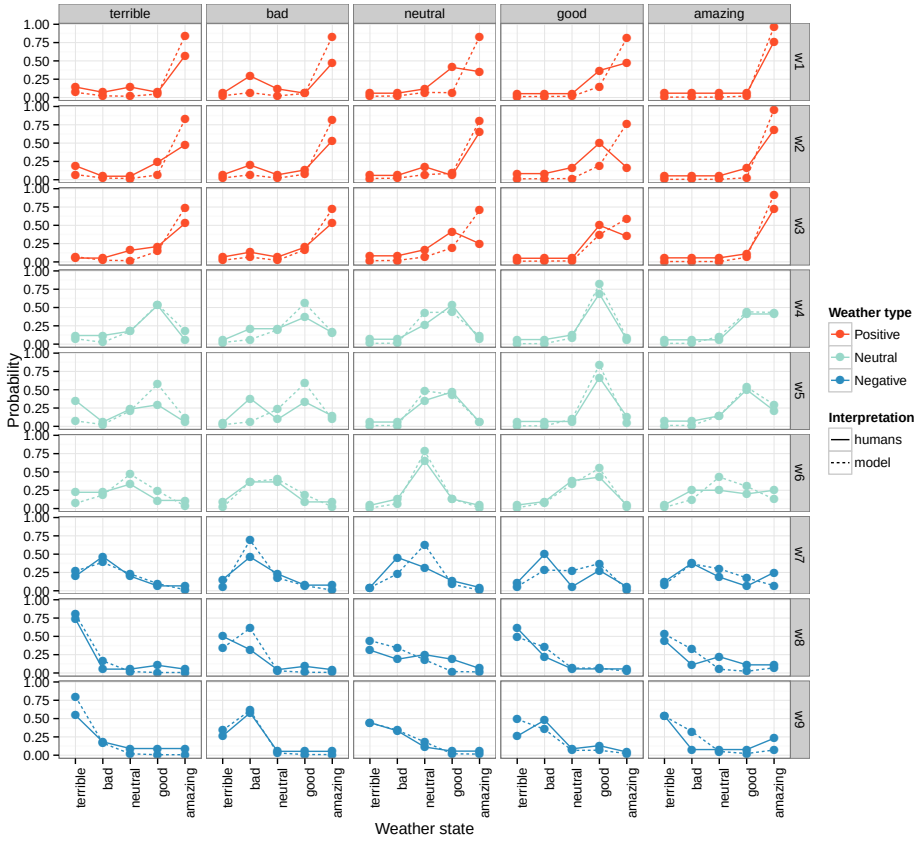

Figure 3.8: Model's and participants' inferences about the weather state (x-axis) given a weather context (row) and an utterance (column). Each panel represents interpretations of an utterance in a weather context. The solid lines are participants' ratings; the dotted lines are model's posterior distributions over weather states.

We first examined participants' irony ratings for each of the weather context and utterance pairs. We found a basic irony effect, where utterances whose polarities are inconsistent with the polarity of the weather context are rated as significantly more ironic than utterances whose polarities are consistent with the weather context ($t(34.16) = -11.12, p < 0.0001$). For example, "The weather is terrible" (a negative utterance) is rated as more ironic in weather context 1 (positive context) ($M = 0.90$, $SD = 0.21$) than in weather context 7 (negative context) ($M = 0.15$, $SD = 0.27$). A linear regression model with the polarity of the utterance, the polarity of the weather context, and their interaction as predictors of irony ratings produced an adjusted R^2 of 0.91, capturing most of the variance in the data. This suggests that participants'

lay judgments of irony align with its basic definition: utterances whose apparent meanings are opposite in polarity to the speaker's intended meaning. Given the fact that participants can identify verbal irony based on its inconsistency with context, how do they then use context to determine the speaker's intended meaning? We examined participants' interpretations of utterances given contexts. For each of the 45 weather context (9) $\times$ utterance (5) pairs, we obtained the number of participants who gave each of the five weather state ratings (`terrible, bad, neutral, good, amazing`). We performed add-one Laplace smoothing on the counts to obtain a smoothed distribution over weather states given each context and utterance (solid lines in Figure 3.8). Results show that participants produce ironic interpretations of utterances, such that the weather is most likely to be `amazing` given that the speaker said "The weather is terrible" in weather context 1. Participants also produce hyperbolic interpretations, such that the weather is most likely to be `bad` given that the speaker said "The weather is terrible" in weather context 7. This confirms the intuition that people are highly sensitive to context and use it both to determine when an utterance is not meant literally and to appropriately recover the intended meaning. Finally, we examine participants' inferences about the speaker's affect given utterances in context. We used the loadings from the PCA on emotion ratings from Experiment 1 to project the emotion ratings from Experiment 2 onto the same dimensions. We then standardized and converted the scores into values between 0 and 1, as before, which gives us probability ratings of the speaker feeling positive valence and high arousal given an utterance and weather context.

3.2.3 Model Evaluation

From Experiment 1, we obtained the prior probability of a weather state given a context ($P(s)$) as well as the probability of affect given a weather state ($P(A|s)$). In addition, we fit three free parameters to maximize correlation with data from Experiment 2: the speaker optimality parameter ($\lambda = 1$) and the prior probability of each of the three QUDs ($P(q_{state}) = 0.3$, $P(q_{valence}) = 0.3$, $P(q_{arousal}) = 0.4$)[4]. For each of the 45 utterance and weather context pairs, the model produced an interpretation consisting of the joint posterior distribution $P(s, A|u)$, where A can

[4]Since $P(q_{state}) + P(q_{valence}) + P(q_{arousal}) = 1$, $P(q_{arousal})$ is determined by the other two QUD parameters and not a free parameter.

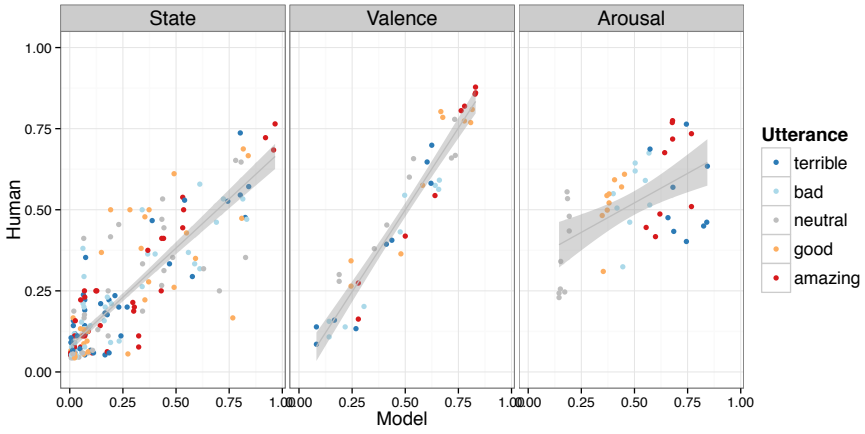

Figure 3.9: Scatter plot showing correlations between model predictions and human ratings for weather state, speaker valence, and speaker affect. Colors indicate utterances.

be further broken down into valence and arousal dimensions. We will examine the model's performance on each of these state and affect dimensions by marginalizing over the other dimensions.

Figure 3.9 shows scatter plots correlating model predictions with human interpretation data for each of the dimensions: weather state, valence, and arousal. The model predictions of weather state given utterance match humans' interpretations, with a correlation of 0.86. Since the split-half correlation for the human data is $\rho = 0.898$ (95%CI $= [0.892, 0.903]$)[5] we find that our model captures much of the explainable variance in human judgements. The model predicts humans' interpretations of valence extremely well, with a correlation of 0.96, capturing essentially all of the explainable variance in the data ($\rho = 0.948 \pm 0.001$). Importantly, the model infers the appropriate valence even when it is inconsistent with the valence of the utterance's literal meaning. The model's predictions for emotional arousal match humans' with a correlation of 0.66, capturing a substantial amount of the explainable variance ($\rho = 0.763 \pm 0.005$). Furthermore, the absolute difference between the model's inferred valence and the valence of the utterance's literal meaning correlates

[5]Split-half correlations ρ were calculated by repeatedly bootstrapping samples from the data (sample each participant with replacement), computing correlation between two halves of the bootstrapped samples, and using the Spearman-Brown prediction formula to estimate predicted reliability with full sample size. Confidence intervals are 95% CI over 1000 iterations of bootstrap sampling.

Model	State	Valence	Arousal	Average
Literal	0.38	0.45	0.49	0.44
Prior	0.79	0.84	0.49	0.71
Valence	0.84	0.79	0.61	0.75
Valence + arousal	0.86	0.96	0.66	0.83
Best possible	0.90	0.95	0.76	0.87

Table 3.1: Correlation coefficients between model predictions and human interpretations of weather state, valence, and arousal given an utterance and weather context from Experiment 2. *Best possible* gives an estimate of the maximum possible correlation given noise in the data (see footnote 5).

significantly with people's irony ratings ($r = 0.86$, $\rho = 0.94 \pm 0.005$), suggesting that the model is able to use inconsistencies between literal and interpreted meanings to identify ironic uses.

We considered a series of simpler models to show that the full model using a two-dimensional affect space best predicts human interpretations. We first examined a model that interprets utterances literally, such that "The weather is terrible" is always interpreted as the weather state being `terrible`, along with the valence and arousal associated with `terrible` weather. We then examined a model that simply ignores the speaker's utterance and takes into account only the state and affect priors associated with each weather context. Finally, we examined the performance of the qRSA model with a unidimensional affect space (valence only). Table 3.1 shows the models' correlations with human judgements for state, valence, and affect. A complete model that takes into account prior knowledge, the literal meaning of the utterance, and a two-dimensional affect space outperforms the other models. This dominance is especially apparent with respect to inferences about valence, which is the most important aspect of understanding an ironic utterance, since the listener must infer the intended positive/negative valence from an ostensibly negative/positive utterance. These comparisons suggest that our full model successfully leverages richer knowledge of affect and uses pragmatic reasoning to produce the appropriate nonliteral interpretations.

In this work, we formalized intuitions about verbal irony understanding and clarified the role of shared prior knowledge in ironic interpretations. We explored the consequences of expanding the space of affect considered by previous Rational Speech Act models to account for verbal irony. By making a minimal extension to (Kao et al.,

2014)'s hyperbole model, we were able to capture people's fine-grained interpretations of ironic utterances in addition to hyperbole. This provides evidence that hyperbole and irony may operate using similar underlying principles of communication—reasoning about shared background knowledge as well as the speaker's affective goals.

3.3 Metaphor

Metaphors are utterances that implicitly compare ideas or concepts from different domains (Gibbs & Colston, 1999; Roberts & Kreuz, 1994). It has inspired a particularly impressive amount of research in cognitive science, spanning topics such as how metaphors structure and shape our thoughts (Ortony, 1993; Lakoff et al., 1993; Thibodeau & Boroditsky, 2011), whether metaphor processing recruits the same strategies as standard language processing (Giora, 1997; Gibbs, 2002; Glucksberg & Keysar, 1993) and what factors determine people's interpretation a novel metaphor (Gentner & Wolff, 1997; Blasko & Connine, 1993; Tourangeau & Sternberg, 1981; Kintsch & Bowles, 2002). This overwhelming interest in metaphor research is due to both the ubiquity of metaphor in everyday language and the potential role of metaphor for helping us understand how the mind creates meaning.

Why do people choose to use metaphors to communicate? What are some characteristics of metaphor that contribute to its popularity? Roberts & Kreuz (1994) examined the discourse goals that people have when they use various figurative tropes. They found that the most common goals for using metaphor were to clarify (82%), to add interest (71%), to compare similarities (35%), to provoke thought (35%), and to be eloquent (35%). Interestingly, although metaphors are defined as implicit comparisons, "to compare similarities" is less frequently listed as a goal than "to clarify" and "to add interest." This suggests that beyond examining the cognitive processes for comparing and aligning concepts that may be involved in metaphor understanding, it is also important to consider other higher-level communicative functions that metaphors may serve.

In addition to Roberts & Kreuz (1994)'s exploration of discourse goals, other researchers have suggested that metaphorical utterances can be used to efficiently express complex meanings (Ortony, 1975; Boerger, 2005; Glucksberg, 1989). In communication tasks where pairs of participants are separated by a screen and asked to

refer to abstract geometrical objects, participants often prefer to describe objects analogically in terms of other known objects rather than use literal analytical descriptions (Clark & Wilkes-Gibbs, 1986; Glucksberg, 1989). Fussell & Krauss (1989) found that these analogical and figurative descriptions tend to be shorter than literal descriptions. In addition, they found that figurative descriptions are used significantly more often when the intended audience is one's self, where presumably there is a great deal of common ground, than when the intended audience is a different person. These findings suggest that people may be balancing efficiency and clarity when choosing figurative versus literal descriptions. Glucksberg & Keysar (1990) wrote, "Metaphors are used to communicate a complex, patterned set of properties in a shorthand that is understood by the members of a speech community who share relevant mutual knowledge" (pp. 16). As a result, a potential benefit of speaking in metaphor may be to utilize common ground to communicate both clearly and efficiently.

In my dissertation, I plan to examine the following (hypothesized) characteristics of metaphorical utterances: (1) metaphors lead to more striking, or extreme, interpretations than literal descriptions (2) metaphors can communicate information along various dimensions with a minimal number of words (3) the interpretation of a metaphor is highly sensitive to the local context (4) the aptness of a metaphor is related to whether there are alternative utterances that can communicate the same amount of information more efficiently. To examine these hypotheses, I will test the qRSA model on different types of metaphorical interpretations.

3.3.1 Model sketch

To reasonably limit the scope of our work, we focus on metaphors of the classic form "X is a Y." Suppose a speaker uses the following utterance to describe a person, Bob: "Bob is a giraffe." How should a listener interpret this utterance? Following the qRSA framework, a listener again assumes that the speaker chooses an utterance to maximize informativeness about a subject along dimensions that are relevant to the QUD. Unlike hyperbole and irony, however, these dimensions are not affective in nature. Rather, they are features associated with the metaphorical source, in this case "giraffe."

We again introduce a literal listener L_0, who interprets the utterances as meaning that Bob is literally a giraffe. Since L_0 believes Bob is a giraffe, she also believes that

Bob is likely to have features associated with giraffes, for example, being tall. The following equation represents the literal listener's interpretation, where c is Bob's category (either a "person" or a "giraffe"), and $\vec{f}$ is a vector representation of Bob's features. $P(\vec{f}|c)$ is thus the prior probability that a member of category c has feature vector $\vec{f}$.

$$L_0(c, \vec{f}|u) = \begin{cases} P(\vec{f}|c) & \text{if } c = u \\ 0 & \text{otherwise} \end{cases}$$

The QUD may be Bob's species category, or Bob's feature(s). We define the speaker's utility as the negative surprisal of the true state under the listener's distribution, projected along the QUD dimension. This leads to the following utility function for speaker S_1:

$$U(u|Q, c, \vec{f}) = \log \sum_{c,\vec{f}} \delta_{Q(c,\vec{f})=Q(c',\vec{f}')} L_0(c', \vec{f}'|u) \tag{3.2}$$

Given this utility function, the speaker chooses an utterance according to a softmax decision rule that describes an approximately rational planner (Sutton & Barto, 1998), where λ is an optimality parameter:

$$S_1(u|Q, c, \vec{f}) \propto e^{\lambda U(u|Q,c,\vec{f})}, \tag{3.3}$$

The pragmatic listener L_1 uses Bayesian inference to guess the intended meaning given prior knowledge and his internal model of the speaker. To determine the speaker's intended meaning, L_1 marginalizes over the possible speaker goals under consideration:

$$L_1(c, \vec{f}|u) \propto P(c)P(\vec{f}|c) \sum_Q P(Q)S_1(u|Q, c, \vec{f})$$

If L_1 believes it is *a priori* very unlikely that Bob is actually a giraffe and that S_1 may want to communicate about Bob's height, she will end up with a posterior distribution where Bob is very likely to be a person who is tall. By combining prior knowledge with reasoning about the speaker's communicative goal, the pragmatics listener can thus arrive at a figurative interpretation of "Bob is a giraffe"—Bob is a very tall person.

3.3.2 Study 1: "Hyperbolic" effects

In Study 1, we use a series of experiments to examine people's interpretations of metaphors such as "Bob is a giraffe," where the salient feature being communicated (e.g. height) lies on a continuous scale. We find that the interpretation of these types of metaphors demonstrates two effects: First, given the utterance "Bob is a giraffe," listeners believe that Bob is taller than they do when given the literal description "Bob is tall." Second, listeners believe that Bob is taller than the average man, but much shorter than the average giraffe. We show that the qRSA model is able to capture these effects.

3.3.3 Study 2: Non-paraphrasability

In Study 2, we use a series of experiments to show that some metaphorical utterances such as "Bob is an ox" allow speakers to communicate about multiple dimensions of Bob at once—for example, that Bob is strong, big, and slow. In order to communicate the same amount of information using a literal description, a speaker would need to choose a much longer and costlier utterance, such as: "Bob is strong, big, and slow." We show that listeners infer information along more dimensions given a metaphorical utterance than given a single literal description. We then show that the qRSA model is able to capture these differences between literal and metaphorical utterances.

3.3.4 Study 3: Context-sensitivity

In Study 3, we use a series of experiments to show that the interpretation of metaphorical utterances such as "Bob is an ox" is highly sensitive to the local conversational context. For example, if Liz asks: "What is Bob like?" and Sam replies: "Bob is an ox," Liz will believe that Bob is perhaps strong, or big, or slow. However, if Liz instead asks: "Is Bob strong?" and Sam replies: "Bob is an ox," Liz will be much more likely to believe that Bob is strong. On the other hand, regardless of Liz's initial question, the interpretation of literal utterances such as "Bob is strong" is unlikely to vary with the conversational context. We show that this effect can be modeled in qRSA using different priors over QUDs.

3.3.5 Study 4: Aptness and alternatives

Many researchers have examined the factors that contribute to the "aptness" of a metaphor as well as how aptness facilities metaphor processing (Tourangeau & Sternberg, 1981; Jones & Estes, 2006; Blasko & Connine, 1993). In Study 4, we propose that aptness may in part be related to two pragmatic factors: (1) The presence of multiple salient features of the metaphorical target, which cannot be described literally without using longer utterance (2) The absence of salient alternative metaphors that the speaker could have used to communicate about the dimension(s) under discussion. We will test these ideas using a series of experiments, and show that the qRSA model can capture some of these effects.

Chapter 4

Discussion

In my dissertation, I will show that an RSA model extended to incorporate QUD inference can account for many important phenomena in figurative language understanding. Despite apparent differences among subtypes of figurative language, the same computational framework can flexibly produce fine-grained interpretations for a range of nonliteral uses, such as hyperbole, irony, and metaphor. I will use this as evidence to suggest that the rich and often affectively-laden meanings expressed by figurative language can be explained by basic principles of communication.

While the RSA framework presents a promising start to examining figurative language understanding, it also invites many more questions. For example, can the model explain metaphor interpretation that requires feature alignment and identification of analogical relations? How might we use the qRSA model to explain higher-level social functions of nonliteral language, such as humor and conveying social closeness? By providing a formal framework for modeling figurative language understanding, I hope that this work can eventually address these questions and shed light on how we add figurative flesh to the bare literal bones of language.

Bibliography

Bergen, L., & Goodman, N. D. (????). The strategic use of noise in pragmatic reasoning.

Bergen, L., Goodman, N. D., & Levy, R. (2012). Thats what she (could have) said: How alternative utterances affect language use. In *Proceedings of the thirty-fourth annual conference of the cognitive science society*.

Blasko, D. G., & Connine, C. M. (1993). Effects of familiarity and aptness on metaphor processing. *Journal of experimental psychology: Learning, memory, and cognition*, *19*(2), 295.

Boerger, M. A. (2005). Variations in figurative language use as a function of mode of communication. *Journal of psycholinguistic research*, *34*(1), 31–49.

Bowdle, B. F., & Gentner, D. (2005). The career of metaphor. *Psychological review*, *112*(1), 193.

Clark, H. H. (1996). *Using language*, vol. 1996. Cambridge University Press Cambridge.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Coulson, S., & Oakley, T. (2005). Blending and coded meaning: Literal and figurative meaning in cognitive semantics. *Journal of Pragmatics*, *37*(10), 1510–1536.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3-4), 169–200.

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

Fussell, S. R., & Krauss, R. M. (1989). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of Experimental Social Psychology*, *25*(3), 203–219.

Gentner, D., & Wolff, P. (1997). Alignment in the processing of metaphor. *Journal of memory and language*, *37*(3), 331–355.

Gibbs, R. (2002). A new look at literal meaning in understanding what is said and implicated. *Journal of Pragmatics*, *34*(4), 457–486.

Gibbs, R., & Colston, H. (1999). Figurative language. *The MIT encyclopedia of the cognitive sciences*, (pp. 314–315).

Gibbs, R. W. (1984). Literal meaning and psychological theory. *Cognitive science*, *8*(3), 275–304.

Gibbs Jr, R. W. (1992). When is metaphor? the idea of understanding in theories of metaphor. *Poetics Today*, (pp. 575–606).

Gildea, P., & Glucksberg, S. (1983). On understanding metaphor: The role of context. *Journal of Verbal Learning and Verbal Behavior*, *22*(5), 577–590.

Giora, R. (1997). Understanding figurative and literal language: The graded salience hypothesis. *Cognitive linguistics*, *8*, 183–206.

Glucksberg, S. (1989). Metaphors in conversation: How are they understood? why are they used? *Metaphor and Symbol*, *4*(3), 125–143.

Glucksberg, S. (2001). *Understanding figurative language: From metaphors to idioms.*. Oxford University Press.

Glucksberg, S. (2003). The psycholinguistics of metaphor. *Trends in cognitive sciences*, *7*(2), 92–96.

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological review*, *97*(1), 3.

Glucksberg, S., & Keysar, B. (1993). How metaphors work.

Goodman, N. D., & Lassiter, D. (2014). Probabilistic semantics and pragmatics: Uncertainty in language and thought. *Handbook of Contemporary Semantic Theory, Wiley-Blackwell*.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Grice, H. P. (1975). Logic and conversation. *The Semantics-Pragmatics Boundary in Philosophy*, (p. 47).

Horn, L. R. (2006). Implicature. *Encyclopedia of Cognitive Science*.

Jones, L. L., & Estes, Z. (2006). Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language*, *55*(1), 18–32.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Kintsch, W., & Bowles, A. R. (2002). Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor and symbol*, *17*(4), 249–262.

Kreuz, R. J., & Roberts, R. M. (1993). The empirical study of figurative language in literature. *Poetics*, *22*(1), 151–169.

Lakoff, G., & Turner, M. (2009). *More than cool reason: A field guide to poetic metaphor*. University of Chicago Press.

Lakoff, G., et al. (1993). The contemporary theory of metaphor. *Metaphor and thought*, *2*, 202–251.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites*, vol. 1. Stanford university press.

Lassiter, D., & Goodman, N. D. (2014). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Proceedings of SALT*, vol. 23, (pp. 587–610).

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press.

Li, L., & Sporleder, C. (2010). Using gaussian mixture models to detect figurative language in context. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (pp. 297–300). Association for Computational Linguistics.

Luce, R. D. (2005). *Individual choice behavior: A theoretical analysis*. Courier Dover Publications.

McCarthy, M., & Carter, R. (2004). there's millions of them: hyperbole in everyday conversation. *Journal of pragmatics*, *36*(2), 149–184.

Ortony, A. (1975). Why metaphors are necessary and not just nice. *Educational theory*, *25*(1), 45–53.

Ortony, A. (1993). *Metaphor and thought*. Cambridge University Press.

Papafragou, A. (1996). Figurative language and the semantics-pragmatics distinction. *Language and Literature*, *5*(3), 179–193.

Pilkington, A. (2000). *Poetic effects: A relevance theory perspective*, vol. 75. John Benjamins Publishing.

Riloff, E., Wiebe, J., & Phillips, W. (2005). Exploiting subjectivity classification to improve information extraction. In *Proceedings of the National Conference On Artificial Intelligence*, vol. 20, (p. 1106). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

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

Roberts, R. M., & Kreuz, R. J. (1994). Why do people use figurative language? *Psychological Science*, *5*(3), 159–163.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Searle, J. (1979). Metaphor in ortony, a.(1979) metaphor and thought.

Sperber, D., & Wilson, D. (2008). A deflationary account of metaphors. *The Cambridge handbook of metaphor and thought*, (pp. 84–105).

Sperber, D., Wilson, D., He, Z. ·., & Ran, Y. ·. (1986). Relevance: Communication and cognition.

Stalnaker, R. (2002). Common ground. *Linguistics and philosophy*, *25*(5), 701–721.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*, vol. 1. Cambridge Univ Press.

Taylor, J. R. (2003). *Linguistic categorization*. Oxford University Press.

Tendahl, M., & Gibbs, R. W. (2008). Complementary perspectives on metaphor: Cognitive linguistics and relevance theory. *Journal of Pragmatics*, *40*(11), 1823–1864.

Tessler, M., & Goodman, N. D. (????). Some arguments are probably valid: Syllogistic reasoning as communication.

Thibodeau, P. H., & Boroditsky, L. (2011). Metaphors we think with: The role of metaphor in reasoning. *PLoS One*, *6*(2), e16782.

Tourangeau, R., & Sternberg, R. J. (1981). Aptness in metaphor. *Cognitive psychology*, *13*(1), 27–55.