MODELING CREATIVE AND SOCIAL USES OF LANGUAGE

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF PSYCHOLOGY
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Justine T. Kao

April 2016

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Noah D. Goodman) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Herbert H. Clark)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Christopher Potts)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Michael C. Frank)

Approved for the Stanford University Committee on Graduate Studies

Abstract

People use language to accomplish many kinds of goals, ranging from communicating information, to expressing personal attitudes, to entertaining and connecting with those around them.

This thesis examines the computational basis for how people understand figurative and creative language as well as the social motivations for using language in unconventional ways. In the first half of the thesis, I describe a computational model that formalizes principles of communication to predict people's interpretations of diverse types of figurative language, such as hyperbole, irony, and metaphor. In the second half, I focus on social motivations for using creative language, such as to highlight shared beliefs and to evoke humor. I present a computational model that predicts social inferences licensed by figurative language as well a model that explains the humor in word play.

Acknowledgments

Thank you for reading this thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Human beings are creatures of play. We begin playing from the cradle, from our first box of legos to the latest games on our smartphones. Play has the magical ability to exercise our imaginations and to satisfy our desire to explore, create, and interact with the physical and social world. As we develop, the games we play increasingly involve abilities that are socially valued and beneficial to our success: physical coordination, logic, planning, social reasoning, cooperation, and even deception. Despite its association with levity and fun, play is seriously and profoundly useful, enabling us to practice a range of skills that prepare us for situations in the real world.

Given the link between play and competency, it is perhaps not a coincidence that one of our favorite things to play with is also one of our most important skills—language. From nursery rhymes to puns to metaphors and poetry, human beings are exposed to and generate various forms of creative and playful uses of language throughout development (Ely & McCabe, 1994; Carter, 1999; Carter & McCarthy, 2004; Cook, 1996; Cazden, 1974). It is difficult to imagine a world in which the sole purpose of language is to transmit objective information to one another. In such a world, we would still be able to describe events, record facts, share knowledge, and work together to accomplish feats ranging from assembling Ikea furniture to curing diseases and ending wars. However, we would probably all suffer from immeasurable boredom. Thankfully, in the world we live in, using language involves much more than passing definitions, descriptions, and commands to each other and processing them like lines of code in a computer program. We use language to share opinions and perspectives, to entertain and be entertained, to create and appreciate works of art—to

play. The power and appeal of language lies not only in its ability to communicate raw information, but also to engage and connect us at a fundamentally human level.

The observation that language is used for purposes beyond purely referential functions is far from a novel one. Philosophers, literary theorists, linguists, and psychologists have been interested in the emotive, social, and aesthetic functions of language for decades and have addressed them from various angles using a diverse set of methods (Aristotle, 1911; Holtgraves, 2013; Halliday, 1971; Potts, 2007; Clark, 1996; Carter, 1999; Hall, 2001; Cook, 1996; Friedrich, 1979; Tannen, 2007; Veale, 2012). In his essay "Closing statement: Linguistics and poetics," linguist and literary theorist Roman Jakobson (1960) wrote: "Language must be evaluated in all the variety of its functions." However, historically the more intangible functions of language have not been the primary focus of a great deal of empirical work. In particular, approaches that involve formal and quantitative tools have largely avoided the playful, emotive, and artistic uses of language because they tend to be more variable, more ambiguous, more subjective, and rely more heavily on sources of information that are external to the language itself (Joos, 1950). In order to make progress on understanding how language and communication work, many scientists and theorists have focused on core or idealized aspects of language that seem more amenable to formal rigor, and focused less on applying the same tools to examine higher-order effects such as linguistic creativity, humor, and aesthetics (Clark, 1997). There is also a prevailing notion that creative language and creativity in general resist rules and quantification, and that the quest for a formula for creative uses of language can only be a fruitless endeavor (Giesen, 2000; Boden, 2009; Veale, 2012).

Fortunately, recent advancements in computational tools and digital resources have blurred the lines between disciplines and introduced new methods to fields that traditionally rely on qualitative approaches. Computational tools are now used to extract patterns in complex data and to generate valuable insights in the humanities and social sciences, giving rise to new fields such as digital humanities and computational social science (Manovich, 2011; Berry, 2012; Jockers, 2013; Lazer et al., 2009; Wallach, 2016). In light of these advancements, the time is ripe for language scientists to broaden our scope and to begin understanding in a precise and quantitative manner the multitude of functions that language performs. This thesis aims to accomplish this goal by using computational models to explore the psychology of creative and social uses of language.

1.1 Taking language play seriously

Why study creative uses of language? Just as studying play in children sheds light on key aspects of cognitive and social development, studying the way that adults play with language may provide insight into how we comprehend language and communicate more generally. Creative uses of language are puzzling because they often deviate from or violate linguistic norms in interesting ways (Ricoeur, 1973; Maybin & Swann, 2007). Figurative language, for example, is intentionally used to communicate meanings that differ dramatically from its standard, literal meaning (e.g. "Juliet is the sun" means that Juliet is a beautiful woman, and not that Juliet is a sphere made of hot plasma) (Gibbs & Colston, 1999). These types of uses may be ideal for helping us examine key questions in language understanding, such as how we construct linguistic meaning from multiple sources of information, and how our language understanding mechanisms work to flexibly accommodate these types of uses and recover the appropriate meanings. Secondly, if we adopt the standard definition of creativity as the production of something both novel and useful (Mumford, 2003; Boden, 2009), a natural question to ask is what might the "uses" of creative language be. What are the consequences of language play, and how do these effects arise naturally through our cognitive and linguistic mechanisms?

In addition to answering these questions, a more practical reason to study linguistic creativity is that it is far from rare (Gibbs, 1994; Carter, 1999; Carter & McCarthy, 2004; Cook, 1996; Tannen, 2007; Maybin & Swann, 2007; Veale, 2012). Creative uses of language constitute the "long tail" of the distribution of language use, highly diverse but together making up a large portion of our everyday conversations as well as our most prized literary works. In order to understand the range of linguistic behaviors that humans demonstrate, we need theories that address the creative aspects of language along with its more standard bread-and-butter functions.

In this thesis, I use computational models to formalize theories of language understanding that aim to explain this long tail. The motivation for adopting a computational approach is twofold. From a scientific perspective, computational models require the model builder to define assumptions in precise terms and to specify interactions among various components of the model. This requirement provides a more rigorous way of describing the underlying theory without relying on vague definitions and hidden assumptions. Formalizing theories and implementing them in computer programs also makes it convenient for model builders to observe the model's behavior in various situations, simply

by providing the model with different inputs and parameter settings. Furthermore, model builders can remove or modify various parts of the model to determine which assumptions and components are necessary to produce outputs that match the observed data. Computational models are therefore useful tools for specifying, testing, and revising scientific theories (M. Frank, 2013). And from an engineering perspective, ideas that are implemented in computer programs are more easily translated into technologies that can be enjoyed and tested by users at a larger scale.

The majority of work on computational approaches to language falls within the fields of natural language processing (NLP) and natural language understanding (NLU), which use sophisticated statistical models to extract patterns from large corpora of natural language. While these techniques have been used to develop technologies that are indispensable to modern society, they are not designed to test theories grounded in psychological reality. Because our motivation is to examine the cognitive mechanisms that support creative language interpretation and appreciation, here we aim to more directly draw insights from psychology and linguistic theory, which we then validate using data from behavioral experiments.

This motivation leads us to the choice of structured Bayesian probabilistic models. These models are useful for specifying the structure and dependencies among variables of interest and using Bayesian inference to recover the values of unobserved variables (MacKay, 2003; Pearl, 2014). In recent years, Bayesian models have been highly productive for constructing and testing formal theories across a wide range of topics in psychology and cognitive science (Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Oaksford & Chater, 2001; Yuille & Kersten, 2006). For the specific purposes of this thesis, two features of these models are particularly desirable. First, the models allow us to explicitly represent and make use of people's rich prior knowledge, which seems intuitively necessary for understanding and appreciating creative uses of language. Second, the models are tolerant of uncertainty and produce graded results, which is a useful property given that creative language often involves ambiguities and subtleties that would be difficult to capture using hard-and-fast rules.

Finally, Bayesian models of language understanding assume that people use their powers of rational probabilistic inference to comprehend language, which is consistent with a long tradition of rational approaches to language comprehension. At the perceptual level, research has shown that people's perception of acoustic signals can be modeled as the rational integration of various cues

(Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Feldman, Griffiths, & Morgan, 2009; Kleinschmidt & Jaeger, 2015). At the sentence-processing level, there is strong evidence that language comprehenders integrate noisy evidence to infer coherent sentence meanings in a rational manner (Levy, 2008; Gibson, Bergen, & Piantadosi, 2013). At the pragmatics level, communication has long been viewed as having a rational basis, where the interlocutors assume each other to be rational, cooperative agents and produce and interpret utterances accordingly (Grice, 1975; García-Carpintero, 2001; Cohen & Levesque, 1985). Recently, a family of models called Rational Speech-acts (RSA) models has formalized the rational basis of communication using Bayesian models (M. C. Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013). At many levels of linguistic analysis, it appears that language users can be modeled as "ideal comprehenders" who use Bayesian inference to infer linguistic meanings given the available evidence.

One contribution this thesis makes is to show that the interpretation and appreciation of creative uses of language can be achieved through many of the same rational principles that govern standard language understanding. We describe three approaches that show how assumptions of rationality lead to socially meaningful and evocative interpretations of creative language. In the first approach, we formalize standard principles of communication using an extended version of the Rational Speech-acts model and show that it produces appropriate interpretations of creative language such as hyperbole, irony, and metaphor as well as their subtexts. In the second, we further extend the RSA model to include inferences about the interlocutors' beliefs about each other. This model predicts rich social inferences that are uniquely licensed by figurative uses of language. Finally, in the third approach, we adopt a simple model of sentence processing that integrates noisy linguistic evidence in a rational manner, and show that we can derive measures from this model that predict people's judgments of humor. Together, these three approaches produce fine-grained, quantitative predictions that may help advance our psychological and linguistic theories of a range of creative and social uses of language.

1.2 Overview

Earlier in this chapter, I posed two questions regarding creative uses of language: (1) How do we understand language that violates certain linguistic norms? (2) What are the motivations for and

consequences of using language in creative ways? This thesis is organized around these two motivating questions. The first half of the thesis examines the computational basis of how people understand nonliteral, figurative uses of language using pragmatic reasoning and principles of communication. The second half examines two social motivations for using creative language in detail: to foster social closeness and to evoke humor.

In Chapter 2, we present a case study of number words to show that basic principles of communication can explain how people arrive at nonliteral interpretations of language. We introduce and formalize a notion of the principle of relevance, where listeners assume that speakers choose utterances that are maximally informative with respect to their communicative goals. We extend the basic Rational Speech-acts (RSA) model to include inferences about speakers' communicative goals, in particular the goal to communicate the speaker's attitudes and emotions. We test the model on nonliteral uses of number words: hyperbole (e.g. saying "It's a million degrees outside" when it is 92 degrees) and pragmatic halo (e.g. saying "It's 70 degrees outside" when it is 72 degrees). We show that the extended RSA model predicts people's interpretations with high accuracy and captures the rhetorical effect of hyperbole, suggesting that this approach may shed light on the computational basis of nonliteral communication.

Chapter 3 shows that the model described in Chapter 2 can be generalized to explain other types of figurative language such as irony and metaphor. We start by reviewing several types of figurative language and articulate the need for a coherent, formal theory of figurative communication. We show that extending the space of communicative goals to include emotional valence and arousal allows the model described in Chapter 2 to capture ironic interpretations. We then show that the same model explains effects observed in the interpretation of metaphor and analogy, suggesting that viewing metaphors and analogies and communicative acts may contribute a different perspective to the existing literature. By testing our model on a broader set of figurative utterances, we provide evidence that diverse figurative interpretations can arise from the same basic principles of communication.

In Chapter 4, we describe an extension to the RSA model that incorporates uncertainty about speakers and listeners' beliefs about each other. By explicitly reasoning about these beliefs given various utterances, the model predicts rich social inferences such as the idea that speakers can use nonliteral utterances to communicate social intimacy with the listener. These inferences are supported by existing evidence in the literature. We show that our model provides a principled

explanation for how speakers may use figurative language to accomplish social goals such as highlighting shared beliefs.

Chapter 5 focuses on one important aspect of creative language: the ability to evoke humor. We use linguistic puns as a case study to examine this aspect in detail. We describe a simple model of sentence processing from which we derive quantitative measures of humor. We then show that the measures correlate with people's funniness judgments on a set of puns and regular sentences. This work provides empirical evidence that humorous language is characterized by internally incongruous meanings and suggests that the experience of humor may arise from general language processing strategies.

Finally, Chapter 6 summarizes and synthesizes the work described and suggests directions for future research.

Chapter 2

Nonliteral Understanding of Number Words

Human communication is rife with nonliteral, figurative uses of language, ranging from metaphor to irony to hyperbole. How do people go so far beyond the literal meaning of an utterance to infer the speaker's intended meaning? In this chapter, we focus on the nonliteral interpretation of number words, in particular hyperbole (interpreting unlikely numbers as exaggerated and conveying affect) and pragmatic halo (interpreting round numbers imprecisely). We provide a computational model of number interpretation as social inference regarding the communicative goal, meaning, and affective subtext of an utterance. We show that our model predicts humans' interpretation of number words with high accuracy, incorporating principles of communication and empirically measured background knowledge to quantitatively predict hyperbolic and pragmatic halo effects in number interpretation*.

2.1 Introduction

Imagine a friend describing a new restaurant where she recently dined. Your friend says, "It took 30 minutes to get a table." You are likely to interpret this to mean she waited approximately 30 minutes. Suppose she says: "It took 32 minutes to get a table." You are more likely to interpret this to mean exactly 32 minutes. Now, suppose she says: "It took a million years to get a table." You

*This chapter is based on J. T. Kao, Wu, Bergen, and Goodman (2014)

will probably interpret this to mean that the wait was shorter than a million years, but importantly that she thinks it took much too long. One of the most fascinating facts about communication is that people do not always mean what they say; speakers often use imprecise, exaggerated, or otherwise literally false descriptions to communicate experiences and attitudes, and a crucial part of the listener's job is to understand an utterance even when its literal meaning is false. People's ability to interpret nonliteral language poses a critical puzzle for research on language understanding.

A rich body of literature in psychology and linguistics has examined how people use and understand nonliteral language (R. M. Roberts & Kreuz, 1994; Dews & Winner, 1999; Glucksberg, 2001; Gibbs & Colston, 1999). However, most of the work has been qualitative, with little focus on analyzing aspects of an utterance that predict the quantitative details of people's figurative interpretations. Here we present a computational model that formalizes and integrates three general principles of language and communication to explain the basis of nonliteral language understanding. First, speakers and listeners communicate with the assumption that their interlocutors are rational and cooperative agents; second, listeners assume that speakers choose utterances to maximize informativeness with respect to their communicative goals; third, speaker and listener utilize common ground—their shared knowledge of the world—to communicate effectively. The first principle has been formalized by a recent body of work on Rational Speech-acts (RSA) models, which views pragmatic language understanding as probabilistic inference over recursive social models and explains a range of phenomena in human pragmatic reasoning (M. C. Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Bergen, Goodman, & Levy, 2012; Jäger & Ebert, 2009). We go beyond the previous formal work and address the second principle by extending the RSA framework. We first extend the space of potential interpretations to include subjective dimensions such as affective opinion. We then assume that the listener is uncertain about the speaker's communicative goal and jointly infers both the goal and the intended meaning. Since the interpretation space has multiple dimensions, a speaker's goal may be to maximize the probability of successfully conveying information along one dimension of meaning but not another. This makes it possible for a literally false utterance to be optimal as long as it is informative along the target dimension. These elements of the model have important connections to Gricean pragmatics (Grice, 1975; Clark, 1996) and relevance theory (Sperber, Wilson, & Ran, 1986), in particular the argument that listeners infer the meaning of metaphors as well as other forms of loose talk by assuming that speakers maximize relevance

(Wilson & Carston, 2006; Sperber & Wilson, 1985). Finally, we address the third principle of communication by empirically measuring people's background knowledge to understand the interaction between nonlinguistic and linguistic knowledge in shaping language understanding. By applying this computational approach to a case study on number words, we show that nonliteral interpretations can arise from basic principles of communication without positing dedicated processing mechanisms for nonliteral language.

At the core of RSA models, a listener and a speaker recursively reason about each other to arrive at pragmatically enriched meanings. Given an intended meaning m, speaker S_1 reasons about a literal listener L_0 and chooses utterance u based on the probability that L_0 will successfully infer the intended meaning (Bergen et al., 2012):

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

Here $C(u)$ is the psychological cost of an utterance, potentially determined by factors such as the utterance's frequency, availability, and complexity. The exponential results from using a Luce-choice rule to model utterance choice, which is used extensively in models of decision-making (Sutton & Barto, 1998). A pragmatic listener L_1 then reasons about S_1 and uses Bayes' Rule to infer the meaning m given utterance u, where $P(m)$ is the prior probability of a meaning*:

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

Since the RSA framework operates under the assumption that speakers optimize informativeness, it predicts that choosing an utterance whose literal meaning directly contradicts the intended meaning is never optimal. However, this contradictory use is precisely the case in nonliteral language. For example, people understand the utterance "It took a million years to get a table" to mean that the wait time was long but not, in fact, a million years, resulting in a contradiction between literal and interpreted meaning. This suggests that the basic RSA model is incomplete and requires additional elements to explain nonliteral communication.

Previous work has examined people's communicative reasons for using figurative language and suggested that certain goals, such as conveying emotion and emphasis, are commonly satisfied by

*While in principle speaker and listener can recurse to arbitrary depth, here we stop at recursive depth 1.

nonliteral language (R. M. Roberts & Kreuz, 1994). A natural extension is thus to add an affective dimension to the meaning of utterances, which has interesting connections to previous work on expressives (Potts, 2007). However, simply adding this dimension is insufficient; it is still unclear how people infer affect from an utterance whose literal semantics is unconnected to affect (such as number terms). Here we additionally extend the RSA framework to represent alternative communicative goals, such that a speaker can want to convey information about one dimension but not another. We show that the combination of these two extensions is sufficient to give rise to nonliteral understanding of language.

We explore the case where the interpretation space has two dimensions: the state of the world and the speaker's affect or opinion*. The speaker is now modeled as:

$$S_1(u|s,a,g) \propto \sum_{s',a'} \delta_{g(s,a)=g(s',a')} L_0(s',a'|u) \cdot e^{-c(u)} \tag{2.3}$$

where the intended meaning includes two dimensions s (the state of the world) and a (the speaker's affect). The function g projects the listener's inferred meaning onto relevant dimensions, meaning the speaker's communicative goal is to be informative (only) along this "topic" dimension. A literal listener interprets utterances literally without reasoning about the speaker, while a pragmatic listener performs joint inference on both the speaker's goal and her intended meaning:

$$L_1(s,a|u) \propto \sum_{g} P_S(s) P_A(a|s) P_G(g) S_1(u|s,a,g) \tag{2.4}$$

The listener utilizes nonlinguistic background knowledge of the probability of a state (P_S) and the probability of having a particular affect given a state (P_A), which we measure empirically (see Experiment 3a and 3b). Based on the listener's linguistic knowledge, the literal semantics of utterance u conveys information about state s and nothing about affect a. However, the common knowledge that affect is usually associated with certain states of the world allows the listener to believe information about a given an assertion about s. If it is known that the speaker's goal is to convey affect, and not the state, then the pragmatic listener will discount information about s but retain information about a—a nonliteral interpretation is obtained. Even when the pragmatic listener is

*In what follows we describe the subtext dimension as "affect," but it could be other kinds of speaker attitude, *mutatis mutandis*.

not certain of the speaker's goal, a joint inference of goal, state, and affect can also result in nonliteral interpretation. Common knowledge of a domain and joint reasoning about communicative goals thus allows the speaker to communicate additional dimensions of meaning without explicitly describing these dimensions.

The incorporation of goal inference and multiple dimensions of meaning is a major change to the existing RSA framework that critically allows it to accommodate nonliteral language understanding. As a case study, we focus on the interpretation of number words. We chose number words because they have precise literal meanings that can be easily modeled, and apply to domains (such as prices) that lend themselves to quantitative measurement. We aim to capture two well-known phenomena regarding number interpretation: hyperbole and pragmatic halo. Hyperbole is a figure of speech that uses exaggeration to convey emphasis and emotion (McCarthy & Carter, 2004). Despite being literally false, hyperbolic utterances are readily understood and serve purposes such as establishing social closeness and expressing opinions (R. M. Roberts & Kreuz, 1994; McCarthy & Carter, 2004; Gibbs, 2000; Gibbs & O'Brien, 1991). Pragmatic halo refers to people's tendency to interpret round numbers such as 100 imprecisely and sharp numbers such as 103 precisely (Lasersohn, 1999). The halo effect has been formalized in game theoretic models as a rational choice given different utterance costs and a possibility of pragmatic slack (Bastiaanse, 2009; Krifka, 2007). Other research has shown that speakers' tendency to choose simple number expressions decreases when more precise information is relevant to the listener (Der Henst, Carles, Sperber, et al., 2002), suggesting that higher-level pragmatic considerations such as communicative goals directly impact the production and interpretation of round versus sharp numbers. Our model uses alternative communicative goals coupled with differential utterance costs to model the pragmatic halo effect. We show that our framework for pragmatic inference makes quantitative predictions for both hyperbole and pragmatic halo in the interpretation of number words.

2.2 Materials and Methods

2.2.1 Model

Let u be an utterance. The meaning of u has two dimensions: the actual price state s and the speaker's affect a. We defined the set of price states $S = \{50, 51, 500, 501, 1000, 1001, 5000, 5001, 10000, 10001\}$.

We assumed that the set of utterances U is identical to S. We defined the set of affect states $A = \{0, 1\}$ (0 means no affect and 1 means with affect—this binarization is purely for simplicity). Given S and A, the set of possible meanings M is given by $M = S \times A$. We denote each meaning as s, a, where $s \in S$ and $a \in A$.

The speaker S_1 is assumed to be a planner whose goal is to be informative about a relevant topic. We write the goal and its topic as g. S_1 chooses utterances according to a softmax decision rule that describes an approximately rational planner (Sutton & Barto, 1998):

$$S_1(u|s, a, g) \propto e^{U_1(u|s,a,g)} \tag{2.5}$$

We wish to capture the notion that the speaker aims to be informative about a topic of discussion while minimizing cost. If the topic is represented by a projection $g : M \to X$ from the full space of meanings to a relevant subspace, then the speaker cares only about the listener's distribution over the subspace,

$$L_0(x|u) = \sum_{s',a'} \delta_{x=g(s',a')} L_0(s', a'|u). \tag{2.6}$$

Following the Rational Speech-acts model, we formalize informativity of an utterance as the negative surprisal of the intended meaning under the listener's distribution; here the listener's distribution over the topical subspace X. Hence:

$$U_1(u|s, a, g) = \log L_0(g(s, a)|u) - C(u), \tag{2.7}$$

where $C(u)$ represents the utterance cost. Substituting into equation 2.5, this gives:

$$S_1(u|s, a, g) \propto \sum_{s',a'} \delta_{g(s,a)=g(s',a')} L_0(s', a'|u) \cdot e^{-C(u)} \tag{2.8}$$

In our situations, the speaker may have the goal to communicate along the price dimension, affect dimension, or both. This gives three possible projections r:

$$r_s(s, a) = s$$
$$r_a(s, a) = a$$
$$r_{s,a}(s, a) = s, a.$$

The speaker may also want to communicate the price either exactly or approximately (we assume that no such distinction exists for affect, since we have already binarized it). When the speaker wants to communicate the price approximately, she projects numbers to their closest round neighbors. For example, such a speaker will represent the prices 51 and 1001 as 50 and 1000, respectively. This gives two projections (exact and approximate), f, defined as:

$$f_e(s) = s$$
$$f_a(s) = \text{Round}(s),$$

where Round(s) denotes the multiple of 10 which is closest to s. The two types of projections, f and r, can be composed to make the goal g of the speaker: $g(s, a) = r(f(s), a)$, which results in $2 \times 3 = 6$ possible goals (though note that $r_a(f_e(s), a)$ and $r_a(f_a(s), a)$ are equivalent).

A literal listener L_0 provides the base case for recursive social reasoning between the speaker and listener. L_0 interprets u literally without taking into account the speaker's communicative goals:

$$L_0(s, a|u) = \begin{cases} P_A(a|s) & \text{if } s = u \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

The pragmatic listener L_1 performs Bayesian inference to guess the intended meaning given the priors P_S and P_A and his internal model of the speaker. To determine the meaning, the listener will marginalize over the possible goals under consideration.

$$L_1(s, a|u) \propto \sum_g P_S(s) P_A(a|s) P_G(g) S_1(u|s, a, g) \tag{2.10}$$

The prior probability of s is taken from an empirically derived price prior P_S, and the probability of a given s is taken from an empirically derived conditional affect prior P_A (see Experiments 3a and 3b). The probability distribution P_G is defined to be uniform. We used $C(u) = 1$ when u is a round number (divisible by 10) and treated the sharp/round cost ratio as a free parameter that we fit to data (see Experiment 1). We obtained a posterior distribution for all possible meanings s, a given an utterance u. Raw data for model predictions are here*. Figure 2.2 shows the full posterior distributions for all utterances.

*http://stanford.edu/~justinek/hyperbole-paper/data/model-predictions.csv

2.2.2 Experiment 1: Halo and hyperbole

120 participants were recruited on Amazon's Mechanical Turk. We restricted participants to those with IP addresses in the United States (same for all experiments reported). Each participant read 15 scenarios in which a person (e.g. Bob) buys an item (e.g. a watch) and is asked by a friend whether the item is expensive. Bob responds by saying "It cost u dollars," where $u \in \{50, 50 \pm k, 500, 500 \pm k, 1000, 1000 \pm k, 5000, 5000 \pm k, 10000, 10000 \pm k\}$, where k was randomly selected from the set $\{1, 2, 3\}$ for each trial. We refer to this set of utterances as U. Given an utterance u, participants rated the probability of Bob thinking that the item was expensive. They then rated the probability of the item costing the following amounts of money: $50, 50 \pm k, 500, 500 \pm k, 1000, 1000 \pm k, 5000, 5000 \pm k, 10000, 10000 \pm k$, where k was randomly selected from $\{1, 2, 3\}$ for each trial. We refer to this set of prices as S. Ratings for each price state were on a continuous scale from "impossible" to "extremely likely," represented as real values between 0 and 1. There are a total of 30 possible trial configurations (3 Items $\times$ 10 Utterances). We randomized the order of the trials as well as the names of the buyers (same for all experiments). See stimuli for Experiment 1 here*.

We normalized participants' ratings across price states for each trial to sum up to 1. The average normalized ratings across participants for each item/utterance pair is shown in Figure 2.3, and the data can be found here†. To adjust for humans' biases against using the extreme ends of the slider bars, we performed a power-law transformation on the model's distribution: We multiplied the predicted probability for each meaning by a free parameter λ and renormalized the probabilities to sum up to 1 for each utterance. We jointly fit λ and the model's cost ratio C to optimize correlation with the behavioral data. The best fit was with $\lambda = 0.36$ and $C = 1.3$, resulting in a correlation of $r = 0.974$ (95% CI = $[0.9675, 0.9793]$). The range of cost ratios that produces correlations within this confidence interval is $[1.1, 3.7]$, which is quite broad, suggesting that the overall model fit is not very sensitive to the cost ratio. To further capture the details of the halo effect, we jointly fit λ and C within this range to a measure that is more sensitive to utterance cost: We computed the difference between the probabilities of exact versus fuzzy interpretations for each utterance, which gives us each utterance's bias towards exact interpretation. We then computed the difference in this bias for sharp versus round numbers, which gives us a "halo" score for each sharp/round pair. We fit λ and C to minimize the mean squared error between the model and humans' halo scores. We

*http://stanford.edu/~justinek/hyperbole-paper/materials/experiment1.html
†http://stanford.edu/~justinek/hyperbole-paper/data/experiment1-normalized.csv

found that the cost ratio that best captures the magnitude and pattern of the halo effect found in participants' data is 3.4, while $\lambda = 0.25$. This produces an overall correlation of 0.9677 with human data from Experiment 1. All figures and analyses that we report in the main text are with these parameter values.

For the analysis reported in Figure 2.6, we computed the probability of a participant interpreting an utterance u as hyperbolic by summing up ratings for each interpreted price state s where $u > s$. Since our analysis of hyperbole does not involve utterance costs, we collapsed across round and sharp versions of utterances and price states. For example, "1001" interpreted as 1000 does not count as hyperbole. Since 50 and 51 are the lowest available price states, the probabilities for hyperbolic interpretation of utterances "50" and "51" are 0. We computed the average probability of a hyperbolic interpretation across subjects for each utterance. We then showed the hyperbole effect with a linear regression model, using prior probabilities for the utterances' literal meanings as predictor and probabilities for hyperbolic interpretation as response. Results indicated that participants were more likely to interpret utterances as hyperbolic when their literal meanings have lower prior probabilities ($F(1, 10) = 44.06, p < 0.0001$). For Figure 2.7, we analyzed the pragmatic halo effect by computing each subject's bias for interpreting an utterance u exactly versus fuzzily. Bias was measured by subtracting the probability of a fuzzy interpretation from the probability of an exact interpretation. We then obtained the average bias for each utterance across subjects. We showed that the average bias for exact interpretation is significantly higher for sharp utterances than for round utterances ($F(1, 28) = 18.94, p < 0.001$).

2.2.3 Experiment 2: Affective subtext

160 participants were recruited on Amazon's Mechanical Turk. Each participant read 30 scenarios in which a person (e.g. Bob) buys an item that costs s dollars and is asked by a friend whether the item is expensive. Bob responds by saying "It cost u dollars," where $u \in U$ and $u \geq s$. Participants then rated how likely Bob thinks the item was expensive on a continuous scale ranging from "impossible" to "absolutely certain," represented as real values between 0 and 1. There are a total of 180 trial configurations (3 Items $\times$ 60 $\{u, s\}$ pairs where $u \geq s$). The stimuli for Experiment 2 can be found here*; the raw data here†. Since our analysis of affective subtext does not involve utterance

*http://stanford.edu/~justinek/hyperbole-paper/materials/experiment2.html
†http://stanford.edu/~justinek/hyperbole-paper/data/experiment2-raw.csv

cost, for the analyses reported in Figure 2.8 and 2.9, we collapsed round and sharp versions of each utterance and price state such that there are a total of 45 utterance/price state pairs under consideration. Utterances u for which $u = s$ are considered literal; utterances u for which $u > s$ are hyperbolic. For the analysis reported in Figure 2.9, we obtained average ratings of affect for each utterance given that it is literal or hyperbolic. A linear regression model showed that hyperbolic utterances are rated as having significantly higher affect than literal utterances across price states ($F(1, 25) = 12.57, p < 0.005$).

2.2.4 Experiment 3a: Price prior

To obtain people's prior knowledge of the price distributions for electric kettles, laptops, and watches, 30 participants were recruited from Amazon's Mechanical Turk. Each participant rated the probability of someone buying an electric kettle, laptop, and watch that cost s dollars ($s \in S$), without any linguistic input from the buyer. Ratings for each price state were on a continuous scale from "impossible" to "extremely likely," represented as real values between 0 and 1. The stimuli for Experiment 3a can be found here*. We normalized participants' ratings across price points for each trial to sum up to 1. The average normalized ratings for each item were taken as the prior probability distribution of item prices. These price distributions were used in the model as P_S to determine the prior probability of each price state. The normalized ratings can be found here†.

2.2.5 Experiment 3b: Affect prior

To obtain people's prior knowledge of the probability of affect given a price state, 30 participants were recruited from Amazon's Mechanical Turk. Each participant read 15 scenarios where someone had just bought an item that cost s dollars ($s \in S$) without any linguistic input from the buyer. They then rated how likely the buyer thinks the item was expensive on a continuous scale from "impossible" to "absolutely certain," represented as real values between 0 and 1. The stimuli for Experiment 3b is here‡. The average ratings for each price state were taken as the prior probability of an affect given a price state and used in the model as P_A. The data can be found here§.

*http://stanford.edu/~justinek/hyperbole-paper/materials/experiment3a.html
†http://stanford.edu/~justinek/hyperbole-paper/data/experiment3a-normalized.csv
‡http://stanford.edu/~justinek/hyperbole-paper/materials/experiment3b.html
§http://stanford.edu/~justinek/hyperbole-paper/data/experiment3b-raw.csv

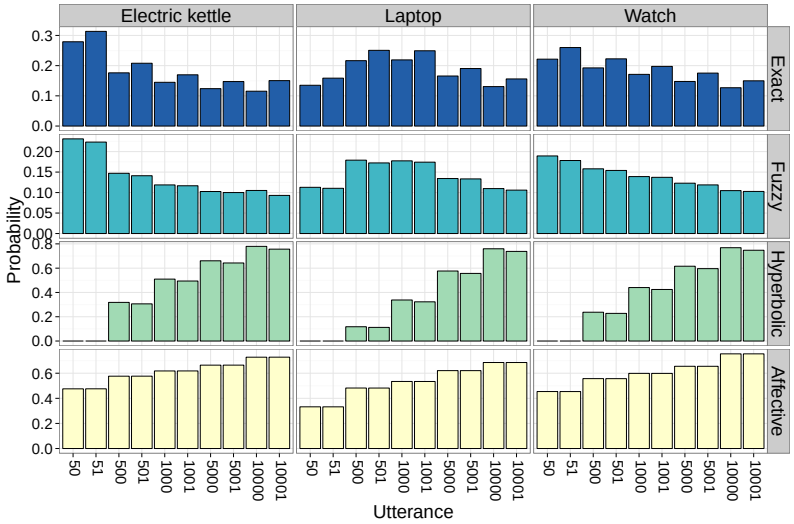

Figure 2.1: Model predictions of interpretations given utterances. Each bar in the first three rows shows the probability of a type of interpretation given an utterance. Exact interpretations are more likely given sharp rather than round utterances; fuzzy interpretations are slightly more likely given round utterances; hyperbolic interpretations are more likely given more extreme utterances. The final row shows the probability of an affective interpretation.

2.3 Results

We tested our model on number words that refer to the prices of three types of everyday items: electric kettles, watches, and laptops. We selected these items because they have distinct price distributions, P_S, which we measured empirically by asking participants to rate the probability of various prices for the three items (see Experiment 3a). We also obtained an affect prior, P_A, by asking participants to rate the probability of a speaker thinking that an item is too expensive given a price state (Experiment 3b). Using these priors, which capture purely nonlinguistic knowledge, we aimed to model people's interpretations of utterances such as, "The electric kettle cost u dollars."

2.3.1 Model simulations

Using the price priors and affect priors measured for each of the three items, we obtained the meaning distributions predicted by the model for all utterances (see Figure 2.2). Figure 2.1 summarizes this distribution into different types of interpretations. The first three are model interpretations regarding the price state: exact (e.g., "1000" interpreted as 1000), fuzzy (e.g. "1000" interpreted as 1001), and

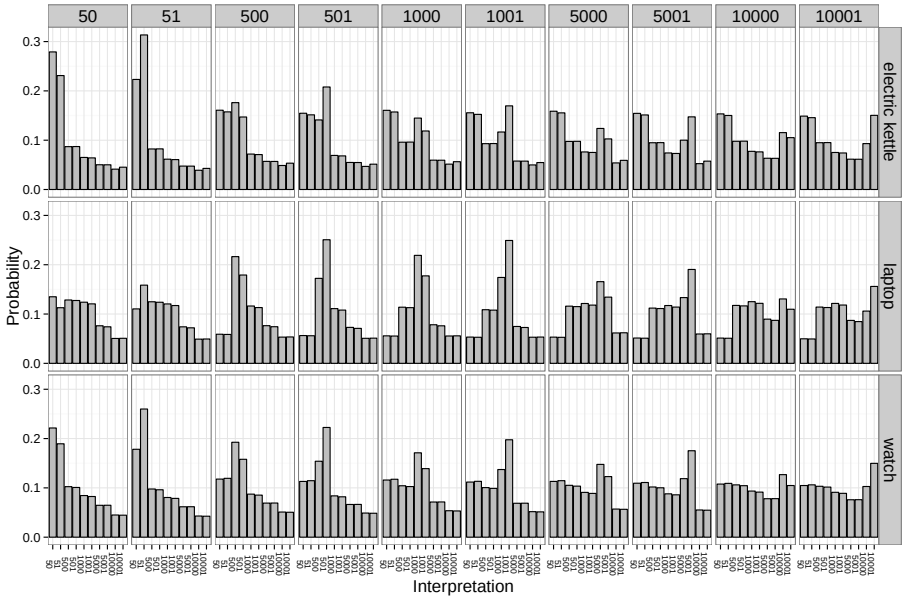

Figure 2.2: Posterior price state distributions predicted by the model given utterances. Each panel shows the interpretation distribution of an utterance.

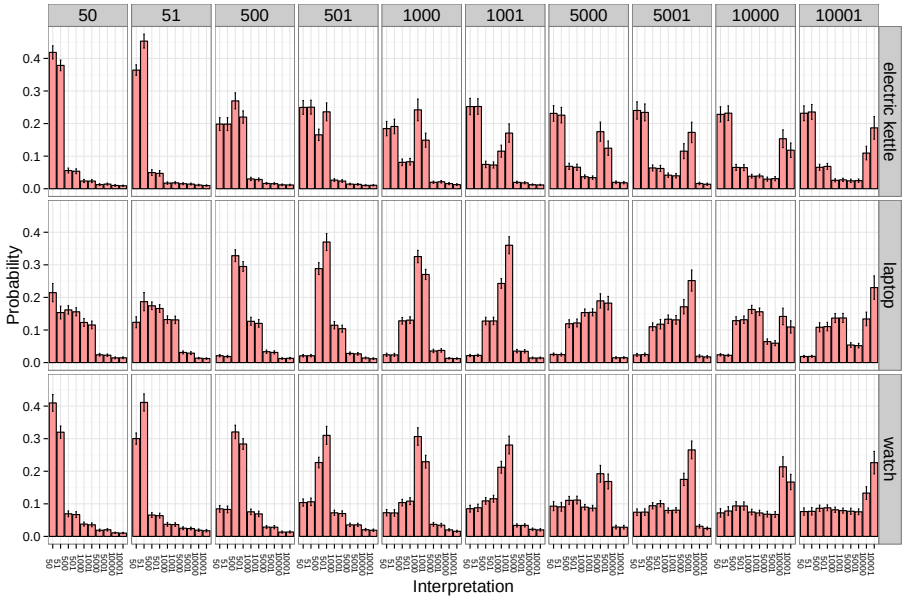

Figure 2.3: Price state distributions rated by participants given utterances. Each panel shows the interpretation distribution of an utterance. Error bars are standard errors.

hyperbolic (e.g. "1000" interpreted as 100). Round utterances (divisible by 10) such as "500" and "1000" are interpreted less exactly and more fuzzily than their sharp counterparts, which captures pragmatic halo. Utterances whose literal meanings are less likely given the price prior are more likely to be interpreted hyperbolically (e.g. "1000" is more likely to be interpreted hyperbolically for electric kettles than laptops), which captures a basic feature of hyperbole. Affective interpretation refers to the probability that an utterance conveys the speaker's opinion that the price is expensive. Utterances whose literal meanings are associated with higher affect priors (such as "10000" and "10001") are more likely to be interpreted as conveying affect, which predicts the affective subtext of hyperbole.

To build intuition for these predictions, consider a pragmatic listener who reasons about a speaker and analyzes her choice of utterance. The pragmatic listener hears "10,000 dollars" and knows that its literal meaning is extremely unlikely. Given that the speaker reasons about a literal listener who interprets "10,000 dollars" literally and believes that the speaker very likely thinks it is expensive, "10,000 dollars" is an informative utterance if the speaker's goal is to communicate an opinion that the kettle is expensive (without concern for the actual price). Since the pragmatic listener uses this information to perform joint inference on the speaker's communicative goal and the meaning of the utterance, he infers that "10,000 dollars" is likely to mean less than 10,000 dollars but that the speaker thinks it is too expensive.

2.3.2 Behavioral experiments

We conducted Experiment 1 to evaluate the model's predictions for the interpreted price. Participants read scenarios in which a buyer produces an utterance about the price of an item he bought, for example: "The electric kettle cost 1000 dollars." Participants then rate the likelihood that the item cost s dollars for $s \in S$ (see Experiment 1). Figure 2.3 shows humans' interpretation distributions across all utterances. Participants were more likely to interpret utterances as hyperbolic when their literal meanings have lower probabilities under the item's prior price distribution ($F(1, 10) = 44.06, p < 0.0001$). To examine the halo effect, we computed the difference between the probability of an exact interpretation and the probability of a fuzzy interpretation for each utterance. This difference is signficantly smaller for round numbers than for sharp numbers ($F(1, 28) = 18.94, p < 0.001$), which indicates that round numbers tend to be interpreted

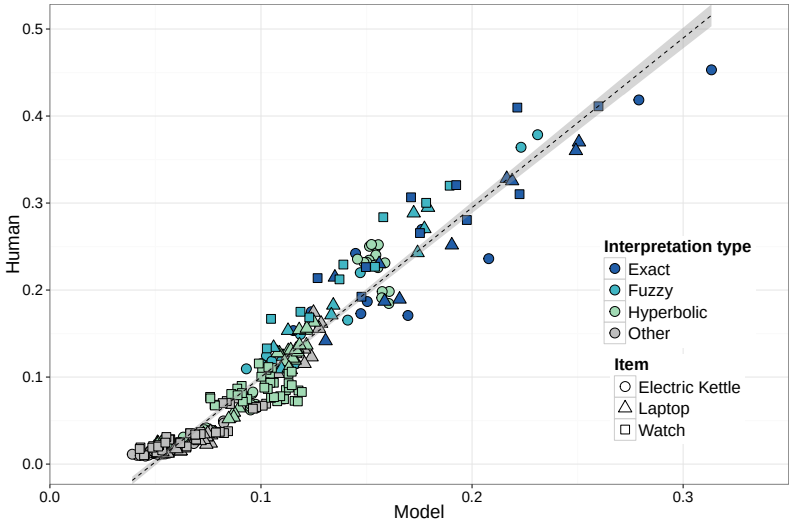

Figure 2.4: Model predictions v.s. average human responses from Experiment 1. Each point represents an utterance and price state pair (u, s). The x-coordinate of each point is the probability of the model interpreting utterance u as meaning price state s; the y-coordinate is the empirical probability. Correlation between model and human interpretations is 0.968 (95% confidence region in grey).

less precisely than sharp numbers. To quantitatively evaluate the model's fit, we compared model and human interpretation probabilities across all utterances and showed that model predictions are highly correlated with human interpretations of number words ($r = 0.968, p < 0.0001$) (Figure 2.4; see Materials and Methods for details).

To show how each component of the proposed model is necessary to capture effects observed in the human data, we explore a series of simpler comparison models. For illustration, Figure 2.5 compares model interpretations of the utterance "The electric kettle cost 1,000 dollars" given inference over different communicative goals. A model that does not consider alternative goals interprets the utterance entirely literally. Note that even though such a model has information about the affect dimension (i.e. P_A), without goal inference it is unable to produce nonliteral interpretations because it assumes that the speaker only wants to maximize informativeness along the same dimension as the utterance, i.e. the price state. A model that considers a speaker whose goal may be to communicate precisely or imprecisely interprets the utterance as meaning either 1000 or 1001. A model that considers a speaker whose goal may be to communicate the price state *or* her affect prefers price states with higher prior probabilities. Finally, a model that considers the full range of goals

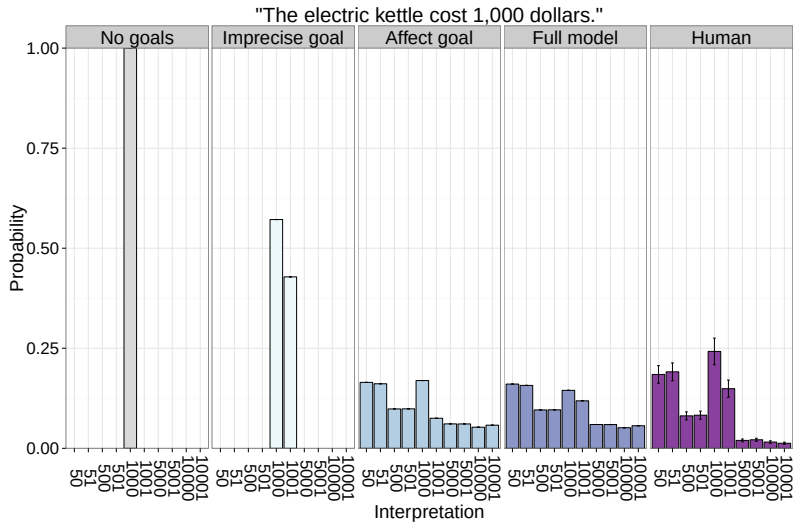

Figure 2.5: Comparison of models with different communicative goals and human interpretations for the utterance "The electric kettle cost 1,000 dollars." A model that considers both affect and precision goals (full model) most closely matches human data.

demonstrates hyperbole and halo effects that closely match humans' interpretations. To demonstrate that our model is able to usefully incorporate nonlinguistic knowledge to infer the meaning of utterances, Figure 2.6 shows the hyperbole effect as measured by the probability that an utterance u is interpreted as price state s such that $u > s$. A full model that uses empirically measured price priors captures humans' interpretations, while a model that takes a uniform distribution over price states does not. To demonstrate that our model is able to utilize utterance costs and goal inference to capture pragmatic halo, Figure 2.7 shows the halo effect as measured by the bias towards exact interpretation for sharp versus round numbers. A full model that assigns higher utterance costs to sharp numbers captures the significant difference in humans' biases for sharp versus round numbers, while a model where utterance costs are uniform does not. These analyses suggest that extending the RSA framework to include goal inference, incorporating empirically measured background knowledge, and including information about utterance costs all contribute to the model's ability to understand nonliteral language.

Does the model capture the rhetorical effect of hyperbole? We conducted Experiment 2 to examine humans' 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

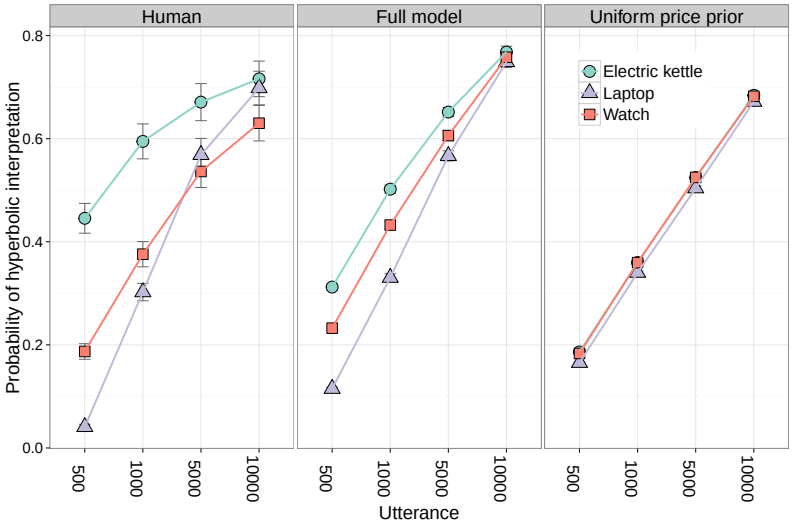

Figure 2.6: Probability of hyperbolic interpretation given utterances. Leftmost panel shows human data (error bars are standard errors). A full model that uses price priors measured in Experiment 3a demonstrates similar hyperbole effects and distinguishes among item types; a model that uses uniform price priors does not.

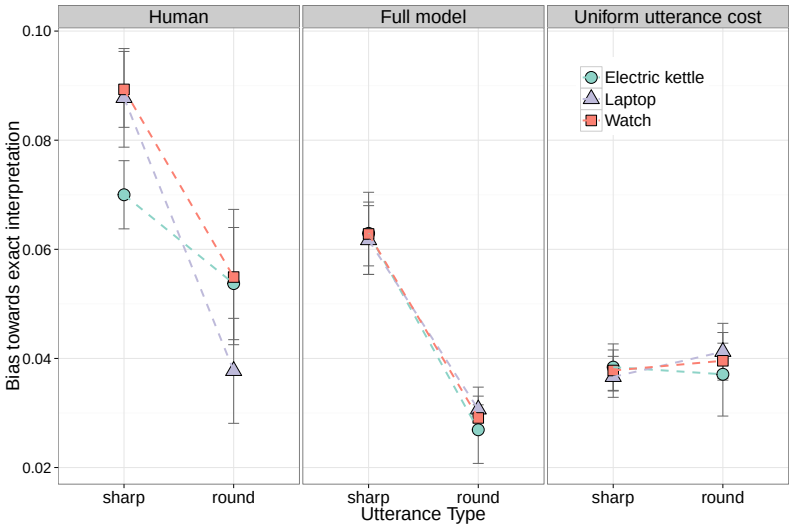

Figure 2.7: Halo effect as measure by bias towards exact interpretation for round/sharp utterance types. Humans' bias towards exact interpretation is significantly higher for sharp numbers. A full model that assigns higher cost to sharp numbers captures this result; a model that uses uniform utterance cost does not.

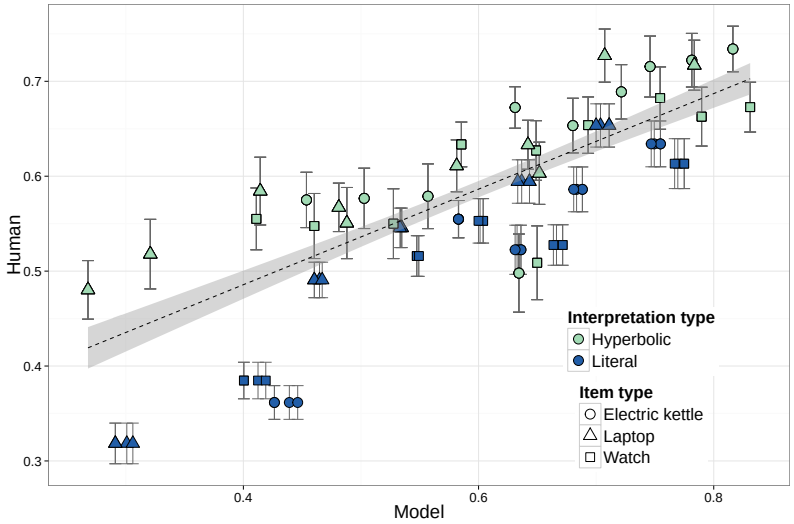

Figure 2.8: Model predictions of affect v.s. human responses from Experiment 2. Each point represents an utterance and price state pair (u, s). For pairs where $u = s$, the utterance is literal; for $u > s$, the utterance is hyperbolic. The x-coordinate of each point is the model's prediction of the 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 (95% confidence region in grey).

$u \geq s$. They then rate how likely it is that the buyer thinks the item was too expensive (see Experiment 2). We focused on the affect of an item being too expensive because previous findings suggest that hyperbole is more often used to communicate negative rather than positive attitudes (R. M. Roberts & Kreuz, 1994; McCarthy & Carter, 2004). Results showed that utterances u where $u > s$ are rated as significantly more likely to convey affect than utterances where $u{=}s$ ($F(1, 25) = 12.57, p < 0.005$). This suggests that listeners infer affect from hyperbolic utterances above and beyond the affect associated a priori with a given price state. Quantitatively, we compared model and human interpretations of affect for each of the 45 utterance and pice 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 2.8).

To demonstrate how our model explains this effect, Figure 2.9 shows probabilities of affect given a price state and a literal or hyperbolic utterance. The human data shows that higher actual price

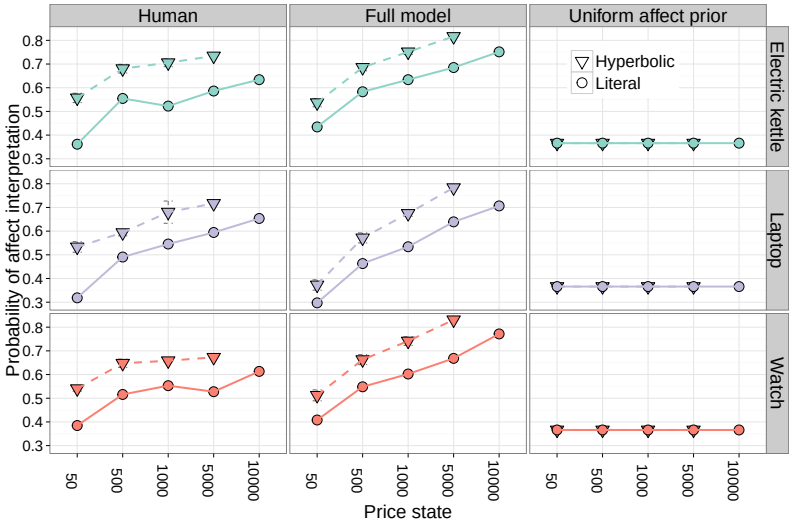

Figure 2.9: Probability of interpreting a hyperbolic/literal utterance as conveying affect. For the same price state, humans infer higher probability of affect given hyperbolic utterances than literal. A model that uses affect priors measured in Experiment 3b captures this result; a model that uses uniform affect priors does not.

states are associated with higher probabilities of affect. Within the same price state, hyperbolic utterances are interpreted as conveying more affect than literal utterances. These effects are replicated by the full model, but not by a model that takes in a uniform affect prior. This analysis suggests that the rhetorical effect of hyperbole is driven in part by people's shared knowledge about prices and associated affect.

2.4 Discussion

We presented the first computational model of nonliteral understanding that quantitatively predicts people's hyperbolic and imprecise interpretations of number words. Our behavioral results show that complex patterns in number interpretation depend on common knowledge between speaker and listener, consideration of communicative efficiency, and, critically, reasoning about the speaker's communicative goal. Our model represents an explicit, computational-level hypothesis about how these factors are integrated to give rise to the particular, graded interpretations that people arrive at. The model's quantitative predictions closely match humans' judgments, including cases of hyperbole, a complex phenomenon previously beyond the scope of computational models.

The current approach has important connections to theories of communication and linguistic meaning. Our speaker aims to be informative, as in Gricean theories of communication, but only with respect to a particular goal or topic—realizing a kind of relevance principle. This relevance is critical for deriving non-literal interpretations in our model. While our model is currently limited to two dimensions of meaning and corresponding goals, in future work we hope to capture dimensions central to other figures of speech such as irony and metaphor, thus extending our model to explain nonliteral language more broadly. We believe that our framework significantly advances the flexibility and richness of formal models of language understanding, such that some day probabilistic models will explain *everything* (hyperbolically speaking).

Chapter 3

Formalizing the Pragmatics of Figurative Language

In the previous chapter, we showed that a Rational Speech-acts model extended to accommodate inferences about the speaker's communicative goal* is able to predict people's interpretations of nonliteral interpretations such as hyperbole and pragmatic halo. In this chapter, we show that the model described can be generalized to capture interpretations of other distinct types of figurative uses such as verbal irony and metaphor. We argue that despite apparent differences among subtypes of figurative language, the same computational framework flexibly produces fine-grained interpretations for a range of figurative utterances. We 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 †.

3.1 Introduction

[ndg: why is figurative language important? very common, perhaps central to creativity, a key challenge to 'logic-like' views of mind, ...] The ability to understand figurative language is necessary in a world where people do not always mean what they say. We implicate, exaggerate, make

*In the previous chapter, we used the term "communicative goals" to refer to a topic of conversation that the speaker aims to address. In this chapter, we will consistently use the term "question under discussion (QUD)" to refer a very similar concept.

†This chapter is based on J. Kao and Goodman (2015) and J. T. Kao, Bergen, and Goodman (2014)

metaphors, and wax poetic. From "Juliet is the sun" to "It took a million years to write this paper," figurative language such as metaphor, irony, and hyperbole are commonplace in everyday communication, creating poetic or humorous effects that add richness to linguistic behavior (Glucksberg, 2001; Pilkington, 2000; Lakoff & Turner, 2009; R. M. Roberts & Kreuz, 1994). Although figurative statements are often false under their literal semantics (Juliet is not literally the sun, and it is infeasible to take a million years to write a paper), people are highly adept at inferring relevant and true information from these utterances (e.g. Romeo thinks Juliet is beautiful; it took unexpectedly long to write this paper). Because the literal meanings of these utterances are insufficient for uncovering the intended meanings, understanding figurative language requires integrating a host of information sources to create meaning. This ability to go beyond direct evidence (the words) to infer unobserved information (the meanings) is a hallmark of human intelligence that underlies many aspects of how we understand and interact with the world. How do our linguistic, cognitive, and social faculties work together to allow us to fluently and accurately understand the communicative intent behind figurative utterances? [ndg: this paragraph seems to say the same thing a few times, but also isn't very clear about what a literal or figurative meaning actually are. we don't actually want to define them yet, but we also don't want people to bump on this... maybe start with the observation that intended, and received, meanings often differ from overt meaning; this seems to be most radically true in the case of figurative language.]

An ocean of ink has been spilled attempting to answer this question, across many disciplines including psychology, linguistics, philosophy, computer science, and literary theory (Glucksberg, 2001; Papafragou, 1996; Li & Sporleder, 2010; Kreuz & Roberts, 1993). Much of the empirical research on figurative language focuses on cognitive mechanisms that underly interpretations of specific types of figurative use. For example, psychologists have proposed various ways in which people align shared properties and analogous relations across different domains in order to understand metaphor, including the domain interaction model (Tourangeau & Sternberg, 1982), structure mapping model (Gentner & Wolff, 1997; Gentner, 1983), and category assertion model (Glucksberg, 2003). To explain other types of figurative language separate explanations are posited, such as pretense (Clark & Gerrig, 1984) or allusion (Sperber & Wilson, 1981) in the case of verbal irony. While these approaches advance our understanding of the cognitive factors invoked by particular types of figurative language, they require a wide array of distinct mechanisms in addition to the process involved in

standard language understanding. Furthermore, these approaches leave open the question of how utterances, such as "My surgeon is a butcher" or "Such lovely weather we're having" (uttered in the middle of a storm), trigger the specialized mechanisms in the first place.

A different approach to studying figurative language focuses on how people use general communicative principles to arrive at contextually appropriate interpretations (Grice, 1975; J. Searle, 1979; Sperber & Wilson, 2008; Ortony, 1993; Tendahl & Gibbs, 2008). Two main theories in the pragmatics literature have taken this approach to explain figurative language: the standard pragmatic theory and relevance theory. The standard pragmatic view analyzes figurative utterances using the cooperative principle and standard Gricean maxims, which state that speakers tend to produce utterances that adhere to principles of quality (truthfulness), quantity (informativeness), relevance, and manner (e.g. brevity, orderliness, clarity) (Grice, 1975; J. Searle, 1979). Under this view, figurative utterances are understood through a three-step process: (1) determine the literal meaning of the utterance (2) determine whether the literal meaning violates the quality maxim by being untruthful (3) reanalyze the utterance to identify implied or figurative meanings that would allow the utterance to adhere to the Gricean maxims. Although the standard pragmatic view is appealing in that it fits parsimoniously within a more general theory of language understanding, it has met with several criticisms. One of the critiques is the fact that many figurative statements do not violate the quality maxim because their literal meanings can also be true. "No man is an island," for example, is a literally true statement (there does not exist a man that is literally a piece of land surrounded by water) in addition to a metaphorically meaningful one (people do not exist in isolation) (Gibbs, 1992). By relying on the violation of the quality maxim, the standard pragmatic view does not provide a satisfying explanation for how figurative meanings arise from these types of utterances. An even more common criticism in the psycholinguistics literature is that the standard pragmatic view requires the listener to first access the literal meaning of the utterance, verify that the literal meaning is false, compute potential figurative interpretations, and then select the interpretation that best satisfies conversational maxims. Given the extra steps involved, this model would predict that people should take longer to interpret figurative utterances than literal utterances. However, many experiments have shown that the figurative meanings of irony and metaphor can be accessed as quickly or even more quickly than their literal meanings given supporting contexts (Glucksberg, 2003; Gildea & Glucksberg, 1983; Gibbs, 1992). These empirical findings suggest

that literal meanings do not have to be explicitly computed and then rejected before appropriate figurative interpretations emerge.

Relevance theory, on the other hand, is a general theory of communication whose central claim is that human cognition is governed by a tendency to maximize relevance (Sperber et al., 1986). Instead of using Grice's four maxims as the guiding principle for figurative interpretation, relevance theory proposes that the principle of relevance is sufficient for explaining a range of phenomena in communication and cognition more generally. More specifically, the interpretation of all language involves maximizing the relevance of the interpretation to a contextually determined topic (Sperber & Wilson, 2008; Tendahl & Gibbs, 2008). As a result, interpretations of the same utterance can vary dramatically given different topics. Suppose two interlocutors, Ann and Bob, are discussing their friend Cam. Ann asks, "Does Cam have a fever?" and Bob replies, "He is boiling." Ann will interpret Bob's utterance to mean that Cam has a very high temperature. If, on the other hand, Ann asks, "Is Cam upset?" and Bob replies, "He's boiling," Ann should interpret Bob's utterance to mean that Cam is very angry. The word "boiling" receives different figurative meanings in these two contexts: in the first, Cam is very hot but not literally at boiling point; in the second, Cam is experiencing intense anger. Ann arrives at the appropriate interpretation by assuming that Bob's utterance provides maximally relevant information regarding her question. Under the relevance theoretic view, figurative uses such as hyperbole and metaphor are not distinct from literal language, but rather lie on a continuum of "loose uses" that all require listeners to use the principle of relevance to recover the intended meanings. This view situates figurative language within a general theory of communication and has been argued to provide a complementary perspective to cognitive linguistics in the study of metaphor (Tendahl & Gibbs, 2008). However, one concern with relevance theory is that the concept of relevance, while intuitively appealing, has not been clearly operationalized or tested in a quantitative manner to determine its specific role in figurative language understanding.

[ndg: maybe some of the details of standard and relevance theories (and arguments about them) should be saved for the next section?]

[ndg: we should argue that a precise, quantitative (and hence computational) approach is to be desired. these exist (somewhat) for the specific cases (eg SME) but not for the pragmatic theories.]

Taking the approach of analyzing general communicative principles that shape interpretations of figurative language, our goal here is to propose an explicit and testable theory of figurative language

understanding that can be validated against empirical data. We describe a computational model that integrates several pragmatic elements (e.g. assumptions that speakers are rational and cooperative; assumptions that utterances tend to be informative and relevant to the topic of conversation; representations of common knowledge and prior beliefs; and inferences about speakers' subjective attitudes) to produce appropriate interpretations of figurative utterances.

[ndg: give a bit more intuition about how our model will work. it formalizes the gricean view, but makes the topic explicit and central (a la relevance). the key insight is that by jointly inferring the topic and interpretation, a listener can deviate from literal meaning through the standard course of utterance interpretation.]

In what follows, we first review core empirical phenomena in figurative language, highlighting existing research and open questions regarding factors that shape figurative language understanding. Next, we describe the ways in which many of these factors can be integrated through a framework of pragmatic reasoning. We introduce Rational Speech-Acts (RSA) models, a family of computational models that formalizes communication as recursive social reasoning. We then show that natural but critical extensions can be made to RSA models to account for figurative language. We adapt the extended model to three types of figurative language: hyperbole, verbal irony, and metaphor, and present behavioral data and modeling results. Finally, we discuss the insights this model reveals about figurative language understanding as well as future research that our modeling approach licenses. We argue that a general model of figurative language enables us to more precisely examine the ways in which semantics and principles of communication interact to generate rich linguistic meaning.

Figurative Language: The Phenomena

Figurative language is often defined as utterances whose intended meanings differ in various ways from their "literal" or standard meanings (Gibbs & Colston, 1999, 2012). At first glance, this definition seems straightforward and corresponds with our intuitions regarding which usages of language are literal and which are figurative; however, it grows murky upon closer inspection. Suppose a speaker Bob says, "I arrived late and the theater was full." Since it is implausible that the entire space of the theater was occupied from floor to ceiling, the sentence's strict literal meaning appears to be false. Instead, Bob most likely intends to communicate that he had difficulty finding an empty

seat at the theater. This is an example of "loose talk," also described as pragmatic slack, where a speaker uses a proposition Q (e.g. the theater was full) in order to communicate a set of propositions that can be derived from Q (e.g. there were a lot of people at the theater, and Bob was unable to find a seat), without being committed to the truthfulness of Q (Sperber & Wilson, 1985; Lasersohn, 1999; Bach, 1994).

On the other hand, consider an utterance such as, "Bob is always late," produced by an annoyed speaker. In order for the literal meaning of the utterance to be true, for all cases in which Bob can be either late or on time, Bob must be late in 100% of the cases. However, one can easily interpret the utterance to mean that the speaker thinks Bob is very often (but not literally always) late, thus arriving at an interpretation that differs from the literal meaning of the utterance. Under these analyses, the intended meanings of both of these utterances ("The theater was full" and "Bob is always late") differ from their "literal" meanings. Indeed, Sperber and Wilson (1985) claim that there is no discontinuity between loose and figurative uses of language; both exploit the principle of relevance in order to express what is derivable from the utterance without committing to the truth of the literal meaning of the utterance. At the same time, a sentence such as "Bob is always late" feels qualitatively different from a sentence such as "The theater was full" and is more easily recognized as hyperbole. In fact, it has been observed that some utterances are intuitively recognized as "figurative" while other are not (Coulson & Oakley, 2005), which suggests that figurative language may be a psychologically meaningful category distinct from most other loose uses.

What factors, if any, distinguish figurative uses from loose uses? One distinction is that figurative utterances are often used with the intention to produce particular effects and in order to accomplish discourse goals beyond relaying objective information about the world. R. M. Roberts and Kreuz (1994) examined the discourse goals that motivate people to use various figurative tropes and identified a taxonomy of 19 discourse goals, such as to convey emotion, to emphasize, to be humorous, or to be eloquent. Colston and Keller (1998) showed that hyperbole and irony are often used to express surprise. In addition, hyperbole and irony are more often used with friends and may signal social intimacy between speaker and listener (Gibbs, 2000; Pexman & Zvaigzne, 2004; Kreuz, 1996). Other work has shown that verbal irony can heighten or soften criticism (Colston, 1997), elicit emotional reactions (Leggitt & Gibbs, 2000), highlight group membership (Gibbs, 2000), and express affective attitudes (Colston & Keller, 1998). Still other researchers have suggested that metaphors

are often used to express subjective attitudes towards the subject (Ortony, 1979), and that subjective sentences frequently contain figures of speech such as metaphor and hyperbole (Riloff, Wiebe, & Phillips, 2005). It is possible that the intuitive judgment of figurativeness involves recognizing elaborate that the speaker's intent is not to communicate objective information about the world, but rather to produce one or more of these rhetorical effects. As a result, it may be important to consider the affective subtexts and social information that figurative language communicates above and beyond most loose uses of language.

Despite many efforts to draw a distinction between literal and figurative language, the line remains blurry (Honeck, 1986; Coulson & Oakley, 2005). In fact, many researchers have argued that the line does not exist, partly due to the fact that literal meaning itself is not a single cohesive notion (Gibbs, 1994; Lakoff, 1986; Giora, 2002; Ariel, 2002). Instead of seeking to define the precise boundary between literal and figurative meanings, here we will focus on cases that are rather uncontroversially categorized as "figurative." In order to identify these cases, we first review the various types of language use that researchers have included within the category of figurative language and extract overlapping cases.

3.1.1 Types of figurative language

In part due to the difficulty of defining figurative language, researchers have not always agreed upon which figures should be included in the category of figurative language. Lanham (1991) created a list of nearly 1000 rhetorical terms; however, R. M. Roberts and Kreuz (1994) pointed out that many of these terms do not seem intuitively figurative, for example *apodioxis*, which means to indignantly reject an argument as false. Kreuz and Roberts (1993) instead identified eight figures of speech, which they believe form the basic categories of figurative language: *indirect requests*, which are commands phrased as comments or questions (e.g. "It would be great if you could keep this a secret"); *idioms*, where the intended meaning of the utterance cannot be derived from the individual words' typical meanings (e.g. "Ann ended up spilling the beans"); *irony*, where the intended meaning is opposite in polarity from the utterance's literal meaning (e.g. "Ann is the best secret keeper ever", in a situation where Ann clearly failed to keep a secret); *understatement*, where the speaker intentionally says something that is less extreme or intense than is actually the case (e.g. "Bob seems a tiny bit upset at Ann", when Bob is clearly furious); *hyperbole*, where a speaker

intentionally says something that is more extreme or intense than is actually the case (e.g. "Bob won't forgive Ann in a million years"); *metaphor*, where concepts from distinct domains are implicitly compared or equated with each other (e.g. "Bob's anger is a tornado"), *simile*, where concepts from distinct domains are explicitly compared (e.g. "Bob's anger is like a fire"), and *rhetorical questions*, which are questions that do not require an answer (e.g ."What was Ann thinking giving away that secret?"). Gibbs and Colston (1999) agreed with most of the figures while excluding *rhetorical questions* and including *metonymy*, *proverbs*, and *oxymora*. Based on these lists and on the amount of attention each figurative trope has received in the psycholinguistics literature, here we will focus on *hyperbole*, *irony*, and *metaphor* as three of the most central and broadly studied figurative tropes. Here we will describe each of the three tropes and review relevant theoretical and empirical research.

Hyperbole

A hyperbole is an exaggerated statement that purposefully presents its subject as more striking or extreme than it actually is (R. M. Roberts & Kreuz, 1994; McCarthy & Carter, 2004). Rhetoric studies in ancient Greece regarded hyperbole as a major figure of speech, often used to persuade and demonstrate power (J. Smith, 1969). 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 in humorous and other affective contexts. Norrick (1982) proposed that hyperbole is characterized by three properties: its affective dimension, its pragmatic nature, and its function as a vertical-scale metaphor where the comparison is between different positions on a scale rather than between discrete concepts. Gibbs (1994) makes a distinction between hyperbole and overstatement, where the former is purposefully produced for rhetorical effect. For a hyperbolic statement to be interpreted successfully, the listener must recognize the non-veridicality of the statement, thus entering an activity of joint pretense (Clark, 1996). Hyperbolic statements often include extreme case formulations (e.g. "It was the biggest storm in the history of the universe") or implausible descriptions (e.g. "It's a thousand degrees outside.") These demonstrations of non-veridicality require the listener to produce what Fogelin (2011) called a "corrective" response that is more in line with reality.

Verbal irony

An ironic statement describes something as contrary to what it actually is: for example, saying "Such beautiful weather we are having" in the middle of a storm (R. M. Roberts & Kreuz, 1994; Gibbs & Colston, 1999). Irony is thought to be related to hyperbole because it also involves a vertical scale (niceness of the weather), where the literal meaning's position on the scale ("beautiful") is different from the position of the intended meaning ("terrible"). Like hyperbole, irony also requires the listener to recognize the non-veridicality of the utterance and enter into joint pretense. However, the required corrective response is one of "kind" (e.g. from "beautiful" to "terrible") instead of degree (e.g. from "drizzling" to "pouring") (McCarthy & Carter, 2004). Clark and Gerrig (1984) propose the pretense theory of irony, where irony involves setting up a pretend world that is contrasted with the actual world to highlight the incongruity between what is and what might have been. Irony usually draws attention to this contrast and more often involves using a positive statement to express a negative attitude. Sperber and Wilson (1981) suggest that this asymmetry is due to the fact that irony is used to remind listeners of jointly held beliefs, social norms, or expectations that are being disrespected, which they call the echoic reminder theory . Since most social norms are positive, it follows naturally that ironic statements are often literally positive (e.g. "Such a fine friend you are") but express negative opinions (e.g. "You are not behaving as a good friend should"). Despite discrepancies among different theories of irony, they generally agree that irony relies heavily on using common ground—beliefs that are shared and known to be shared—to ensure that the listener produces a corrective response and recovers the speaker's intended meaning.

Metaphor

Metaphors are utterances that implicitly compare ideas or concepts from different domains. They are extremely prevalent in both literary and everyday language (Gibbs & Colston, 1999; R. M. Roberts & Kreuz, 1994). For example, "Juliet is the sun" expresses Juliet's beauty; "My lawyer is a shark" communicates the lawyer's ruthlessness; and "Art washes away from the soul the dust of everyday life" allows Picasso to compare "art" to a cleansing fluid and "the soul" to a physical object that collects dust, which gracefully accomplishes two poetic metaphors at once. One can find traces of metaphoricity even in mundane utterances such as "I waited for a long time," where the spatial term "long" is used to describe the abstract domain of time (Lakoff, 1993). Due in part to its

ubiquity and in part to the possibility that metaphor is intimately tied to our ability to create mappings between concrete experiences and abstract concepts (Lakoff & Johnson, 2008), metaphor is by far the most widely studied trope in cognitive science and related fields (Gibbs & Colston, 2012). Researchers have suggested that metaphor requires aligning analogical structures between two domains and can shape our reasoning and inferences (Gentner & Wolff, 1997; Thibodeau & Boroditsky, 2011). Evidence that metaphors are often processed as quickly as literal statements suggests that metaphor understanding does not require first accessing literal meanings or necessarily involve different processing mechanisms from literal language (Glucksberg, 2003; Gibbs & Colston, 2012).

Factors that shape figurative interpretation

In reviewing these three figurative tropes, some common features emerge. First, each example from these three tropes produces multiple interpretations that are distinct and highly different from each other (e.g. "It's a thousand degrees outside" (literal) v.s. "It's unexpectedly hot outside, like 90 degrees"; "The weather is amazing" (literal) v.s. "The weather is terrible"; "Juliet is made of hot plasma" (literal) v.s. "Juliet is beautiful"). Second, the intended meanings of these utterances are related to their "literal" meanings in non-arbitrary ways (e.g. a thousand degrees and 90 degrees are both unexpectedly high; "beautiful" and "terrible" both describe an extreme attitude towards the weather; the sun and Juliet are both very important and appealing to Romeo). Third, these utterances tend to express speakers' subjective experiences and attitudes rather than objective information about the world. Finally, a great deal of common ground is required to successfully interpret these utterances. For example, interpretation of an utterance such as "Such beautiful weather we are having," depends upon the speaker and listener's mutual beliefs about the relevant state of the world (e.g. it is raining), their shared background knowledge (e.g. sunshine is usually preferable to rain), and mutual awareness of potential discourse goals (e.g. the speaker wants to convey her opinion about the weather). Because the interpretation of such utterances depends upon these different flavors of common ground, it tends to be highly sensitive to changes in context. Here we will examine the various factors that together shape the interpretation of a figurative utterance in more detail.

Literal meaning

Although the relationship between a sentence's literal meaning and its intended meaning is not always clear, it is fairly uncontroversial that the intended meanings of utterances depend upon the literal semantics in a non-arbitrary manner (Coulson & Oakley, 2005). One cannot simply say *any* sentence and expect the context to make one's meaning clear (e.g. saying "I had eggs for breakfast today" in the middle of a storm and expect to be understood as expressing that the weather is terrible). Instead, the literal meaning of an utterance such as "Such beautiful weather we're having" contributes to the intended ironic meaning by drawing attention to the weather as well as the speaker's evaluation of it. The puzzle, then is *how* the intended meaning of a figurative utterance could be derived from its literal semantics.

add a note here about how the literal meanings we're looking at are rather simple and uncontroversial?

Encyclopedic knowledge

One way in which literal meaning gives rise to intended meaning is through encyclopedic knowledge, which includes a network of background knowledge shared among people in a community (J. R. Taylor, 2003; Langacker, 1987). Indeed, some researchers propose that the meaning of a word itself includes encyclopedic knowledge. J. R. Searle (1978) argued that literal meaning is not entirely independent of extra-linguistic information and instead relies heavily on this kind of encyclopedic knowledge. For example, the literal meanings of "Sally cut the cake" and "Sally cut the grass" depend on the manners in which cake and grass are usually cut, which is encoded in background encyclopedic knowledge (Gibbs, 1984). However, other linguists and philosophers argue that Searle "demands too much from literal meaning" and conflates the literal meaning of a sentence with its intended speaker meaning (Dascal, 1981; J. J. Katz, 1981; Gibbs, 1984).

Despite the fact that the distinction between literal meaning and encyclopedic knowledge is not always clear, encyclopedic knowledge often goes beyond the strict literal meanings of utterances to include stereotypes, conventions, and a community's beliefs and practices, which in turn shape the interpretation of language. For example, suppose Ann asks Bob, "Is Cam an honest person?" and Bob replies, "He's a politician." Ann will likely interpret Bob's utterance to mean no, he does not believe that Cam is an honest person. This interpretation arises because while the dictionary meaning of "politician" is "a person who is professionally involved in politics, especially as a holder of

or a candidate for an elected office," the encyclopedic meaning of the word can encompass many more features and connotations, such as dishonesty and corruption. Bob's utterance not only asserts Cam's profession (the literal, dictionary meaning of "politician"), but also attributes features associated with that profession to Cam (e.g. dishonesty, corruption). Ann is able to successfully interpret Bob's utterance, and Bob is able to successfully use this utterance, because they both have access to the relevant encyclopedic meaning of "politician." Naturally, Ann's interpretation is sensitive to the contents of the background knowledge they share. If Ann and Bob belong to a community where politicians are associated with honesty, then Ann would interpret Bob's reply to mean that yes, he believes Cam is an honest person. Similarly, "It's a thousand degrees outside" is interpreted as "It's unbearably hot outside" partly based on the encyclopedic knowledge that "a thousand degrees" is exceedingly hot, and that one is unlikely to survive under that temperature. As a result, the encyclopedic knowledge that interlocutors share can heavily influence the interpretation of figurative utterances.

Prior beliefs

In addition to encyclopedic knowledge, interpretation of language depends upon the listener's prior beliefs and expectations about various states of the world (Clark, 1991). Hörmann (1983) showed that people's interpretation of quantifiers such as "several" and "few" vary based on the kinds of objects to which they refer. For example, "several crumbs" is interpreted to mean around 10 crumbs, while "several mountains" is interpreted to mean around 5 mountains. Clark (1991) explains this phenomena using the "principle of possibilities:" to interpret language, people make use of their prior expectations about what situations or worlds are possible, as well as how likely those worlds are. To interpret "several crumbs" and "several mountains," people consider the number of crumbs and mountains that typically inhabit a scene or situation. Since a typical situation involving crumbs is likely to contain more crumbs than a typical situation involving mountains, the interpretation of "several" results in a higher number for "several crumbs" than in "several mountains." Given that prior beliefs affect the interpretation of superficially straightforward terms such as "several," it is unsurprising that prior beliefs factor into the interpretation of figurative language as well. In the dialogue between Ann and Bob, Ann's interpretation of the utterance "He's a politician" is sensitive to her prior beliefs about Cam. Suppose prior to her conversation with Bob, Ann did not know what

Cam does for a living. She will have learned two facts about Cam from Bob's utterance: Cam is a politician, and Cam is not an honest person. Suppose, on the other hand, that Ann knew beforehand that Cam is a politician, and knows that Bob knows that she knows that Cam is a politician. She will not have learned anything new about Cam's profession from Bob; however, even though she already knows that politicians in general are believed to be dishonest, Bob's utterance makes her more certain that Bob thinks Cam *in particular* is dishonest, because that is the most informative and relevant interpretation given her question about Cam's honesty and given that she already knows Cam's profession. Finally, suppose Ann knows that Cam is not professionally involved in politics at all. How will Ann interpret Bob's utterance? Instead of updating her beliefs about Cam using the dictionary meaning of "politician," she will rely on its encyclopedic meaning to conclude that Cam is dishonest (but not professionally involved in politics), resulting in a metaphorical interpretation. These examples show that interpretation of the same utterance in the same local context can vary in a rich and subtle manner based on the speaker and listener's prior beliefs.

Local context

A great deal of psycholinguistics research has shown that the interpretation of figurative utterances is highly sensitive to the local context (A. N. Katz & Ferretti, 2001; Giora, 2003; Coulson & Oakley, 2005). Within a discourse, context helps specify the topic of conversation as well as the particular communicative goals a speaker brings to a situation, which C. Roberts (1996) calls the "question under discussion" (hereafter QUD). Roberts argues that utterances are expected to be relevant to the QUD and are interpreted with respect to it. The QUD can be determined by an explicit question, for example Ann's question about Cam's honesty, which guides her interpretation of Bob's response because she expects Bob to communicate information that is relevant to her question. If, on the other hand, Ann had instead asked, "Is Cam a persuasive speaker?" then Bob's utterance may now be interpreted as a compliment: Cam is indeed a persuasive speaker (note, however, that in this case Bob's utterance still carries the connotation that Cam is not to be trusted, even though Ann's did not explicitly ask about Cam's honesty). Often, the QUD that a speaker's utterance addresses is not clearly specified to the listener and does not take the form of an explicit question. A speaker may produce an utterance in order to introduce a new QUD, which the listener must then infer based on the utterance itself as well as her expectations about which QUDs the speaker may plausibly wish

more examples from the literature?

to introduce. Given the importance of local context in shaping interpretation, a model of figurative language understanding should flexibly integrate this type of contextual information.

Pragmatic reasoning

A critical insight in communication is that a speaker does not produce utterances in a social vacuum; he considers the listener's beliefs, goals, and disposition to determine which utterance is most effective in a given situation, which Clark and Murphy (1982) termed "audience design." In turn, a listener considers the speaker's beliefs, goals, and disposition as well as the speaker's representation of *her* beliefs, goals, and disposition in order to select the most likely meaning of an utterance (Clark, 1996; Levinson, 2000; Grice, 1975). Furthermore, 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 recursive social 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).

Listeners can make many powerful inferences about utterances by representing speakers as rational and intentional agents who choose utterances in order to accomplish a specific communicative goal. Consider again the conversation between Ann and Bob. Ann may have several hypotheses about Bob's communicative goal and what QUD his utterance aims to address. Bob's goal could be to inform Ann about Cam's honesty, which is likely given Ann's question. His goal could be to inform Ann of Cam's profession, which is likely if Ann does not know Cam's profession, but less likely if Cam's profession is in common ground. Given each of these possible communicative goals, Ann can make inferences about what information Bob intends for her to glean from his utterance. The array of implicatures derived from a novel metaphor also depends on alternative utterances that the speaker could have said. The fact that Bob could have said "Yes, he's a persuasive speaker" but chose to say "He's a politician" makes it likely that Bob wants to communicate information beyond Cam's persuasiveness. Furthermore, the fact that Cam chose the metaphor "He's a politician" instead of "He's a salesman," both of which convey persuasiveness, suggests that Bob wants to communicate specific features about Cam such as deceptiveness and cunning, rather than pushiness.

Reasoning about the speaker's choice of utterance and available alternatives allows the listener to derive rich figurative meanings as well as their subtleties using basic principles of communication. A theory of figurative language as a communicative act should thus incorporate the speaker's intent as well as how the listener reasons about this intent in various communicative contexts.

Putting it all together

While many researchers have suggested that the construction of meaning involves an interplay of the components outlined above (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, quantitative, and fine-grained predictions that can be evaluated against empirical data. Here we propose a formal modeling framework for figurative language understanding that incorporates these components and captures the recursive social nature of communication. We show that these components are sufficient for producing appropriate interpretations of figurative utterances as well as rich affective and social subtexts.

elaborate on why it's useful to have a formal model

3.2 Probabilistic Models of Language Understanding

In recent years, a family of computational models have emerged that use probabilistic tools to formalize principles of communication, called Rational Speech-acts (RSA) models (M. C. Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Goodman & Lassiter, 2014). These models formalize the Cooperative Principle to explain how people arrive at pragmatically enriched meanings of utterances through recursive social reasoning. By representing listeners as agents who reason about the intentions of a rational and cooperative speaker, these models predict pragmatic enrichments that allow the listener to make inferences beyond the strict literal meaning of an utterance. To date, RSA models have 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, 2015), effects in syllogistic reasoning (Tessler & Goodman, under review), and more (Goodman & Lassiter, 2014).

add note that these components are not unique to figurative language understanding

The basic structure of RSA models is simple and usually involves three "agents:" a naive literal listener L_0, a speaker S_1, and a sophisticated, pragmatic listener L_1. S_1 reasons about L_0 and determines which utterance u to choose in order to efficiently communicate a meaning m to L_0. The

more sophisticated listener L_1 then reasons about which meaning m most likely led S_1 to choose u and uses Bayesian inference to recover m given u. More formally, the probability that S_1 will choose an utterance u given an intended meaning m is given by the following equation:

$$S_1(u|m) \propto L_0(m|u) \cdot e^{-Cost(u)} \tag{3.1}$$

where $L_0(m|u)$ is the probability that L_0 will arrive at meaning m given utterance u, and $Cost(u)$ is the psychological cost of producing utterance u given its length, difficulty, or availability. The term $e^{-Cost(u)}$ thus implements the Luce-choice rule, which is widely used to model rational decision-making (Luce, 2005). Using Bayes' rule to infer S_1's intended meaning given a generative model of S_1's utterance choice, L_1's interpretation distribution of u is given by the following equation:

$$L_1(m|u) \propto P(m)S_1(u|m) \tag{3.2}$$

where $P(m)$ is the prior probability of the meaning m, or how likely it is that meaning m is true. L_1 is thus used to model people's pragmatic interpretation of various utterances *.

M. C. Frank and Goodman (2012) tested the RSA framework on humans' pragmatic judgments in a simple reference game. In this paradigm, participants see three objects and are asked to choose which one the speaker is referring to (see Figure 3.1). The speaker can only use one word to identity the intended object, which often results in ambiguous references. For example, the word "blue" may refer to either the blue square or the blue circle in Figure 3.1. M. C. Frank and Goodman (2012) asked participants how likely a speaker is to use a particular word to refer to an object: for example, how likely a speaker is to use the word "blue" or "square" to refer to the blue square. This experiment yields the likelihood term $S_1(u|m)$. Other participants were asked how likely a speaker is to refer to a particular object using an unknown word, which measures $P(m)$, or what the authors refer to as the object's contextual salience. Using these two pieces of information, the RSA model computes $L_1(m|u)$, which is the probability that the referent is a particular object given a particular utterance. The model correctly predicts that listeners are more likely to judge the word "blue" as referring to the blue square, even though the word is technically ambiguous. This is

*In principle, the speaker and listener can recursively reason about each other to 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.

because a sophisticated listener who reasons about the speaker knows that if the speaker had meant the blue circle, he would have used "circle" instead because it is more informative. The model's predictions matched participants' judgments extremely well ($r = 0.99, p < 0.0001$), suggesting that people may be incorporating the speaker's choices and prior probabilities of meanings in a similar rational manner. Using a simple reference game paradigm, this work showed that incorporating recursive social reasoning and prior knowledge allows the listener to go beyond the strict literal meaning of a word to infer the intended meaning in context.

Goodman and Stuhlmüller (2013) made more explicit the fact that in addition to formalizing the rationality principle, the RSA model can also flexibly capture background knowledge and common ground. Imagine Bob has three apples, which Ann cannot see. Bob says, "Some of the apples are red." Ann makes the inference that *not all* of the apples are red, because if all of the apples are red, then Bob would have said "All of the apples are red" in order to be maximally informative. The pragmatic strengthening of "some" to "some but not all"—termed scalar implicature—can arise based on the same principles that allow a listener to infer *blue square* from "blue" in Figure 3.1 (M. C. Frank & Goodman, 2012). However, what happens when the speaker and listener both know that the speaker's knowledge of the world is incomplete? Suppose Bob can only see two of the three apples. To choose an utterance that is maximally informative, Bob needs to consider the possible states of the world and compute the expected utility of different utterances. His choice of utterance is captured with this equation:

$$S_1(u|s,a) = \sum_o S_1(u|o,a)P(o|a,s) \tag{3.3}$$

where u is the utterance, s is the true state of the three apples, a is Bob's perceptual access to the three apples, and o is what he observed. Given that Ann knows Bob's perceptual access to the apples, (i.e. a is common knowledge between Ann and Bob), her inference is captured by the following:

$$L_1(s|u,a) \propto S_1(u|s,a)P(s) \tag{3.4}$$

The model closely matches participants' interpretations of utterances given different combinations of observations and perceptual access ($r = 0.96$). This suggests that by explicitly incorporating common ground about what the speaker knows and does not know, listeners can interpret utterances

in principled ways even when the speaker has imperfect knowledge of the world.

While the RSA framework provides an intuitive and empirically validated way to model the interaction between literal meaning and background knowledge, it requires significant and theoretically important extensions to explain figurative communication. In most of the cases that RSA handles, the pragmatically strengthened interpretations produced by L_1 do not stray very far from the literal meanings of utterances. While interpreting "blue" to mean *blue square* requires pragmatic enrichment, the interpreted meaning is simply more specific than the literal meaning, and not distinct from the literal meaning as is the case in many figurative uses. This is because one of the key assumptions in the RSA model is that S_1 chooses an utterance that most efficiently communicates the intended meaning to L_0. Since L_0 interprets utterances literally, it is never optimal for S_1 to choose an utterance whose literal meaning directly contradicts the intended meaning. For example, suppose S_1 wants to communicate that the weather is terrible. 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. Because L_0 is a literal listener, she would interpret the utterance "The weather is amazing" to mean that S_1 believes the weather is literally amazing. She would thus *not* arrive at the interpretation that the speaker believes the weather is terrible. As a result, S_1 has no reason to say "The weather is amazing" to communicate that the weather is terrible (because L_0 would not receive the intended meaning). Consequently, a pragmatic listener who reasons about why the speaker chose various utterances will not interpret "The weather is amazing" to mean that the weather is terrible. The RSA model in its basic form is unable to explain many cases of figurative language use.

3.2.1 Rational Speech-acts Model with QUD inference

We extend the RSA framework to address the ways in literal meaning, encyclopedic knowledge, prior beliefs, and contextual information shape language understanding through reasoning about relevance to the QUD. The basic RSA models already naturally incorporate aspects of background knowledge and prior beliefs. For example, consider the utterance: "Cam is a wolf." To compute the probability that Cam is a wolf given this utterance, the pragmatic listener considers the prior probability of Cam being a wolf. However, believing that Cam is a wolf is more than believing that Cam is a large wild animal that often hunts in groups. Once you believe that Cam is a wolf, you are more likely to believe that Cam is furry, fierce, loyal, fast, hungry, etc. These beliefs are graded;

one 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 encyclopedic 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 easily accessible through association and are closely tied to the literal meaning. As a result, we assume that the literal listener L_0 also has access to these dimensions of meaning. Given a literal meaning l, associated encyclopedic meanings $\vec{E}$, and an utterance u, the literal listener's interpretation of u is now given by the following:

$$L_0(l, \vec{E}|u) = \begin{cases} P(\vec{E}|l) & \text{if } l \text{ is compatible with the literal meaning of } u \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

We thus provide a formal way of enriching literal meaning with encyclopedic knowledge. However, incorporating encyclopedic knowledge alone is insufficient for explaining figurative language understanding. Although the literal listener now has access to the associated encyclopedic meanings, she still assigns 0 probability to all interpretations that are incompatible with the literal meaning of the utterance. Given the utterance "Bob is a wolf," the literal listener will believe that Bob is a fierce, furry, and loyal wolf with some probability ($P(\vec{E}|l)$); 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 100%.

For figurative meaning to arise, the speaker and pragmatic listener must reason about which dimension of meaning is relevant to the 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. In other words, the speaker does not care about whether the literal listener derives true information regarding any of the other dimensions; she chooses an utterance only to maximize informativeness along the QUD dimension(s). This leads to the following utility function for speaker S_1:

$$U(u|l, \vec{E}, Q) = \log \sum_{l, \vec{E}} \delta_{Q(l,\vec{E})=Q(l',\vec{E}')} L_0(l', \vec{E}'|u) \tag{3.6}$$

Based on this utility function, the speaker's choice of utterance is specified by the following:

$$S_1(u|l, \vec{E}, Q) \propto e^{\lambda U(u|l,\vec{E},Q)}, \tag{3.7}$$

where λ is a rationality parameter that determines the speaker's tendency to choose the optimally informative utterance (Luce, 2005). Consistent with the basic RSA models, the pragmatic listener L_1 performs Bayesian inference to guess the intended meaning given prior knowledge and her internal model of the speaker. Since L_1 is uncertain about the precise QUD that the speaker is trying to address, she marginalizes over the possible QUDs under consideration:

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

This equation now includes multiple dimensions of meaning, the QUD, a model of the speaker's choice given he wants to be relevant to the QUD as well as informative, and the listener's prior beliefs. Something quite magical happens when all of these elements are combined, which we will illustrate with the example of "Cam is a wolf" and a set of QUDs that includes Cam's personality characteristics. Since the literal listener is likely to believe that Cam is fierce if she believes that Cam is a wolf, the speaker is motivated to say "Cam is a wolf" to get her to believe that Cam is a wolf and thus fierce. Furthermore, a speaker who only cares to communicate Cam's fierceness and not which species Cam belongs to will not mind that the literal listener will believe that Cam is actually a wolf. The pragmatic listener simulates the speaker's choice of utterance given different QUDs. Combining this simulation with the prior belief that Cam is very unlikely to actually be a wolf, the pragmatic listener ultimately believes that Cam 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 with encyclopedic knowledge, the RSA model is able to produce figurative interpretations of utterances that match our intuitions.

In what follows, we 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, we will show that the model captures several desired effects in the interpretation of *hyperbole*, *irony*, and *metaphor*: (1) figurative interpretation (2) sensitivity to encyclopedic knowledge (3) sensitivity to prior beliefs (4) sensitivity to utterance cost (5) sensitivity to local context (6) sensitivity to alternative utterances.

3.3 Modeling Figurative Language

Our first attempt at testing the qRSA model on figurative language focused on cases where the literal semantics are simple to quantify and relatively uncontroversial: number words. Although numbers have precise meanings in mathematics, they can be interpreted in various nonliteral ways in natural language. For example, "It's 90 degrees outside" is likely to be interpreted as approximately 90 degrees, while "It's 92 degrees outside" is more likely to be interpreted as exactly 92 degrees, an effect known as pragmatic halo. Even more dramatically, an utterance such as "It's 1000 degrees outside" is likely to receive a hyperbolic interpretation: it is very hot outside, but the temperature is much less than 1000 degrees.

In J. T. Kao, Wu, et al. (2014), we examined how people arrive at the appropriate interpretations and affective subtexts of numeric utterances about prices. To empirically measure people's prior beliefs, we asked participants to rate the probabilities that different items (electric kettles, watches, and laptops) cost various amounts of money (e.g. \$50, \$51, \$1,000, \$10,000). To measure people's encyclopedic knowledge in this domain, we asked participants to rate the probability that someone would think an item that costs $\$x$ is expensive (e.g., a watch that costs \$1,000). We chose expensiveness as the associated dimension of interest, because utterances about cost seem to naturally evoke judgments of expensiveness . Using the empirically measured prior beliefs and background knowledge, we used the qRSA model to obtain predicted interpretations for each utterance. The model reasons about different types of QUDs that the speaker may wish to address, including a QUD concerning affect about the price of the item. By reasoning about relevance to a set of QUDs, 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."

explain this more

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 ($r = 0.968, p < 0.0001$), suggesting that the qRSA model is able to combine linguistic information, background knowledge, and reasoning about the speaker's goals to interpret hyperbolic utterances.

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 $u{=}s$ (literal utterances) ($F(1, 25) = 12.57, p < 0.005$). Moreover, if a watch actually cost 100 dollars and Sam produces a hyperbolic utterance such as "The watch cost 1000 dollars," participants are more likely to believe that Sam thinks the watch is expensive than if the watch *actually* cost 1000 dollars and Sam produces an identical (but in this case literal) utterance: "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.

Results from (J. T. Kao, Wu, 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 very simple space of affect, namely the presence or absence of negative feeling. This simplification overlooks the range of attitudes and emotions that speakers could express using figurative utterances. 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.

3.3.1 Verbal Irony

An ironic statement describes something as contrary to what it actually is (R. M. Roberts & Kreuz, 1994; Gibbs & Colston, 1999). For example, a speaker who says "Such beautiful weather we are having" in the middle of a storm means that the weather is *not* beautiful 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, for example, showing participants a picture of a blue, cloudless sky and asking them to judge how likely it is that someone would perceive the weather to be amazing or terrible. Finally, given the critical role that context plays in understanding irony, we can vary the context (a picture of a gray, cloudy sky instead of blue sky) to observe how the same utterance is interpreted differently given different contextual knowledge.

In what follows, we first explore how an enriched space of affect affects the qRSA model and show that it produces ironic interpretations in a simple simulation. 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'.

Model

Following the qRSA model described previously, a speaker chooses an utterance that most effectively communicates information regarding the 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. Following the formulation described in the modeling section, 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:

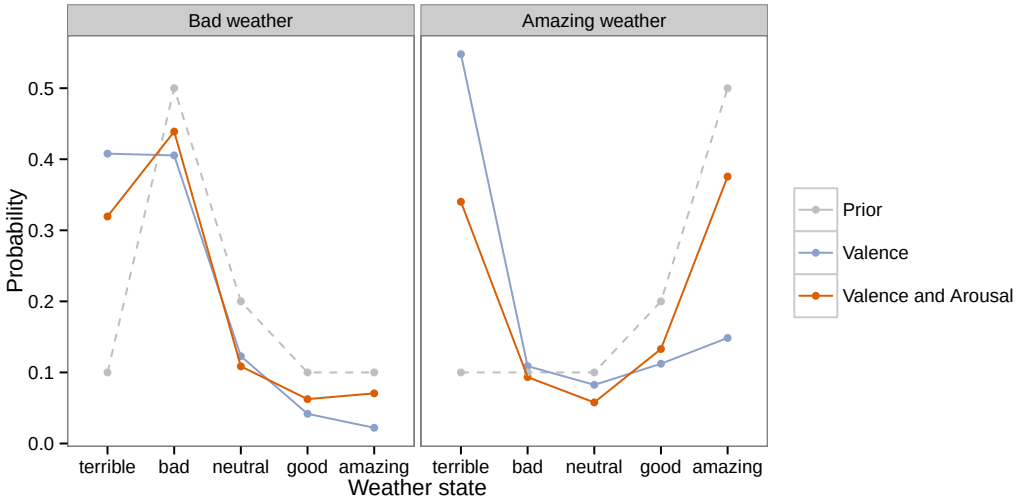

Figure 3.1: Model interpretations of "The weather is terrible" given different prior beliefs about the weather state and affect dimensions. Gray dotted lines indicate prior beliefs about weather states given a weather context; blue lines indicate interpretations when reasoning only about the speaker's valence; orange lines indicate interpretations when reasoning about both valence and arousal.

$$U(u|s,A,Q) = \log \sum_{s',A'} \delta_{Q(s,A)=Q(s',A')} L_0(s',A'|u) \tag{3.8}$$

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 choice of utterance u given state s, his affect A towards the state, and the QUD is then described by the following: $S_1(u|s,A,Q) \propto e^{\lambda U(u|s,A,Q)}$, where λ is the rationality parameter. A pragmatic listener L_1 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, marginalizing over the possible QUDs under consideration:

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

We characterize the interpretation of an utterance as the resulting posterior distribution over world states and speaker affects.

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 3.1). 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 3.1 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. This is because a pragmatic listener who 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). As a result, a pragmatic listener that only considers one dimension of affect–emotional valence—is unlikely to infer a positive world state from a negative utterance (and vice versa), thus failing to evidence verbal irony.

This model simulation reveals a critical puzzle in the interpretation of 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 a dimension of emotional arousal in the space of affects a listener considers. The orange lines in Figure 3.1 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 ironic interpretation of "The weather is terrible," meaning that the weather is likely *amazing*. This is because, with the enriched

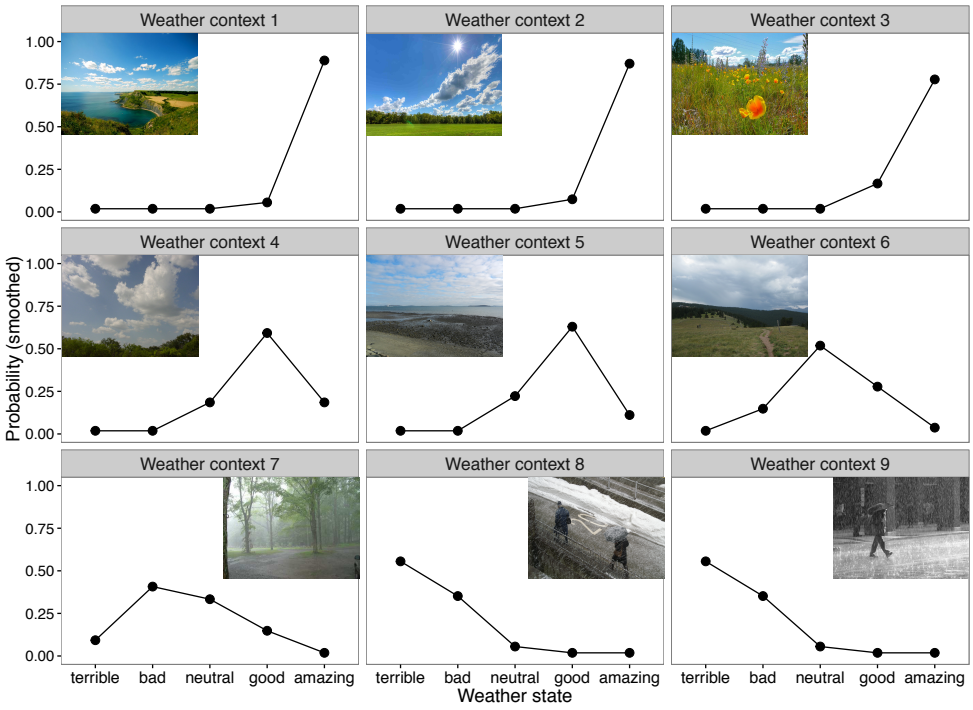

Figure 3.2: Smoothed prior probability distributions over weather states for each of the nine weather contexts. Participants saw each image and chose a state label from the set: *terrible, bad, neutral, good, amazing*. Probability distributions over weather states were computed by performing Laplace smoothing on the counts for each state label given a weather context and normalizing the counts to sum up to 1.

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 more psychologically realistic, two-dimensional affect space enables the qRSA model to interpret ironic utterances in addition to hyperbolic ones.

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 1a 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.

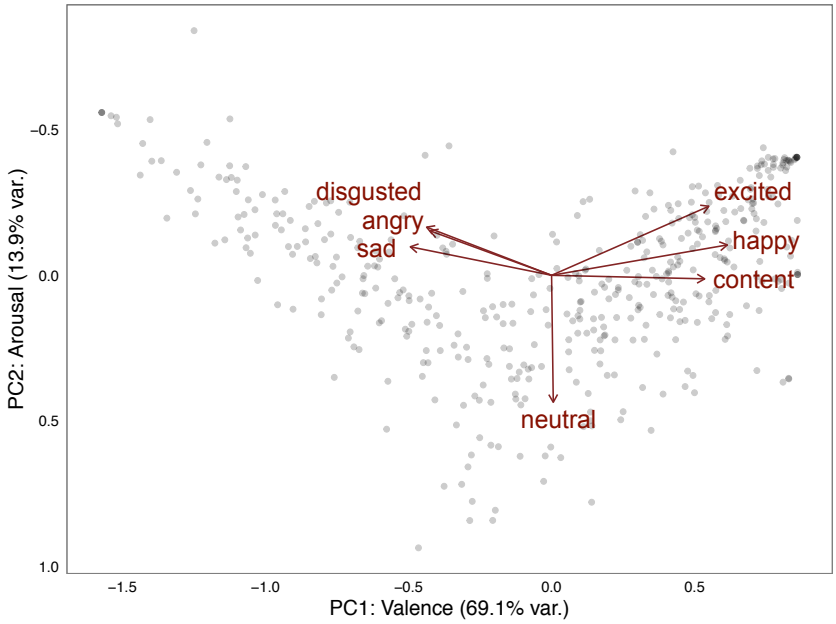

Figure 3.3: Biplot of the first two principle components of the seven emotion ratings. The first two PCs correspond roughly to valence and arousal, with positively valenced emotions (*excited, happy, content*)clustering on the right, and more high arousal emotions (*disgusted, excited*) appearing appearing at the top.

Information about emotions associated with each weather state allows us to empirically derive the affective space and priors, $P(A|s)$, for this domain. In Experiment 1b, we 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 1a: Background knowledge for verbal irony

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. Each of these images represent what we will call a "weather context," as shown in Figure 3.2.

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.

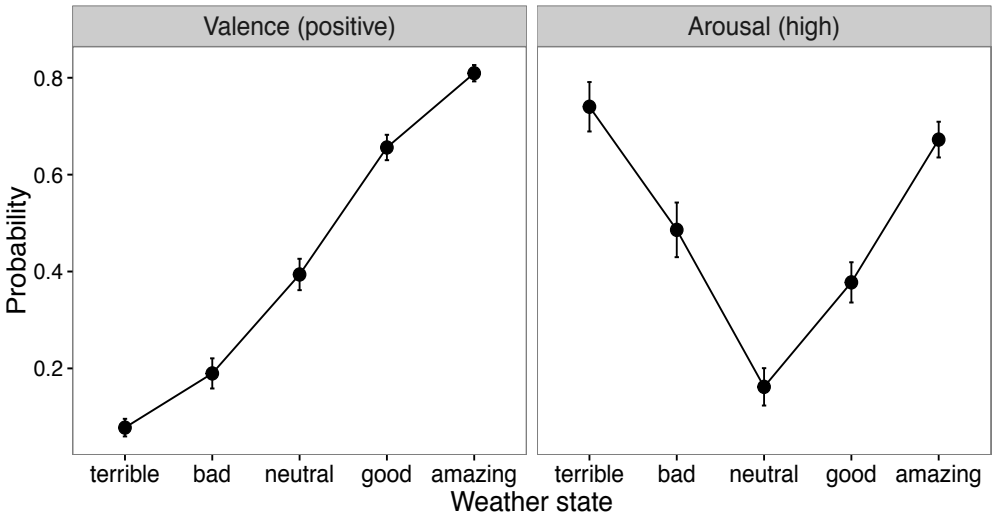

Figure 3.4: Average probabilities of positive valence and high arousal given each weather state. Error bars are 95% confidence intervals. Probability of positive valence increases monotonically over the five weather states; probability of high arousal follows a symmetric U-shaped curve and does not differ significantly for the *terrible* and *amazing* weather states.

They then indicated how Ann would rate the weather using a labeled 5-point Likert scale, ranging from *terrible*, *bad*, *neutral*, *good*, to *amazing*. Participants also 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 common emotion categories (Ekman, 1992)*. The order of the emotions was randomized for each participant but remained consistent across trials †.

For each of the nine weather contexts, we obtained the number of participants who gave each of the weather state ratings. We performed add-one Laplace smoothing on the counts to compute a smoothed prior distribution over weather states given each context, namely $P(s)$ (Figure 3.2). 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 studies of emotion ratings (Russell, 1980). We found that the first two principal components corresponded to the dimensions of emotional valence

*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.

†Link to Experiment 1a: `http://stanford.edu/~justinek/irony_exp/priors/priors.html`

and emotional arousal, accounting for 69.14% and 13.86% of the variance in the data, respectively. As Figure 3.3 shows, the first two principle components successfully distinguish positively valenced emotions (*excited, happy, content* from negatively valenced emotions (*disgusted, angry, sad)*, as well as high arousal emotions (*excited, disgusted*) from low arousal emotions (*content, neutral, sad*).

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 3.4). We observe that the probability of positive valence given a weather state increases monotonically across the ordered set of states: *terrible, bad, neutral, good*, and *amazing*, where the probability of positive valence given a *terrible* state is significantly lower than the probability given an *amazing* state. However, the probability of high arousal given each weather state follows a U-shape curve, where the probability of high arousal given a *terrible* state is approximately equivalent to the probability of high arousal given an *amazing* state. In other words, while the valences associated with *terrible* and *amazing* differ significantly, the arousals evoked by these states are very similar.

Is this figure useful?

Experiment 1b: Interpreting verbal irony

We conducted Experiment 1b to elicit people's interpretations of utterances, which we then use to evaluate model predictions. 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 **??** 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 Ann is to feel each of seven emotions about the weather *.

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 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 different 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 3.6). 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 1a to project the emotion ratings from Experiment 1b onto the same dimensions. We then standardized and converted the scores into

*Link to Experiment 1b: `http://stanford.edu/~justinek/irony_exp/interpretation/interpretation_askIrony.html`

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.

Irony model evaluation

From Experiment 1a, 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 1b: the speaker optimality parameter ($\lambda = 1$) and the prior probability of each of the three QUDs ($P(q_{state}) = 0.3$, $P(q_{valence}) = 0.3$, $P(q_{arousal}) = 0.4$)*. 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 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 3.8 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 $= [0.892, 0.903]$)† 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 for utterances that are judged as highly ironic (the darker dots in Figure 3.8). Thus, the model is able to to recover the intended 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 ($\rho = 0.763 \pm 0.005$). Finally, the model's inferences about the QUD varies systematically across the utterances and weather contexts (Figure 3.7. For utterances with high irony ratings (e.g. "terrible" uttered in WC1 and "amazing" uttered in WC9), the model infers that the QUD is most likely the speaker's emotional arousal, whereas for utterances with low irony ratings (e.g. "terrible" uttered in WC9 and "amazing" uttered

*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.

†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 †).

in WC1), the model infers that the QUD is most likely the weather state. In fact, the model's posterior probability of an arousal QUD captures participants' graded judgments of irony and is highly correlated with irony ratings ($r = 0.88$, $\rho = 0.94 \pm 0.005$). This suggests that the model may be able to use inferences about the QUD to identify ironic uses, and also that the degree of perceived irony may be associated with the probability of affective QUDs such as emotional arousal.

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 figurative interpretations.

Discussion

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 RSA to account for verbal irony. By making a minimal extension to J. T. Kao, Wu, 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, namely reasoning about shared background knowledge as well as the speaker's affective goals.

There remain important qualities of verbal irony to account for. For example, speakers often use verbal irony to remind the listener of previous utterances that turned out to be false, or of positive norms that were violated (Sperber & Wilson, 1981; Jorgensen, Miller, & Sperber, 1984). On the other hand, pretense theory argues that when a speaker produces an ironic utterance, she is only pretending to be someone who would make such an utterance (Clark & Gerrig, 1984). While our model is able to capture the main characteristics of verbal irony, it does not account for the intuitions behind echoic mention or pretense theories. We hope to enrich our model's understanding of the social aspects of irony by addressing these intuitions in future research. In addition, we aim to further examine how people identify the particular dimensions of meaning that may be under discussion in a given context. For example, affective dimensions such as valence and arousal may be particularly relevant in domains that involve evaluation (e.g. "good" or "terrible" weather), while non-affective dimensions may be more salient in other domains (J. T. Kao, Bergen, & Goodman, 2014).

3.3.2 Metaphor

Todo: Better segue into metaphor that highlights the differences

Todo: Run interpretation experiment with free-response features to show that features relevant for interpretation are pretty consistent with features of the source domain (at least for the metaphors we're looking at)

In the work described above, we assumed that the set of QUDs under consideration included the speaker's affect, which is supported by previous research on the rhetorical effect of hyperbole and verbal irony. In what follows, we will explore ways to systematically elicit the set of QUDs that listeners consider, as well as to manipulate prior probabilities over QUDs using discourse context. We show that considering these additional, non-affective QUDs allows the model to capture interesting effects of metaphor interpretation. In particular, we will focus on three aspects of the pragmatics

of metaphor understanding that the qRSA model naturally captures. First, interpretation of the same metaphor differs systematically given different discourse contexts, which can be modeled as different prior probabilities over QUDs (Experiment 2). Second, metaphors are able to communicate information efficiently along several dimensions and address multiple QUDs at once, which may serve as an advantage over literal statements (Experiment 2). Finally, the specific interpretations of a metaphor are sensitive to the alternative utterances that a speaker could have chosen to address the QUDs under consideration (Experiment 3).

While metaphoricity can arise from sentences with various types of syntactic forms, to reasonably limit the scope of our work, we focus on nominal metaphors of the classic form "X is a Y." For example, suppose a speaker uses the following utterance to describe a person, Bob: "Cam is a shark." 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 "shark."

Model

We again introduce a literal listener L_0, who interprets the utterances as meaning that Cam is literally a shark. Since L_0 believes Cam is a shark, she also believes that Cam is likely to have features associated with sharks, for example, being scary or fierce. The following equation represents the literal listener's interpretation, where c is Cam's category (either a "person" or a "shark"), and $\vec{f}$ is a vector representation of Cam'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 Cam's species category, or Cam'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:

$$U(u|Q, c, \vec{f}) = \log \sum_{c,\vec{f}} \delta_{Q(c,\vec{f})=Q(c',\vec{f}')} L_0(c', \vec{f}'|u) \tag{3.9}$$

Given this utility function, the speaker chooses an utterance according to a softmax decision rule, where λ is an optimality parameter:

$$S_1(u|Q,c,\vec{f}) \propto e^{\lambda U(u|Q,c,\vec{f})}, \tag{3.10}$$

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 Cam is actually a shark and that S_1 may want to communicate about Cam's scariness, she will end up with a posterior distribution where Cam is very likely to be a person who is scary. By combining prior knowledge with reasoning about the speaker's communicative goal, the pragmatics listener can thus arrive at a figurative interpretation of "Cam is a shark"—Cam is a very scary person.

In our first exploration of the model's behavior, we made a number of simplifying assumptions. First, we restricted the number of possible categories to which a member may belong to c_a and c_p, denoting an animal category (in this case *shark*) or a person category, respectively. We also restricted the possible features of Cam under consideration to a vector of size three: $\vec{f} = [f_1, f_2, f_3]$, where f_i is either 0 or 1. Finally, we assumed a small and rather impoverished set of alternative utterances that the speaker could have said: the utterance she did say (e.g. "Cam is a shark"), and a grammatically similar and literally true utterance (e.g. "Cam is a person.").*

Based on this formulation, the listener needs to consider the following prior probabilities to arrive at an interpretation:

(1) $P(c)$: the prior probability that the entity discussed belongings to category c. We assume that the listener is extremely confident that the person under discussion (e.g. John) is a person, but that there is a non-zero probability that John is actually a non-human animal. We fit $P(c_a)$ to data with the assumption that $10^{-4} \leq P(c_a) \leq 10^{-1}$.

*In principle, the model can be extended to accommodate more categories, features, and alternative utterances. In Experiment 3, in particular, we explore the model's behavior given more animal categories and alternatives.

(2) $P(\vec{f}|c)$: the prior probability that a member of category c has feature values $\vec{f}$. This is empirically estimated in Experiment 1.

(3) $P(g)$: the probability a speaker has goal g. This prior can change based on the conversational context that a question sets up. For example, if the speaker is responding to a vague question about Cam, e.g. "What is Cam like?", the prior over goals is uniform. If the question targets a specific features, such as "Is Cam scary?", then she is much more likely to have the goal of communicating John's scariness. However, she may still want to communicate other features about Cam that were not asked about. We assume that when the question is specific, the prior probability that S_n's goal is to answer the specific question is greater than 0.5, fitting the value to data below.

To evaluate our model's interpretation of metaphorical utterances, we selected a set of 32 metaphors comparing human males to various non-human animals. We conducted Experiment 2a and 2b to elicit feature probabilities for the categories of interest. We then conducted Experiment 2c to measure people's interpretations of the set of metaphors.

Experiment 2a: Feature Elicitation

We selected 32 common non-human animal categories from an online resource for learning English (`www.englishclub.com`). The full list is shown in Table 1. 100 native English speakers with IP addresses in the United States were recruited on Amazon's Mechanical Turk. Each participant read 32 animal category names presented in random order, e.g. "whale", "ant", "sheep". For each animal category, participants were asked to type the first adjective that came to mind in a text box. Using participants' responses, we constructed a list of adjectives for each animal category and ordered them by the number of times they were given by a different participant (i.e. their popularity). We removed all color adjectives, such as "white" and "black," to eliminate the possibility of interpreting these adjectives as racial descriptions. To avoid redundancy in the feature set, we used WordNet (Miller, 1995) to identify synonymous adjectives and only kept the most popular adjective among a set of synonyms. We then took the three most popular adjectives for each animal category and used them as the set of features. In what follows, f_1 is the most popular adjective, f_2 the second, and f_3 the third. Table 1 shows the animal categories and their respective features.

Experiment 2b: Feature Prior Elicitation

Using the features collected from Experiment 1a, we elicit the prior probability of a feature vector given an animal or person category (i.e. $P(\vec{f}|c)$). We assume that the adjective corresponding to a feature (e.g. *scary*) indicates that the value of that feature is 1 (present), while the adjective's antonym indicates that the value of that feature is 0 (not present). We used WordNet to construct antonyms for each of the adjective features produced in Experiment 1a. When multiple antonyms existed or when no antonym could be found on WordNet, the first author used her judgment to choose the appropriate antonym. Table 1 shows the resulting list of antonyms. For each animal category, eight possible feature combinations were constructed from the three features and their antonyms. For example, the possible feature combinations for a member of the category "ant" are {small, strong, busy}, {small, strong, idle}, {small, weak, busy}, and so on.

60 native English speakers with IP addresses in the United States were recruited on Amazon's Mechanical Turk. Each participant completed 16 trials in random order. Each trial consisted of the eight feature combinations for a particular animal category. Using slider bars with ends marked by "Impossible" and "Absolutely certain," participants were asked to rate how likely it is for a member of the animal category to have each of the eight feature combinations. Participants also rated the probabilities of the feature combinations for a male person. We only elicited priors for males to minimize gender variation and to maintain consistency with Experiment 2c.

We normalized each participant's ratings for the eight feature combinations in a trial to sum up to 1 based on the assumption that the feature combinations exhaustively describe a member of a particular category. Using the Spearman-Brown prediction equation, reliability of the ratings was 0.941 (95% CI = $[0.9408, 0.9414]$). Averaging across participants' normalized ratings, we obtained feature priors $P(\vec{f}|c)$ for $c = c_a$ (animal) and $c = c_p$ (person). Since the features were created using the animal categories in Experiment 1a, by construction features are rated as significantly more likely to be present in the animal category than in the person category ($F(1, 190) = 207.1$, $p < 0.0001$). These results confirm that participants are fairly confident that each animal category has certain distinguishing features (mean= 0.61, sd= 0.06), while those same features are rated as appearing in people less often (mean= 0.48, sd= 0.06).

Experiment 2c: Metaphor Interpretation

We created 32 scenarios based on the animal categories and results from Experiment 1. In each scenario, a person (e.g. Bob) is having a conversation with his friend about a person that he recently met. Since we are interested in how the communicative goals set up by context affect metaphor interpretation as well as the effectiveness of metaphorical versus literal utterances, we created four conditions for each scenario by crossing vague/specific goals and literal/metaphorical utterances. In vague goal conditions, Bob's friend asks a vague question about the person Bob recently met: "What is he like?" In specific goal conditions, Bob's friend targets f_1 and asks a specific question about the person: "Is he f_1?," where f_1 is the most popular adjective for a given animal category c_a. In literal conditions, Bob replies with a literal utterance, either by saying "He is f_1." to the question "What is he like?" or "Yes." to the question "Is he f_1?". In Metaphorical conditions, Bob replies with a metaphorical statement, e.g. "He is a c_a." where c_a is an animal category. See Table 2 for examples of each condition.

49 native English speakers with IP addresses in the United States were recruited on Amazon's Mechanical Turk. Each participant completed 32 trials in random order. The 32 trials were randomly and evenly assigned to one of the four conditions, i.e. each participant read 8 scenarios for each condition. For each trial, participants used sliders to indicate the probabilities that the person described has features f_1, f_2, and f_3, respectively.

For each condition of each scenario, we obtained the average probability ratings for the three features. Figure **??** shows the average ratings for each feature across animal categories given a vague or specific goal and a literal or metaphorical utterance. When the speaker gives a literal statement directly affirming the presence of f_1, participants rate f_1 as significantly more likely than when the speaker gives a metaphorical statement ($F(1, 126) = 52.6$, $p < 0.00001$). However, participants rate f_2 and f_3 as significantly more likely when the speaker produces a metaphorical utterance than when the utterance is literal ($F(1, 126) = 23.7$, $p < 0.0001$; $F(1, 126) = 13.66$, $p < 0.0005$). Comparing feature probability ratings in Experiment 2 to the feature priors obtained in Experiment 1b, we can measure how literal and metaphorical utterances change listeners' inferences about a person's features. Given a literal utterance that directly confirms the existence of f_1, probability ratings for f_1 are significantly higher than the prior probabilities of f_1 for a person ($t(63) = 59.19, p < 0.00001$). However, probability ratings for f_2 and f_3 are not significantly

different from their prior probabilities ($t(63) = -0.13, p = 0.89$; $t(63) = 0.03, p = 0.97$). Given a metaphorical utterance, probability ratings for all three features are significantly higher than the prior probabilities ($t(63) = 15.74, p < 0.0001$; $t(63) = 7.29, p < 0.0001$; $t(63) = 5.91, p < 0.0001$). This analysis suggests that metaphorical utterances may convey richer information and update listeners' beliefs along more dimensions than literal utterances.

We now analyze the effect of the speaker's communicative goal on the interpretation of literal or metaphorical utterances. When the speaker's utterance is literal, the probability ratings for f_1, f_2, and f_3 are not significantly different given a vague or a specific question (($F(1, 62) = 2.73, p = 0.1; F(1, 62) = 0.0001, p = 0.99; F(1, 62) < 0.0001, p = 0.99$). For metaphorical utterances, however, the question type has an effect on participants' interpretations: participants rate the probability of f_1 as significantly higher when the question is specifically about f_1 than when it is vague ($F(1, 62) = 10.16$, $p < 0.005$). The probabilities of f_2 and f_3 are not significantly different given a vague question or a specific question about f_1 ($F(1, 62) = 0.04, p > 0.05$; $F(1, 62) = 0.8285, p > 0.05$). This suggests that people's interpretation of metaphor may be more sensitive to the communicative goals set up by context than their interpretation of literal utterances.

Metaphor model evaluation

We used the feature priors obtained in Experiment 2b to compute model interpretations of the 32 metaphors. As discussed in the previous section, the behavioral results in Experiment 2c show evidence that the context set up by a question changes participants' interpretation of a metaphor. Our model naturally accounts for this using the speaker's prior over communicative goals $P(g)$. When a speaker is responding to a vague question, we set the prior distribution for $P(g)$ as uniform. When the speaker is responding to a question specifically about f_1, we assume that $P(g_1) > 0.5$ and equal between $P(g_2) = P(g_3)$. Fitting the goal prior parameter to data yields a prior of $P(g_1) = 0.9$ when responding to a specific question about f_1. We fit the category prior $P(c_a) = 0.01$ and the speaker optimality parameter $\lambda = 4$.

Using these parameters, we obtained interpretation probabilities for each of the 32 metaphors under both vague and specific goal conditions. For each metaphor and goal condition, the model produces a joint posterior distribution $P(c, \vec{f}|u)$. We first show a basic but important qualitative result, which is that the model is able to interpret utterances metaphorically. Marginalized over

values of $\vec{f}$, the probability of the person category given the utterance is close to one ($P(c_p|u) = 0.994$), indicating that the pragmatic listener successfully infers that the person described as an animal is actually a person and not an animal. This shows that the model is able to combine prior knowledge and reason about the speaker's communicative goal to arrive at nonliteral interpretations of utterances.

We now turn to the second component of the interpretation, $P(\vec{f}|u)$. To quantitatively evaluate the model's performance, we correlated model predictions with human interpretations of the metaphorical utterances. Given a metaphorical utterance and a vague or specific goal condition, we computed the model's marginal posterior probabilities for f_1, f_2, and f_3. We then correlate these posterior probabilities with participants' probability ratings from Experiment 2c. Figure 3.10 plots model interpretations for all metaphors, features, and goal conditions against human judgments. Correlation across the 192 items (32 metaphors $\times$ 3 features $\times$ 2 goal conditions) is 0.64 ($p < 0.001$)*. The predicted reliability of participants' ratings using the Spearman-Brown prediction formula is 0.828 (95% CI = $[0.827, 0.829]$), suggesting first that people do not agree perfectly on metaphorical interpretations, and second that our model captures a significant amount of the reliable variance in the behavioral data. In particular, our model does especially well at predicting participants' judgments of f_1, which are the most salient features of the animal categories and were targeted by specific questions in Experiment 2. Correlation between model predictions and human judgments for f_1 is 0.77 ($p < 0.0001$), while the predicted reliability of participants' ratings for f_1 is 0.82 (95% CI = $[0.818, 0.823]$).

We now compare our model's performance to a baseline model that also considers the feature priors and the conversational context. We constructed a linear regression model of participants' feature ratings that takes as predictors the marginal feature priors for the animal category, the marginal feature priors for the person category, whether the QUD is vague or specific, and their interactions. With eight parameters, this model produced a fit of $r = 0.48$, which is significantly worse than our model ($p < 0.0001$ on a Cox test). This suggests that our computational model adequately combines people's prior knowledge as well as principles of pragmatics to produce metaphorical interpretations that closely fit behavioral data.

*We also fit a free parameter that determines a power-law transformation of the feature priors. The power-law transformation was introduced based on analyses of the feature priors showing that participants tend to overestimate unlikely features and underestimate likely features, likely due to a tendency to avoid the extreme ends of the slider bars. Without performing the power-law transformation and using the raw normalized feature priors obtained in Experiment 2a, the model correlation is 0.6 ($p < 0.001$)

While our model predictions provide a reasonable fit to behavioral data and outperforms a linear regression model using fewer parameters, we observed residual variance that can be further addressed. Previous work has shown that alternative utterances—what the speaker could have said—can strongly affect listeners' interpretation of what the speaker *did* say (Bergen et al., 2012). In this experiment, we did not take into account the range of alternative utterances (both literal and metaphorical) that a listener considers when interpreting a speaker's utterance. We posit that other plausible alternative utterances may account for some of the variance in the data that our model does not capture. Consider the metaphor "He is an ant" and the corresponding features *small, strong*, and *busy*. Our model currently assigns a high probability to the feature *strong* given the metaphor, while participants assign it a lower probability. Indeed, this data point has the highest residual in our model fit. To test the idea that alternative utterances may account for this discrepancy, we construct a model that has "He is an ox" as an alternative utterance. "Ox" has features that roughly align with the features of "ant": *strong, big*, and *slow*. Since *strong* is a higher probability feature for "ox" than for "ant," the listener reasons that if the speaker had intended to communicate the feature *strong*, she would have said "He is an ox" since it optimally satisfies that goal. Since the speaker did *not* produce the utterance "He is an ox," the listener infers that *strong* is a less probable feature. Adding this alternative utterance to the model indeed lowers the marginal posterior probability of *strong* given the utterance "He is an ant." Based on this simple error analysis, we posit that adding alternative utterances across all animal categories may significantly improve our model's performance.

Given the large space of animal categories and features selected for Experiment 2, constructing a complete set of alternative utterances using the full set of 32 metaphors was not feasible and would result in a very large and unwieldy set of animals and features. Instead, for the purposes of this current experiment, we focused on a smaller set of initial animal categories as well as a smaller set of features in order to elicit the set of alternative metaphors a listener may reasonably expect a speaker to produce. In the next section, we describe Experiment 3, where we begin with a set of 12 animal metaphors and elicit alternative metaphors to examine their effect on the model's behavior.

Experiment 3a: Feature Elicitation

We selected 12 common animals that have various distinguishing features: *ant, whale, bird, elephant, panda, monkey, penguin, giraffe, cheetah, turtle, lion*, and *rabbit*, which we will call the core animals. 50 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk. Each participant read 12 animal category names presented in random order. Following the same paradigm as Experiment 2a, participants were asked to type the first adjective that came to mind in a text box. We removed color adjectives and combined synonymous adjectives by consulting WordNet (Miller, 1995). We then identified the two most popular adjectives for each animal category and used those as features [*].

Experiment 3b: Alternative Elicitation

In order to adequately interpret an utterance that a speaker *did* say, listeners often need to reason about a set of alternative utterances that the speaker could have said but did not. While the importance of alternative utterances is supported by psycholinguistic evidence (Bergen et al., 2012; Krifka, 2007; Degen & Tanenhaus, 2015), how listeners arrive at a reasonable set of alternative utterances is still an open question. For our purposes, we take the approach of constructing alternative utterances based on the set of QUDs under consideration. Since we assume that the QUDs considered for metaphor interpretation are related to the features associated with the metaphorical source, we assume that the alternative utterances a listener considers to interpret a metaphor are in turn associated with those features. More concretely, suppose a speaker produced the metaphor, "Cam is a bird." We assume that the interpretation of this metaphor (how likely it is that Cam is fast, small, sings, etc) will be influenced by other animal metaphors the speaker could have produced to communicate information about Cam's speed, size, and singing ability. Based on this reasoning, we conducted Experiment 3b to elicit alternative animal metaphors for each of the features associated with the core animals.

Experiment 3a yielded 15 unique features elicited from the 12 core animals: 2 features for each core animal, with 9 overlapping features across the 12 animals such as "small," "big," and "strong." 50 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk

[*]In Experiment 3, we decided to use two features for each category instead of three due in part to the difficulty of obtaining reliable ratings from human participants for $2^3 = 8$ possible feature combinations, and in part to restrict the set of alternative utterances we elicit in Experiment 3b to a reasonable size.

to produce animals associated with each of the 15 features. Each participant read the 15 feature adjectives presented in random order, with the prompt: "Write down the first animal you can think of that is _____," where _____ is a feature adjective. We tallied the number of times an animal was produced for each of the features and identified the animals most strongly associated with each feature.

For each of the 12 core animals we started with, we now have 2 associated features, as well as a set of animals associated with those features. For each core animal, we constructed a set of alternative animals by selecting 2 animals most strongly associated with each feature, excluding the core animal itself. If there were animals that were strongly associated with both of the two features, we selected an additional animal associated with the first feature, in order to ensure that each core animal had exactly 4 alternative associated animals. The full set of core animals, features, and alternative animals are shown in Table 3.3.2.

Experiment 3c: Feature Prior Elicitation

Given the features and alternative animals collected from Experiment 3a and 3b, we now elicit the prior probability of a feature vector given an animal category (i.e $P(\vec{f}|c)$). For each set of two features (e.g. {small, industrious}), we again constructed antonyms to indicate when the feature is absent, resulting in four possible feature combinations (e.g. {small, industrious}, {small, lazy}, {big, industrious}, and {big, lazy}). For each feature set, we elicit prior feature probabilities for the following categories: the core animal (e.g. *ant*); the alternative animals (e.g. *mouse, dog, beaver*, and *monkey*); and a human male.

27 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk. Each participant completed 24 trials in random order. Each trial consisted of the four feature combinations for a particular animal category. Using slider bars with end points marked by "very unlikely" and "very likely," participants were asked to rate how likely it is for a member of the category to have each of the four feature combinations. Based on the assumption that the feature combinations exhaustively describe a member of a particular category, we normalized each participant's ratings for the four feature combinations within a trial to sum up to 1.

Experiment 3d: Metaphor Interpretation

45 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk.

Describe results.

Model evaluation

Model comparison with and without alternative animals.

Discussion

In this section, we showed that rich metaphorical interpretations can be captured by the the general communicative principles formalized in the qRSA model. In particular, our model successfully captures several important aspects of metaphor interpretation. First, the model goes beyond the literal meanings of utterances to infer non-literal interpretations (e.g., Cam is a person and not an ant). Second, the model provides quantitative judgments about the person's features based on prior knowledge of the metaphorical source (e.g. ants are very likely small and industrious) as well as on the local conversational context (e.g. is the topic of conversation specifically about Cam's size, or his characteristics more generally?). Third, the model successfully captures the empirical finding that metaphors tend to communicate information along more dimensions than literal statements. Finally, a model that takes into account the alternative utterances (both literal and metaphorical) that a speaker could have produced to address specific QUDs predicts people's interpretations with higher quantitative accuracy than one that does not. Together, these results suggest that basic principles of communication shape metaphor interpretation in important ways, which can be formalized in a general computational framework that assumes speakers to be rational and cooperative agents.

3.3.3 Hyperbolic Metaphor

Metaphors often elicit hyperbolic interpretations and result in more extreme beliefs about the world. In this section, we examine people's interpretations of metaphors such as "Cam 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 "Cam is a giraffe," listeners believe that Cam is taller than they do when given the literal description "Cam

is tall." Second, listeners believe that Cam is taller than the average person, but much shorter than the average giraffe. We show that the qRSA model naturally captures these effects.

Model

Experiment 4a: Prior Elicitation

Experiment 4b: Metaphor Interpretation

Model Evaluation

Discussion

3.4 General Discussion

In this chapter, we reviewed approaches to studying figurative language understanding from the perspective of pragmatics and communication. We described the communicative principles and extralinguistic factors that shape figurative understanding and articulated the need to clarify the interactions among these components in order to more precisely understand how people arrive at figurative meanings in context. Building upon the RSA framework, we proposed and evaluated a computational model that explicitly describes how people use different sources of information—literal meaning, background knowledge, and contextual information—to produce figurative interpretations. Through a series of behavioral experiments, we showed that the model closely matches people's interpretation of hyperbole, verbal irony, and metaphor, suggesting that different types of figurative language may share the same underlying communicative principles.

We believe that the qRSA model makes several critical advancements to formal models of human language understanding. The model captures key intuitions about communication, including the role of common ground between listener and speaker, the assumption that speakers produce utterances that maximize informativeness, and the idea that informativeness should be considered with respect to the question under discussion and the speakers' communicative goals. By formalizing these intuitions, the model is able to go beyond the literal meanings of utterances and predict subtleties in interpretation that are sensitive to background knowledge, communicative efficiency, local context, and alternative utterances. From a scientific perspective, our work provides a formal definition of QUD as well as the relevance principle, which allows us to empirically test the prediction that

listeners reason about utterances' relevance to the QUD in order to arrive at appropriate interpretations. This also allows us to show that QUD inference is critical for many instances of language understanding, in particular figurative language, where the literal meanings of utterances are often false. By exploiting listeners' assumption that speakers choose utterances in order to communicate information relevant to the QUD (and not irrelevant information or information that is already in common ground), speakers can choose utterances that are not literally true (e.g. "Cam is a giraffe") in order to effectively communicate relevant information (e.g. that Cam is unusually tall), without leading the listener to believe false information (e.g. that Cam is a giraffe). From an engineering perspective, our model provides a step towards building systems that use pragmatic reasoning and representations of the QUD to better interpret and generate utterances in context. We believe that our approach may contribute to the design of artificial agents that use language in more flexible and creative ways.

In addition to advancing general models of language understanding, our work may provide explanations for phenomena that particularly interest researchers of figurative language. One distinctive aspect of figurative utterances is that they communicate multiple dimensions of meaning at once and cite are often difficult to paraphrase . By considering encyclopedic knowledge such as affects associated with different states and features associated with different categories, our model naturally accounts for the multi-dimensional quality of figurative meaning. In addition, many researchers have observed that the interpretation of figurative language is especially sensitive to common ground (Pexman & Zvaigzne, 2004; Gibbs, 2000). Given that the literal meanings of figurative utterances are often implausible or false, listeners rely on background knowledge in order to reason about and recover speakers' intended meanings. As a result, it is important for listeners and speakers to share background knowledge (both encyclopedic knowledge and specific prior beliefs) in order to communicate successfully using figurative utterances. Indeed, researchers have found that willingness to use ironic utterances is positively correlated with social intimacy between interlocutors (Kreuz, 1996), and interlocutors who use metaphorical speech with each other are perceived as having a closer relationship(Horton, 2007). Kreuz (1996) attributes the effects of social closeness to what he calls a principle of inferability: speakers are more likely to use a non-literal utterance when they are more certain that it will be understood appropriately, which in turn is more likely when the speakers and listeners share similar background knowledge and beliefs. For example, suppose Ann saw a watch

that cost $1000 and wants to tell Bob that she thinks it is very expensive. In order to communicate her meaning effectively using a hyperbolic utterance: "That watch cost a million dollars!", Ann must be fairly confident that Bob's prior beliefs would not lead him to believe that the watch literally cost a million dollars. And since Bob reasons about Ann's motivation for choosing an utterance, given such a hyperbolic utterance, Bob arrives at the inference that Ann must be fairly confident about his prior beliefs. Our model provides a natural way to incorporate inferences about common ground in order to predict and model effects of social closeness. In its current form, our model assumes that the listener is certain that the speaker has perfect knowledge of the listener's background knowledge and prior beliefs. If we relax this assumption to incorporate uncertainty about the speaker's knowledge of the listener, the pragmatic listener is able to derive information about the speaker's knowledge of the listener. Preliminary explorations of our model show that it naturally affords these types of social inferences, which we plan to examine more thoroughly in future work.

One of the most important contributions of our work is that it unifies the interpretation of diverse types of figurative language in a single computational model. The generality of this model suggests that separate processing mechanisms may not be necessary to derive different types of figurative interpretations, or to derive figurative interpretations at all. The qRSA model does not distinguish *a priori* between figurative language and other types of language use; instead, the same pragmatic reasoning takes place regardless of whether the model ultimately produces a figurative or a literal interpretation. While our model is a computational-level account of language understanding and makes no process-level claims, we note that the model engages the same reasoning mechanism to interpret both figurative and literal utterances, which is consistent with psycholinguistic evidence that figurative language does not take reliably longer to process. On the other hand, some evidence suggests that prior context reduces the processing time for figurative utterances. Although we again do not make claims with our model regarding processing times, we suggest that this type of contextual effect may be modeled as a reduction of uncertainty in the QUD, which may lead to faster processing. Overall, our work suggests that the rich meanings expressed by figurative language can be explained by basic principles of communication, thus demonstrating the importance of considering pragmatics in theories of figurative language.

Not sure whether it's dangerous to mention processing times at all, but I think reviewers will likely ask about it, so might as well say something...

3.4.1 Future directions

The work we described invites many directions for future research, both for computational models of pragmatics and figurative language understanding. While our model explicitly reasons over a set of QUDs and alternative utterances, we have not precisely defined how listeners select the particular set of QUDs and alternatives over which to reason. For example, when should listeners consider speakers' affects and subjective attitudes to be potential QUDs? How do listeners decide that certain features of the metaphorical source are likely QUDs (e.g. *scary* in the case of "Cam is a shark"), while others are not (e.g. *has teeth* or *swims*)? In the work described in this paper, we made the simplifying assumption that QUDs arise from the literal meanings of utterances in an associative manner, and relied on intuition and previous research to focus on affective QUDs in the case of hyperbole and irony and feature QUDs in the case of metaphor. We then defined a set of QUDs based on the affects and features associated with various states and categories. Future research should examine whether QUDs can be more systematically inferred from prior context or the utterance itself, thus removing the need to predefine a set of QUDs that the model should consider for a given type of figurative utterance. A more flexible and general way to define a set of QUDs could also enable us to determine which types of QUDs are most effectively addressed by different kinds of figurative utterances.

Metaphor understanding is related to many other complex issues such as analogy, conventionality, and embodied cognition. To reasonably limit the scope of our work, in this paper we focused on simple nominal metaphors such as "Cam is a shark," where both the metaphor source and target are concrete objects, and where the source is relatively unconventional (for example, we did not include idiomatic metaphors such as "Cam is a chicken"). We also only considered attributional metaphors, where the source and target have certain features in common (e.g. "fierce" and "scary"). We have not yet shown that our model can account for other types of metaphors, such as (1) verbal metaphors, where verbs instead of nouns are used non-literally (e.g. "The ice-skater *flew* across the rink") (2) conventional metaphors, or metaphors that appear frequently in everyday language to the point of becoming "dead" or lexicalized (e.g. "The man was *drowning* in sorrow"; "She's the *head* of the household"). (3) relational metaphors, where the source and target share the same relational structure but not necessarily the same simple attributes (e.g. "A child's brain is a *sponge*" means that a child's brain absorbs information the way sponge absorbs water) (4) abstract attributional

metaphors, where the source and target only share attributes in an abstracted sense (e.g. "Cam is a *rock*" means that Cam is stable and dependable in his personality, but not necessarily physically static).

Future work will need to look into these other types...need to fill in ideas regarding how.

(3) and (4) may turn out to be the same thing.

While our model provides a unified explanation for three diverse uses language: hyperbole, irony, and metaphor, future work should examine whether the model extends to other types of figurative language. For example, understatement refers to the use of a mild statement to describe an extreme situation, such as saying "It's a bit rainy today" in the middle of a rainstorm in order to draw attention to the extreme raininess. Because the literal meaning of "a bit rainy" is associated with both neutral valence and low arousal, our model finds very little reason for a speaker to choose this utterance to communicate either negative valence or high arousal. What information, then, is a speaker trying to communicate using a statement that doesn't match the speaker's valence, arousal, or the objective state of the world? We observe that intuitively, understatements such as "It's a bit rainy today" draw attention to the common ground between speaker and listener, such as the fact that they both know that it is extremely rainy. In order to explain other types of figurative language such as understatement, it may be necessary to incorporate these types of common ground and social inferences. .

idioms?

Thus far, our work has focused on information-theoretic motivations for using figurative language, such as communicating efficiently about multiple QUDs at once, as well as communicating extreme states. In future work, we plan to further explore the social motivations for using figurative language, such as to communicate common ground with the listener or to evoke humor. Earlier in the discussion, we briefly described how figurative language may lead to inferences about common ground and outlined a way to incorporate these inferences in our model. In addition, we observe that figurative language is often funny, which also has social consequences. We believe the humor of figurative language can be explained by the Incongruity Theory, which posits that situations that afford multiple incongruous at the same time are more likely to be funny. We speculate that figurative utterances are often funny because they give rise to different interpretations given different QUDs, and hypothesize that the humor of figurative utterances can be predicted by formalizations of incongruity, which we plan explore in future work.

politeness?

3.5 Conclusion

Figurative language presents a puzzle for communication. While the information encoded in the literal semantics is false or trivial, figurative utterances are often evocative, socially meaningful, and highly informative. In this paper, we provided an account of figurative language understanding that partially solves this puzzle using a computational model of communication. We showed that a Rational Speech-acts model extended to accommodate inferences about the QUD successfully recovers true and relevant information from literally false utterances and predicts people's interpretations of hyperbole, irony, and metaphor with high accuracy. We argue that the qRSA model incorporates information in a principled and theoretically motivated manner, providing a useful framework for testing and modeling various phenomena in language understanding. By formalizing principles of communication to explain figurative language, our work sheds light on how our linguistic, cognitive, and social faculties work together to produce meanings that go far beyond the reach of words.

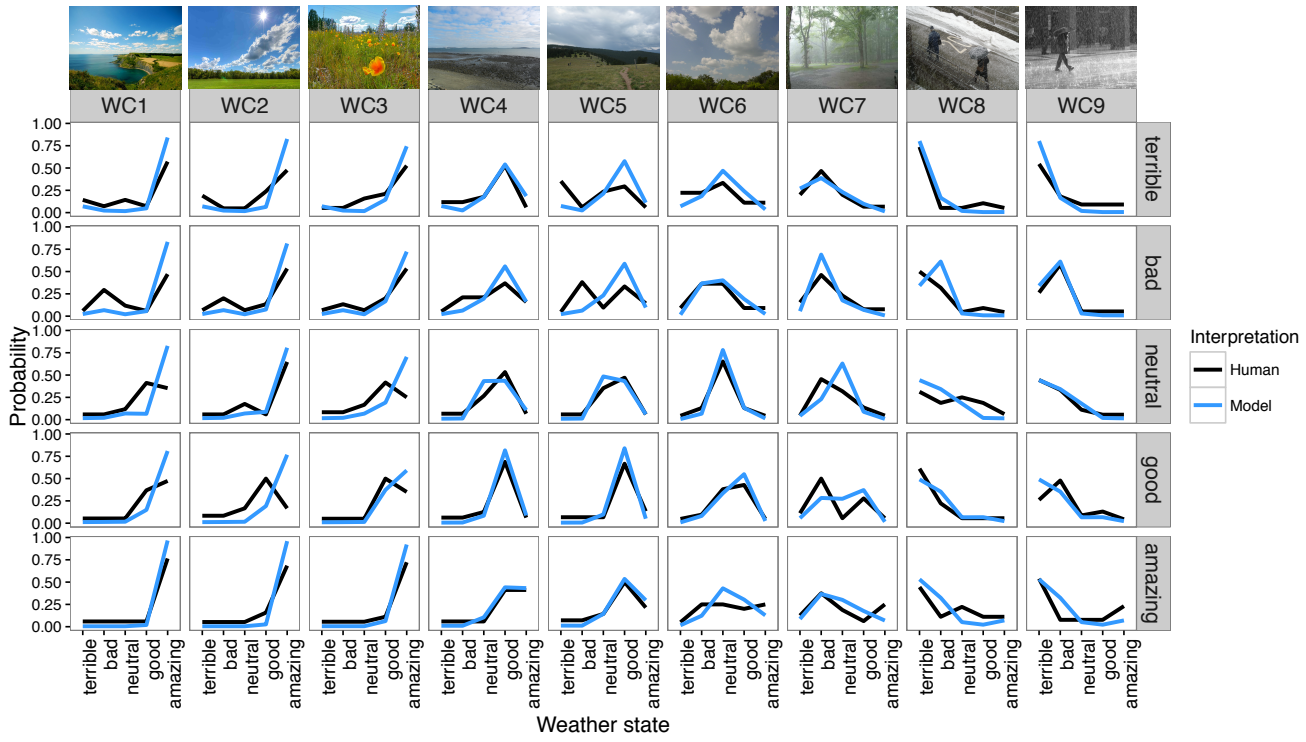

Figure 3.5: Model's and participants' inferences about the weather state (x-axis) given a weather context (column) and an utterance (row). Each panel represents an interpretation given an utterance in a weather context. The dark lines are participants' ratings; the light lines are the model's posterior distributions over weather states.

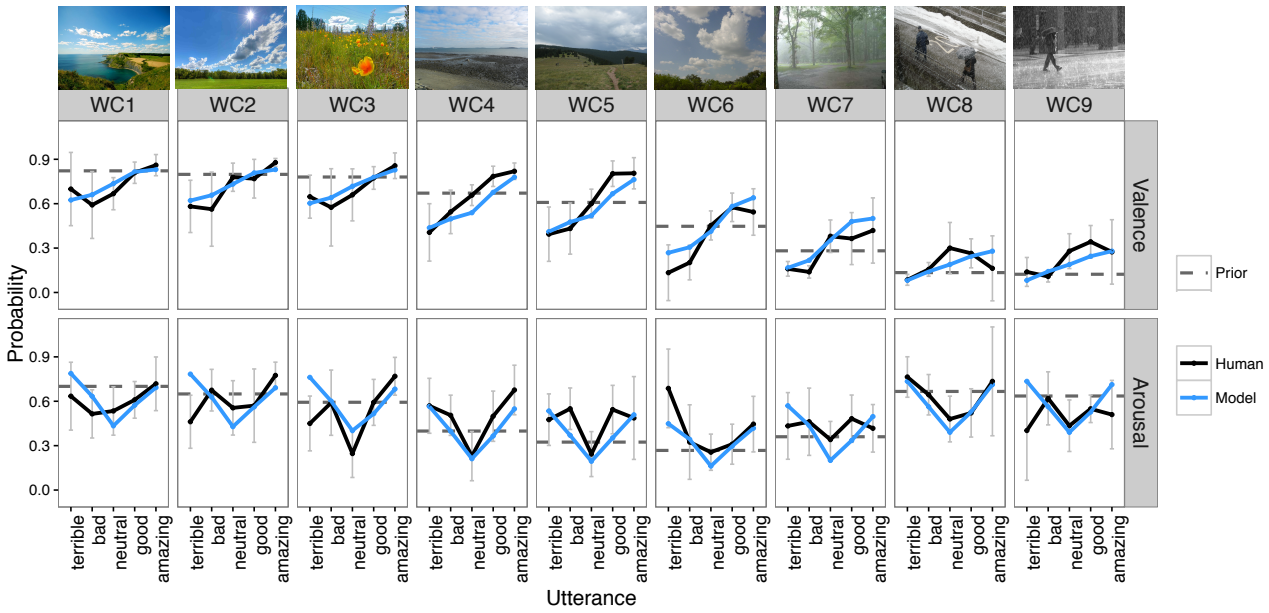

Figure 3.6: Model's and participants' inferences about the probability of valence and arousal (row) given a weather context (column) and an utterance (x-axis). The dark lines are participants' ratings; the light lines are the model's posterior probabilities of positive valence and high arousal given an utterance in a weather context. The dotted lines are prior probabilities of positive valence and high arousal for each weather context. Error bars are 95% confidence intervals on the participants' ratings.

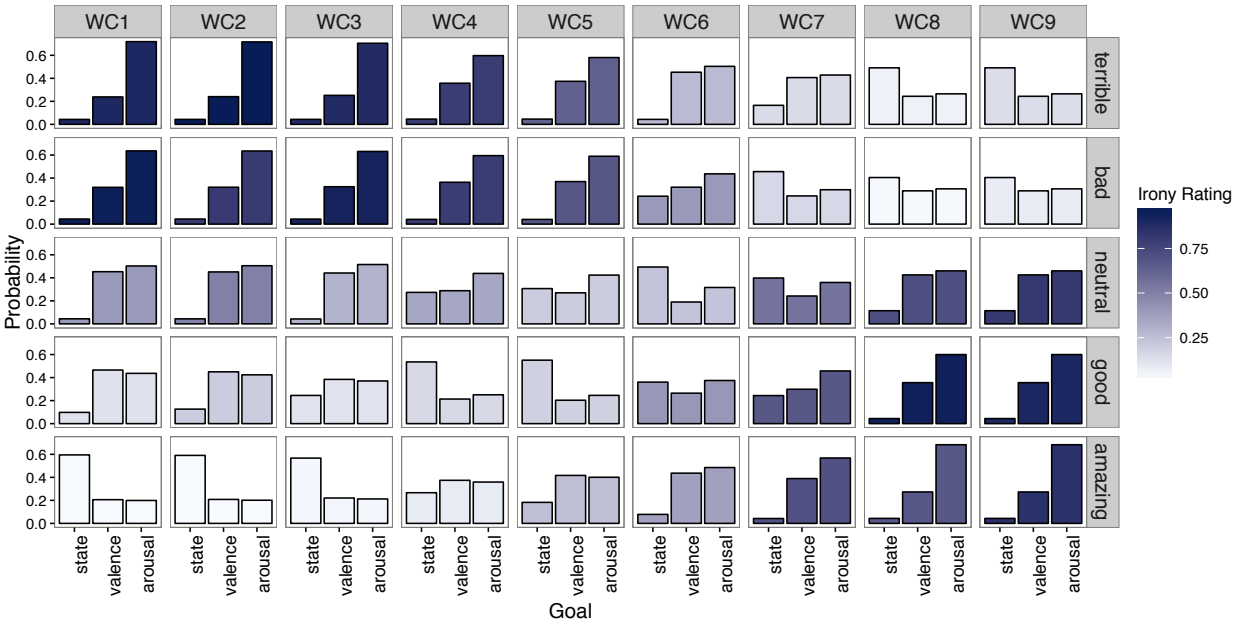

Figure 3.7: Model's posterior distributions over QUDs given an utterance (row) in a weather context (column). The darkness of the bars indicate participants' irony ratings for the utterances.

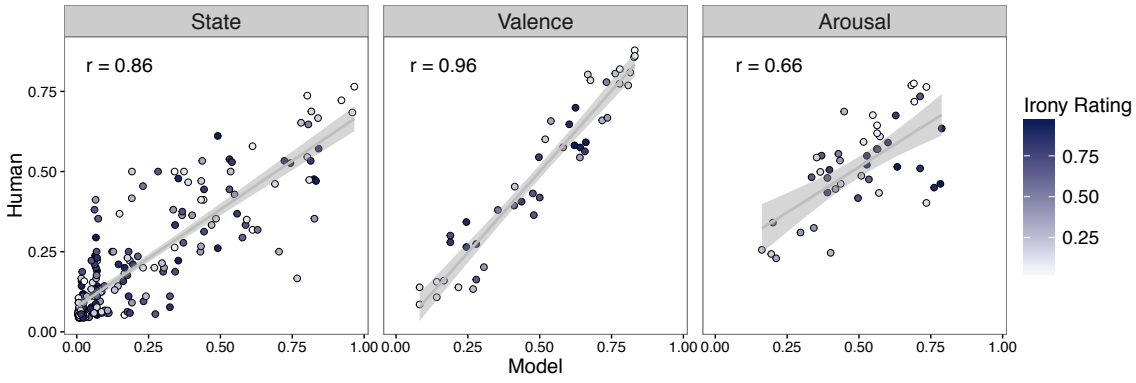

Figure 3.8: Scatter plot showing correlations between model predictions and human ratings for weather state, speaker valence, and speaker affect. Each dot in a panel represents the interpretation of an utterance in a weather situation, along the dimensions of weather state, valence, and arousal. The darkness of the dots indicate participants' irony ratings for the utterances.

Animal	$f_1 = 1$	$f_2 = 1$	$f_3 = 1$	$f_1 = 0$	$f_2 = 0$	$f_3 = 0$
ANT	small	strong	busy	large	weak	idle
BAT	scary	blind	nocturnal	unalarming	sighted	diurnal
BEAR	scary	big	fierce	unalarming	small	nonviolent
BEE	busy	small	angry	idle	large	unangry
BIRD	free	graceful	small	unfree	awkward	large
BUFFALO	big	strong	wild	small	weak	tame
CAT	independent	lazy	soft	dependent	fast	hard
COW	fat	dumb	lazy	thin	smart	fast
DOG	loyal	friendly	happy	disloyal	unfriendly	unhappy
DOLPHIN	smart	friendly	playful	stupid	unfriendly	unplayful
DUCK	loud	cute	quacking	quiet	unattractive	non-quacking
ELEPHANT	huge	smart	heavy	small	stupid	light
FISH	scaly	wet	smelly	smooth	dry	fragrant
FOX	sly	smart	pretty	artless	stupid	ugly
FROG	slimy	noisy	jumpy	nonslippery	quiet	relaxed
GOAT	funny	hungry	loud	humorless	full	quiet
GOOSE	loud	mean	annoying	quiet	nice	agreeable
HORSE	fast	strong	beautiful	slow	weak	ugly
KANGAROO	jumpy	bouncy	cute	relaxed	inelastic	unattractive
LION	ferocious	scary	strong	nonviolent	unalarming	weak
MONKEY	funny	smart	playful	humorless	stupid	unplayful
OWL	wise	quiet	nocturnal	foolish	loud	diurnal
OX	strong	big	slow	weak	small	fast
PENGUIN	cold	cute	funny	hot	unattractive	humorless
PIG	dirty	fat	smelly	clean	thin	fragrant
RABBIT	fast	furry	cute	slow	hairless	unattractive
SHARK	scary	dangerous	mean	unalarming	safe	nice
SHEEP	wooly	fluffy	dumb	hairless	hard	smart
TIGER	striped	fierce	scary	unpatterned	nonviolent	unalarming
WHALE	large	graceful	majestic	small	awkward	inferior
WOLF	scary	mean	angry	unalarming	nice	unangry
ZEBRA	striped	exotic	fast	unpatterned	native	slow

Table 3.2: 32 animal categories, feature adjectives, and their antonyms. Feature adjectives were elicited from Experiment 1a and indicate when a feature is present ($f_i = 1$). Antonyms were generated using WordNet and indicate when a feature is not present ($f_i = 0$). Feature sets shown in Experiment 2b were created with this table, where $\vec{f} = [1, 0, 0]$ for category "ant" is represented by the words {small, weak, idle}. There are $2^3 = 8$ possible feature combinations for each animal category.

QUD	Utterance	Example question	Example utterance
General	Literal	"What is he like?"	"He is scary."
Specific	Literal	"Is he scary?"	"Yes."
General	Metaphorical	"What is he like?"	"He is a shark."
Specific	Metaphorical	"Is he scary?"	"He is a shark."

Table 3.3: Example questions and utterances for each of the four experimental conditions in Experiment 2c.

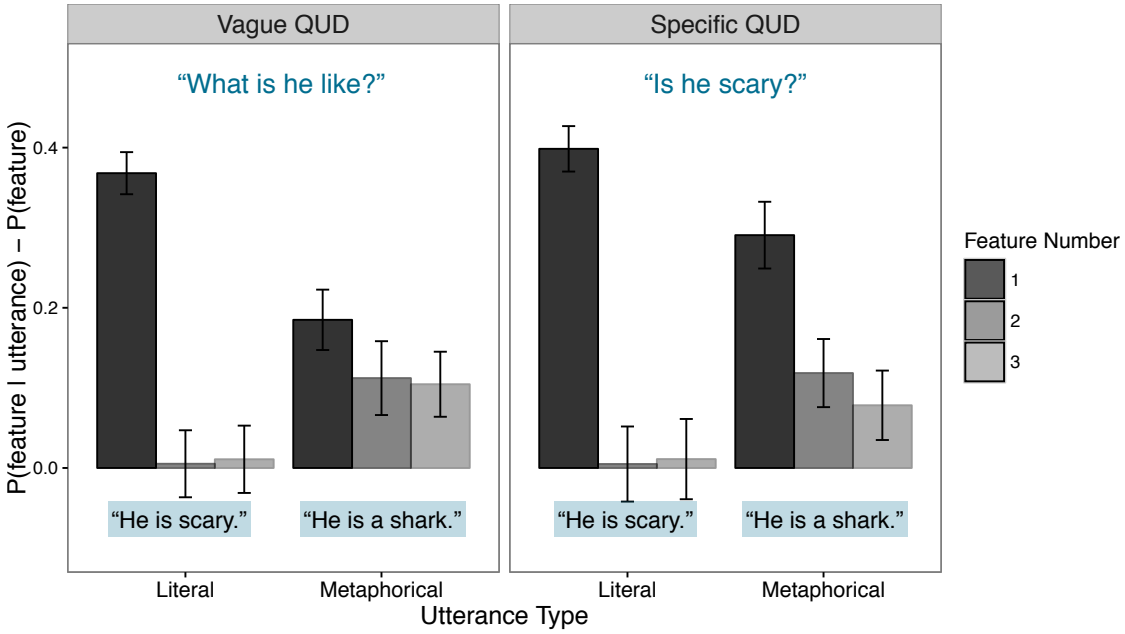

Figure 3.9: Average probability ratings for each of the three features given a vague/specific QUD and a literal/metaphorical utterance (Experiment 2c), subtracted by the average probabilities of a person having each of the features *a priori* (Experiment 2b). Error bars are 95% confidence intervals.

animal	features	alternatives
ant	small, industrious	mouse, dog, beaver, monkey
whale	big, majestic	elephant, hippo, horse, lion
bird	fast, small	cheetah, jaguar, ant, mouse
elephant	big, hard	whale, hippo, turtle, armadillo
panda	cute, big	cat, dog, elephant,whale
monkey	funny, smart	cat, dog, hyena, dolphin
penguin	funny, cute	monkey, cat, dog, kitten
giraffe	tall, long	horse, flamingo, snake, whale
cheetah	fast, agile	jaguar, leopard, monkey, cat
turtle	slow, strong	sloth, snail, elephant, horse
lion	fierce, strong	tiger, shark, elephant, horse
rabbit	fast, cute	cheetah, jaguar, cat, dog

Table 3.4: Core animals, their top two features, and four alternative animals that are strongly associated with those features.

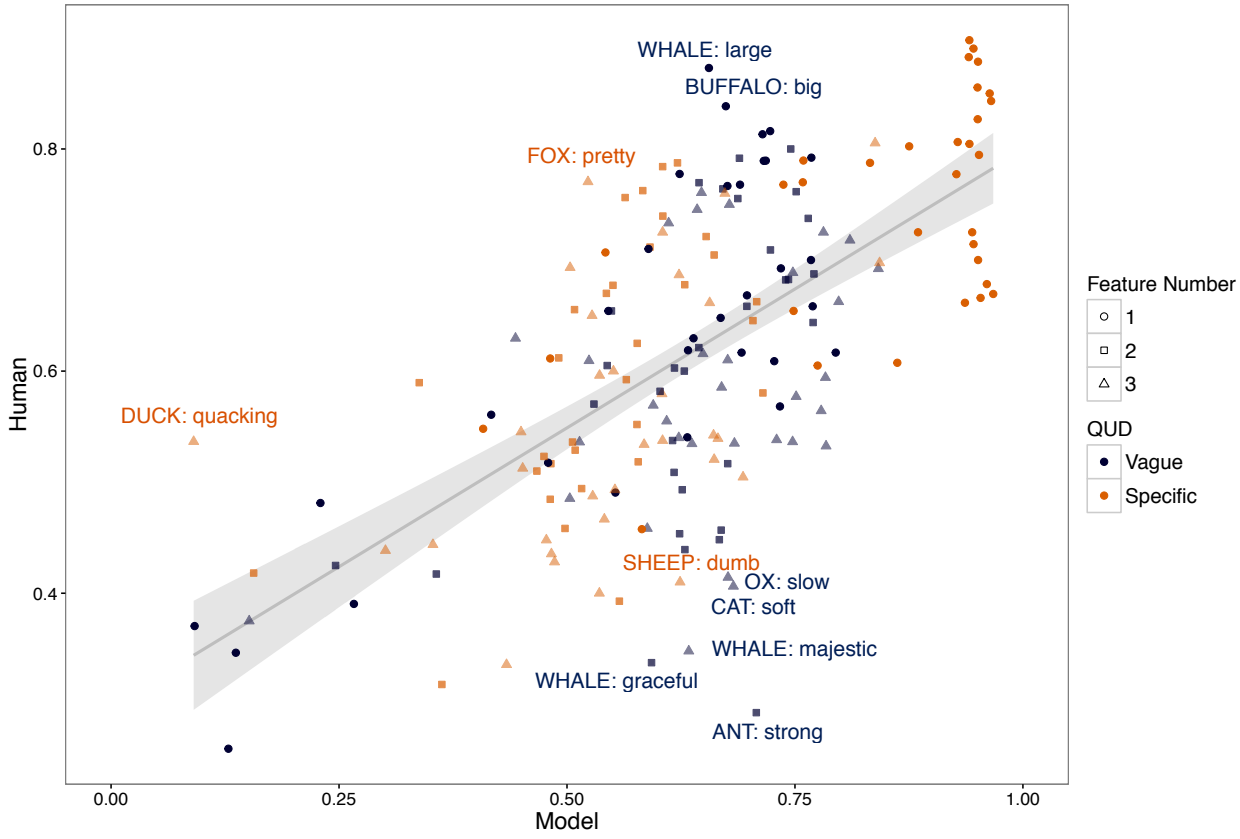

Figure 3.10: Model predictions (x axis) vs participants' probability ratings (y axis) for 192 items (32 metaphors $\times$ 3 features $\times$ 2 goal conditions). Shape of points indicates goal condition and color indicates feature number.

Chapter 4

Social Inferences and Figurative Language

People frequently use figurative language such as sarcasm and hyperbole to communicate attitudes and opinions. However, figurative utterances are often false under their literal semantics and require sufficient common ground to be appropriately understood. Why would speakers produce these types of utterances at the risk of serious misunderstanding, and what are some of the social consequences and benefits of nonliteral communication? In the last two chapters, I described how reasoning about speakers' communicative goals enables listeners to interpret figurative interpretations. In this chapter, I examine the motivations for producing figurative utterances by modeling the social inferences that arise during figurative communication. Unlike the previous chapters, this chapter does not report original behavioral experiments or data. Instead, I will reference existing literature to support the behavior of the models proposed. The goal of this chapter is to then use these models to demonstrate how pragmatic reasoning could in principle explain empirical and intuitive results regarding the relationship between figurative language and social intimacy.

4.1 Introduction

Everyday communication is full of creative and figurative uses of language. People often use nonliteral language such as hyperbole and sarcasm to communicate their attitudes, opinions, and subjective

experiences of the world, repurposing language that usually means one thing (e.g. "What a wonderful idea") to express meanings that are completely different (e.g. "That is a terrible idea"). The prevalence of figurative language is puzzling to rational accounts of communication. Why would a rational, cooperative speaker risk serious misunderstanding by producing utterances that are literally false?

A body of work in pragmatics and psycholinguistics has suggested that there may be important benefits for using figurative utterances that can offset the risk of misunderstanding, such as humor, face protection, and effects of keeping exchanges "off-record" and thus less face-threatening (R. M. Roberts & Kreuz, 1994; Brown & Levinson, 1987; Pinker, 2007; Jorgensen, 1996; Brown, 1995; Dews, Kaplan, & Winner, 1995). In addition, research has also shown interesting connections between nonliteral language and the social relationship between speaker and listener. Given that verbal irony requires a great deal of common ground to be interpreted appropriately (Clark & Gerrig, 1984; Gibbs, 1986; Pexman & Zvaigzne, 2004), Jorgensen (1996) proposed that verbal irony could be used to strengthen close relationships by highlighting shared background between interlocutors. Kreuz (1996) made the logic behind this idea explicit by proposing the principle of inferability: people are more likely to use figurative language when they are confident that the audience will interpret it correctly. Several pieces of empirical evidence regarding verbal irony appear consistent with this principle. For example, people are more likely to use sarcasm with people they are close to (Kreuz, 1996); people judge sarcasm as being more appropriate when it is used with close others (Jorgensen, 1996; Kreuz, Kassler, Coppenrath, & Allen, 1999); and utterances are read faster and rated as more ironic when the interlocutors are described as having a close relationship (Kreuz & Link, 2002). Together, these studies point to one possibility why people choose to communicate non-literally—to highlight the fact that their listeners have privileged knowledge that is required for successful nonliteral communication (Fowler et al., 1926). In other words, figurative language may be a high-reward form of communication precisely because it is high-risk, and can be used only with select audiences.

While these studies lend support to the idea that signaling common ground may be one reason why people communicate non-literally, it is unclear precisely how listeners arrive at inferences about social intimacy based on a patently neutral linguistic input. In this chapter, we clarify how

this process could work by describing a computational model that formalizes communication as recursive social reasoning between speaker and listener. We extend the Rational Speech-acts model, which has previously been shown to predict ironic interpretations (J. Kao & Goodman, 2015), to accommodate inferences about common ground between speaker and listener. We focus on verbal irony partly because it has generated the most empirical research regarding social implications, and partly because ironic utterances often run the highest risk of being interpreted literally (e.g. sarcastic utterances such as "What a wonderful idea" could be interpreted literally by an uninitiated listener, while hyperboles such as "It's a million degrees outside" are often too implausible under their literal semantics to be misinterpreted as literal).

In what follows, we use a series of model simulations to show that listeners can infer rich information about the beliefs and common ground shared with speakers when encountering sarcastic utterances. We also show that a rational speaker may choose to communicate non-literally in order to signal common ground to the listener. Finally, we discuss the implications that these findings may have on the social functions of nonliteral language.

4.2 Modeling Social Inferences

We explore a series of scenarios in which speaker and listener share varying degrees of common ground and reason about each other at varying levels of recursive depth. We use the model of verbal irony described in Chapter 3 as a starting point to show how increasingly sophisticated social inferences can arise as we extend the model to accommodate uncertainty about common ground and socially-oriented communicative goals. Instead of using weather scenarios as our example domain, here we introduce a domain that involves more subjectivity, where judgments about the same topic may vary dramatically across individuals. This inclusion of subjective beliefs allows us to explore the effect of mutual knowledge between speaker and listener on communication outcomes.

Let us make this more concrete with an example. Suppose Ann and Bob just finished watching a mainstream blockbuster movie together. Regardless of Ann's personal opinion about the movie, she may know Bob well enough to know that he is likely to feel negatively about the movie, because he doesn't like blockbuster movies in general. She may not know exactly how negative Bob feels towards this particular movie, but prior to him saying anything, she knows that his distribution over judgements about this movie is skewed towards the negative. When Bob produces an utterance, for

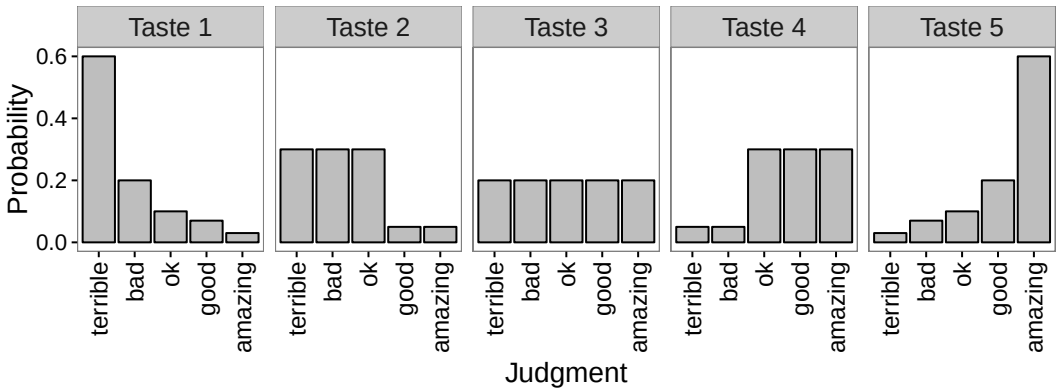

Figure 4.1: Five different "taste profiles" for mainstream blockbuster movies. Each panel is a taste profile, where the x axis represents judgement about a given blockbuster movie. Taste 1 (The Hipster): very likely to judge a given blockbuster movie as strongly negative; Taste 2 (The Picky Person): very unlikely to judge a blockbuster movie as positive; Taste 3 (The Unbiased Person): equally likely to give a blockbuster movie any judgement; Taste 4 (The Tolerant Person): very unlikely to judge a blockbuster movie negatively; Taste 5 (The Mainstream Person): very likely to judge a blockbuster movie as strongly positive.

example, "That movie was terrible," Ann interprets the utterance with respect to her knowledge of Bob's distribution. Because Bob's utterance is consistent with Ann's knowledge of his taste in movies, Ann interprets the utterance literally and infers that Bob believes the movie to be terrible. Even if Ann herself strongly believes that the movie is amazing, knowing Bob's prior disposition towards movies, she will not interpret his utterance as sarcastic. This is because she interprets the utterance not based on her own beliefs about the movie, but on her prior knowledge of *Bob's* likely beliefs about the movie. This distinction between speaker's and listener's prior beliefs is important to note for our purposes. To make the terminology consistent, in what follows we will refer to the speaker's prior probability distribution for feeling a certain way about a movie as his "taste profile."

4.2.1 Perfect common ground

We first revisit the case where the listener assumes herself to have perfect knowledge of the speaker's taste in movies, and also assumes that the speaker knows the listener to have perfect knowledge of his taste. In other words, the speaker's taste profile is in common ground. For simplicity, we assume that there are five different kinds of taste profiles about blockbuster movies in the population, as shown in Figure 4.1.

Following the qRSA model described in Chapter 3, the speaker chooses an utterance that most effectively communicates information regarding the question under discussion (QUD) to a literal listener, which could be his judgment of the movie or his emotional valence and arousal towards it. We consider a meaning space that consists of the variables j, A, where j is the judgment, and A represents the speaker's affect associated with the judgment. Identical to the verbal irony model described in Chapter 3, the pragmatic listener L_1 takes into account the speaker's prior probability of judgments and affects as well as his internal model of the speaker to interpret the utterance. The only difference is that we specify the speaker's taste profile t, from which the listener derives the prior probability of the speaker's judgment $P(j|t)$. We assume that people's taste profiles only affect the prior probability of having certain judgments towards mainstream movies. The probability of having certain affects given a judgment $P(A|j)$ is the same across taste profiles.

$$L_1(j, A|u, t) \propto P(j|t)P(A|j)\sum_q P(q)S_1(u|j, A, q, t) \tag{4.1}$$

Suppose the listener knows that the speaker's taste profile is Taste 1: Hipster. In this case, the speaker is *a priori* very likely to judge a blockbuster movie as *terrible*, moderately likely to judge it as *bad*, and unlikely to judge it as *ok, good*, or *amazing*. Given common knowledge that the speaker's taste profile is such that he is *a priori* highly likely to judge a mainstream blockbuster movie to be *terrible*, the listener will interpret the speaker's utterance "That movie was amazing" sarcastically to mean "That movie was terrible" (Figure 4.2). Overall, this model is able to reason about the speaker's communicative goals and use common knowledge about his taste profile to produce interpretations that are consistent with intuition.

4.2.2 Inferring speaker's priors

In the last section, we showed that the model uses common ground regarding the speaker's taste profile to produce sarcastic and otherwise nonliteral interpretations. In this section, we examine a situation where the listener has no prior knowledge of the speaker's taste profile, and explore the inferences that she makes given different utterances.

Suppose Amy and her blind date Ben just finished watching a mainstream blockbuster movie. Amy doesn't know Ben's taste in movies and has no idea how he might feel about the movie.

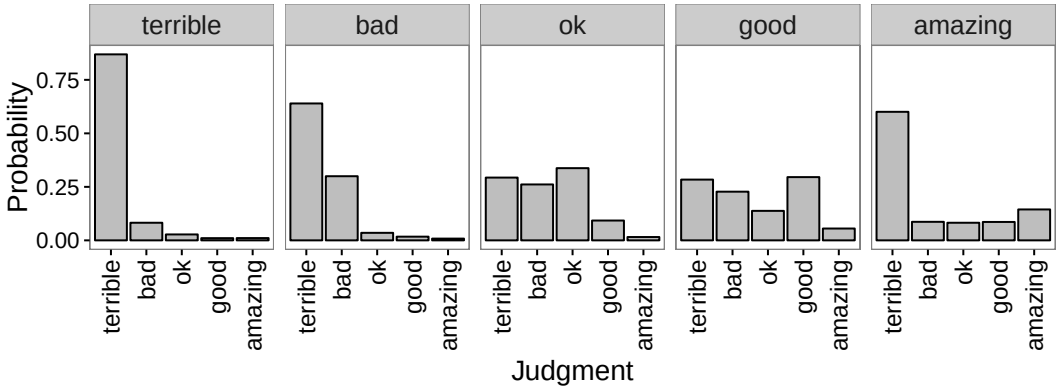

Figure 4.2: Interpretation of utterances given common knowledge that the speaker's taste profile is "Taste 1: Hipster" (see Fig. 4.1). Each panel is the interpretation of an utterance. The x axis represents the speaker's judgment.

Ben turns to Amy and says, "That movie was amazing." Intuitively, Amy receives two pieces of information from this utterance: Ben judges the movie to be *amazing*, and Ben's taste in movies is probably such that he is fond of these types of movies in general*. To model these intuitive inferences about Ben's judgment and his taste profile, we introduce a simple modification to Equation 4.1:

$$L_1(j, A, t|u) \propto P(t)P(j|t)P(A|j)\sum_q P(q)S_1(u|j, A, q, t) \tag{4.2}$$

Instead of being known to the listener, the variable t now needs to be inferred given the utterance. Note that the speaker's choice of utterance $S_1(u|j, A, q, t)$ also depends on the listener's belief about his taste profile, t. Since Amy has no prior knowledge about her blind date's taste profiles, $P(t)$ is uniform.

Figure 4.3 shows Amy's inferences about Ben's judgement about the movie given different utterances he says. Across all five utterances, given no prior knowledge of Ben's taste profile, Amy is most likely to interpret each utterance as literal, such that "That movie was amazing" is interpreted as literally meaning that Ben judges the movie to be amazing. We can also query the model for Amy's inferences about Ben's taste profile given each utterance (Figure 4.4). If Ben says "That movie was amazing," Amy will end up with the posterior belief that Ben is likely to have a mainstream taste profile.

*Another possibility is that Ben is simply trying to be agreeable and polite; however, that is a different type of social goal that is beyond the scope of this chapter.

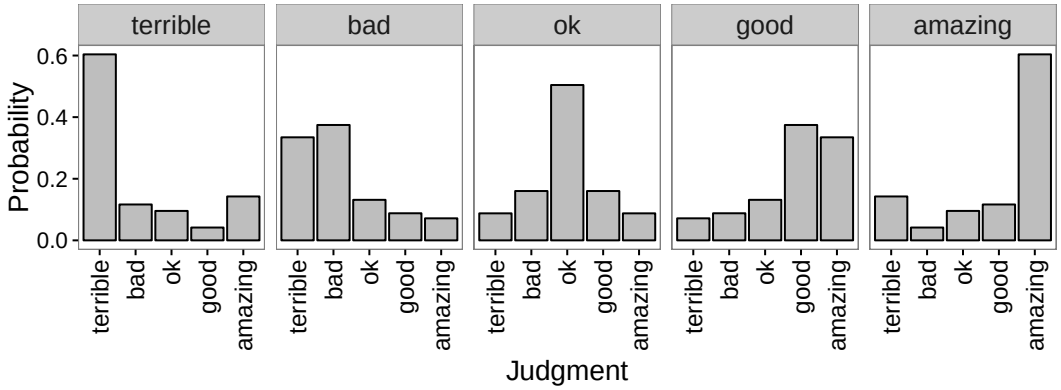

Figure 4.3: Interpretation of utterances given an uninformative prior over the speaker's taste profiles. Each panel is the interpretation of an utterance. The x axis represents the speaker's judgment. The most likely interpretation of each utterance is the literal one.

Contrasting Amy's interpretation of the utterance "That movie was amazing" in Figure 4.2 with the interpretation of the same utterance in Figure 4.3, we find that extending the model to accommodate inferences about the speaker's taste profile reveals an interesting intuition about figurative utterances. Unless Ben is highly confident that Amy knows his taste profile to be "hipster," it is unwise for him to say "That movie was amazing" and expect to be interpreted sarcastically, because Amy would end up with both the wrong interpretation and the wrong inference about his tastes. This model result is consistent with the finding that people are more likely to use verbal irony with close others, who presumably have more knowledge about the speakers' prior dispositions (Kreuz, 1996).

More interestingly, contrasting the model described here (where the listener has no prior knowledge of the speaker's taste profile) and the one in the previous section (where the speaker's taste profile is in common ground), we can reason as follows. If the listener has no prior knowledge of the speaker's taste profile, she will interpret utterances literally. If the speaker's taste profile is in common ground, however, the listener will produce appropriate nonliteral interpretations. As a rational speaker, Ben should only say "That movie was amazing" sarcastically if he believes his taste profile to be in common ground. In the next section, we allow the pragmatic listener to reason about the speaker at a higher level of recursion and show that she is able to use this intuition to produce inferences about the speaker's beliefs about common ground given his utterance.

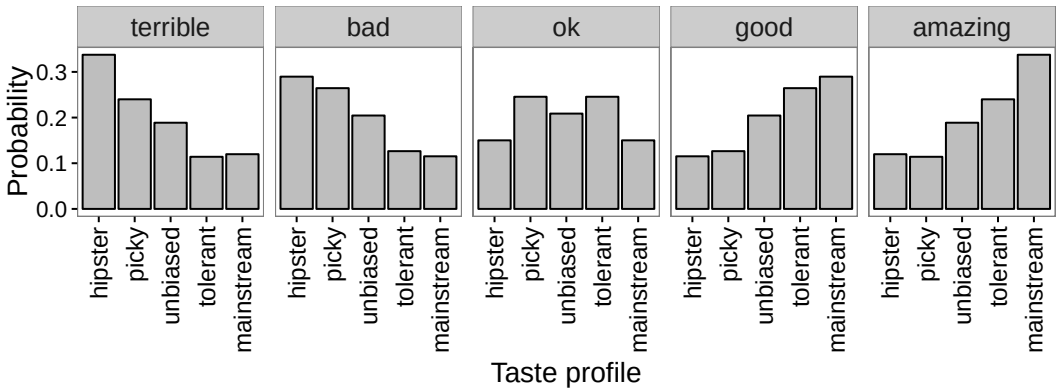

Figure 4.4: Inferences about the speaker's taste profile given different utterances. Each panel represents the posterior distribution over the speaker's taste profiles given an utterance, where the x axis specifies the taste profile described in Fig. 4.1.

4.2.3 Inferring speaker's beliefs about common ground

Suppose Alice and her husband Bill just finished watching a mainstream blockbuster movie. Alice knows Bill's taste in movies very well. She is certain that his taste profile is best described by "Taste 1: Hipster," which means he is likely to judge the movie to be terrible. Bill turns to Alice and says, "That movie was amazing."

In this scenario, Alice could do one of two things. She could interpret the utterance literally, which would require her to revise her beliefs about Bill's taste in movies and question how well she really knows her husband of several decades. Alternatively, Alice could interpret Bill's utterance as sarcastic, which seems intuitively more likely. Interpreting Bill's utterance sarcastically has several implications. If Bill indeed intended to be sarcastic, and if he is a rational, cooperative speaker, he would only choose to be sarcastic with the assumption that Alice knows his taste profile well enough to interpret his utterance correctly. This leads Alice to believe that Bill knows that she knows his taste profile—in other words, she will believe that Bill believes his taste profile to be in common ground. Given such an utterance, Alice still ends up with the initial belief that Bill has a hipster's taste profile; however, she now gains important information about Bill's confidence in her knowledge of him.

Let us formalize this intuition by further extending the model described in the last section. Because Alice now reasons about Bill reasoning about her pragmatic interpretation, Alice needs to

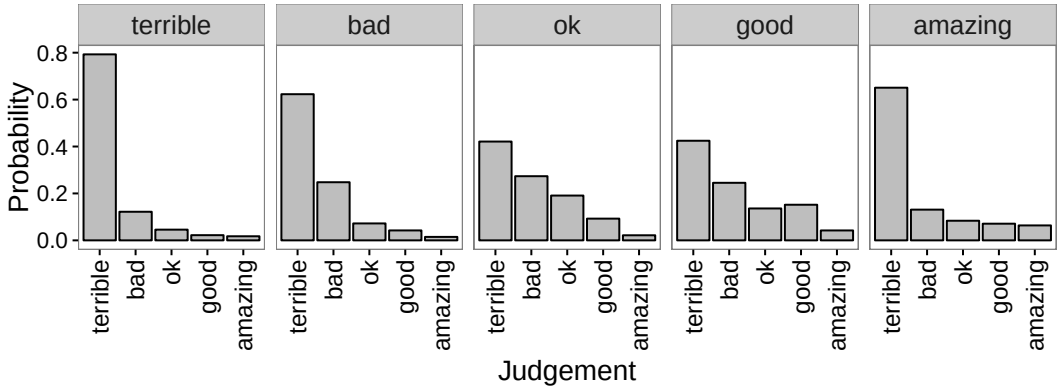

Figure 4.5: Interpretation of utterances given the listener knows that the speaker's taste profile is "Taste 1: Hipster," but with uncertainty about whether this knowledge is in common ground. Each panel represents an utterance.

reason one level further than Ann did in the last section, and will be modeled as an L_2 listener*. We now introduce a boolean variable ζ, which indicates whether the speaker believes his taste profile to be in common ground with the listener. This variable allows us to capture the intuition that while Alice is certain that Bill has a particular taste profile t, she is uncertain about whether Bill believes his taste profile to be in common ground. Alice's inferences can be described by the following equation:

$$L_2(j, A, \zeta|u, t) \propto \begin{cases} P(\zeta)P(j|t)P(A|j)\sum_q P(q)S_2(u|j, A, q, t) & \text{if } \zeta = 1 \\ P(\zeta)P(j|t)P(A|j)\sum_{t',q} P(t')P(q)S_2(u|j, A, q, t') & \text{if } \zeta = 0 \end{cases} \tag{4.3}$$

where Alice's belief about Bill's taste profile t is passed down and used by S_2 if Bill believes it to be in common ground, and is resampled uniformly as t' from all five possible taste profiles if not. In each case, L_2 reasons about S_2's choice of utterance given L_1's beliefs about Bill's taste profile.

Figure 4.5 shows L_2's interpretation of each utterance given knowledge of the speaker's taste profile, but with uncertainty about whether the speaker believes it to be in common ground. We see that L_2 again successfully interprets the utterance "That movie was amazing" as sarcastically meaning that the speaker thinks the movie was terrible. More interestingly, L_2 can also make inferences about the variable ζ—whether the speaker believes his taste profile to be in common

*Although higher levels of recursion have not been necessary in previous RSA models, given that the phenomenon we are interested in is a higher-order social inference derived from pragmatic interpretation, an L_2 level model seems warranted.

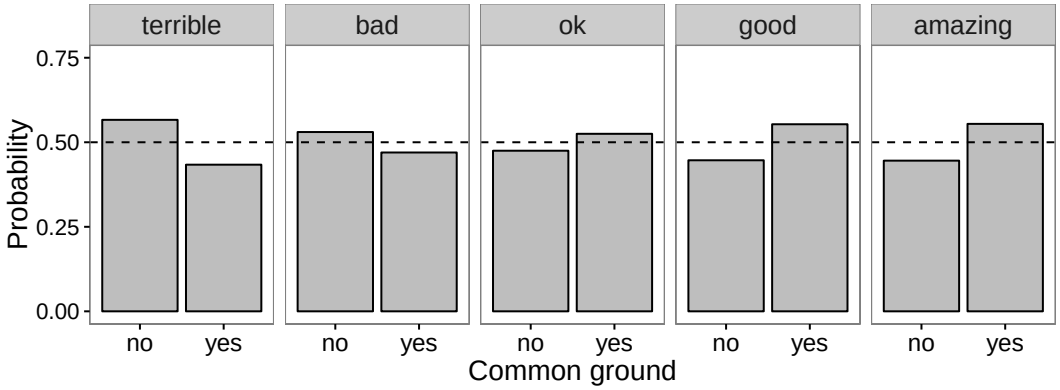

Figure 4.6: Inferences about the probability that the speaker believes his taste profile to be in common ground given different utterances. Each panel represents an utterance. The prior probability of common ground is set to be 0.5.

ground—given different utterances (Figure 4.6). With the prior probability of common ground set as 0.5, L_2 infers a higher posterior probability that the speaker believes his taste profile to be in common ground given the sarcastic utterance "That movie was amazing," and a lower posterior probability of common ground given the literal utterance "That movie was terrible." This results in a formal derivation of the the social inferences licensed by nonliteral, sarcastic uses of language: a strengthened belief that the speaker believes his taste profile to be in common ground, which in turn implies that he believes his relationship with the listener to be close.

4.2.4 Communicating beliefs about common ground

Thus far in this chapter, we have shown the inferences that listeners can make given literal or sarcastic utterances. In this section, we explore the motivations that could lead a speaker to choose a sarcastic versus a literal utterance. In particular, we show how a higher-level pragmatic speaker S_3 could choose utterances in order to communicate common ground to the listener, specified as follows:

$$S_3(u|j, A, t, \zeta) \propto e^{\lambda U_3(u|j,A,t,\zeta)} \tag{4.4}$$

where the utility function is defined as the negative surprisal of ζ under L_2's interpretation distribution:

$$U_3(u|j, A, t, \zeta) = logL_2(\zeta|u, t) \tag{4.5}$$

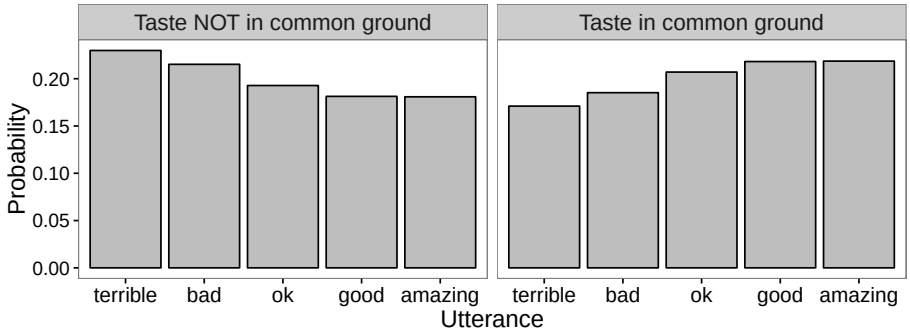

Figure 4.7: Probability that speaker would choose an utterance, given whether he wants to communicate to the listener that he believes his taste profile (Taste 1: Hipster) is in common ground. The speaker is more likely to choose nonliteral utterances when he wants to communicate to the listener that his taste profile is in common ground.

Intuitively, S_3 chooses an utterance to maximize the probability that L_2 will infer the correct common ground variable ζ given the speaker's taste profile. Figure 4.7 shows that the speaker is more likely to choose a nonliteral utterance (in this case, a positive utterance) in order to communicate his belief that his taste profile is in common ground (in this case, a taste profile that is strongly skewed towards negative judgments).

4.3 Discussion

In this chapter, we described a series of model simulations that show how inferences about common ground and social intimacy could be both a natural consequence of and a motivation for using figurative language. The models we present contain natural but critical extensions to the RSA models introduced in earlier chapters, highlighting the flexibility and generality of the modeling framework. The recursive nature of RSA models allows us to demonstrate how uncertainty about common ground could affect language interpretation, and in turn how interpretation constrains inferences about common ground.

While these models shed light on the social motivations for nonliteral communication and indirect language more generally, much work remains to be done to elucidate the details of these social inferences and the specific contexts that facilitate them. In the simulations presented here, we chose simple and idealized scenarios where the listener either has perfect knowledge of the speaker's taste profile or no knowledge at all. In real-world situations, listeners are usually somewhere in between,

juggling beliefs about speakers while simultaneously reasoning about speakers' beliefs about *them*, all under a great deal of uncertainty. It is possible that under these circumstances, other social and linguistic cues are necessary to guide listeners' inferences about what speakers really mean. Indeed, these cues are abundant in many exchanges involving figurative language (Clark & Gerrig, 1984; Caucci & Kreuz, 2012). In the scenario with Amy and her blind date Ben, Ben could accompany the utterance "That movie was amazing" with an eye-roll and a laugh, which would signal to Amy that he is being playful and sarcastic, even if she has no prior knowledge about his taste profile. The specific ways in which such paralinguistic cues factor into pragmatic inferences remains to be examined in future research. On the other hand, our models suggest that common ground allows speakers to use figurative language successfully without the aid of paralinguistic cues, which predicts that sarcastic utterances are most often produced in a "deadpan," unmarked manner among very close friends. This prediction is supported by anecdotal evidence, and also by research on Twitter showing that illocutionary markers such as #sarcasm are used more often with a distant audience (Bamman & Smith, 2015). Without detailed behavioral data to evaluate our models, we can only speculate that the kind of pragmatic reasoning described here is at least partially responsible for the range of phenomena we see in figurative communication in the wild.

Finally, our models make an interesting connection to pretense theory, a classic theory of verbal irony. Clark and Gerrig (1984) first described the pretense theory of irony and attributed its roots to Grice (1975), who wrote: "To be ironical is, among other things, to pretend (as the etymology suggests), and while one wants the pretense to be recognized as such, to announce it as a pretense would spoil the effect" (p. 125). Clark (1997) describes pretense as a kind of "layering," where the speaker describes and refers to layers of action that are not grounded in the present reality, as in the case in story-telling and other kinds of performances (Clark, 1996). We can imagine that the recursive reasoning at the heart of RSA models represents a type of layering, where higher-level agents simulate the behaviors and interpretations of lower-level agents who are performing actions at different layers. When a higher-level speaker produces a sarcastic utterance such as "That movie was amazing," he is simulating (or pretending to be) a person who would find the movie to be amazing, and imagining a listener who would interpret the utterance literally. He also recognizes that a higher-level pragmatic listener would use her knowledge of the speaker to interpret the utterance sarcastically and "be in on" the joint pretense. While the connection at the moment is admittedly

tenuous, the RSA modeling framework may be able to formalize the ideas behind pretense theory in an intuitively satisfying manner.

We believe the work described here is a promising step towards linking creative language use with social motivations such as increasing closeness and solidarity. Building upon earlier models of pragmatic reasoning and introducing natural extensions through inferences about common ground, this work provides a unified computational basis for a range of social effects in figurative language use. By reasoning about speakers' communicative goals and beliefs about common ground, the RSA model shows how figurative language could accomplish one of the most important goals in communication of all—bringing people closer together.

Chapter 5

A Computational Model of Humor in Puns

Creative uses of language are associated with several desirable consequences. In the last chapter, I used the Rational Speech-acts modeling framework to explore how figurative language could communicate speakers' beliefs about common ground and social intimacy. In this chapter, I will use a different modeling approach to address a distinctive consequence of creative language: humor. While humor plays an essential role in human interactions, precisely what makes something funny remains elusive. Here we propose two information-theoretic measures—ambiguity and distinctiveness—derived from a simple model of sentence processing. We test these measures on a set of puns and regular sentences and show that they correlate significantly with human judgments of funniness. Moreover, within a set of puns, the distinctiveness measure distinguishes exceptionally funny puns from mediocre ones. Our work is the first, to our knowledge, to integrate a computational model of general language understanding and humor theory to quantitatively predict humor at a fine-grained level. We present it as an example of a framework for applying models of language processing to understand higher-level linguistic and cognitive phenomena*.

*This chapter is based on J. T. Kao, Levy, and Goodman (2015)

5.1 Introduction

Love may make the world go round, but humor is the glue that keeps it together. Our everyday experiences serve as evidence that humor plays a critical role in human interactions and composes a significant part of our linguistic, cognitive, and social lives. Previous research has shown that humor is ubiquitous across cultures (Martin, 2010; Kruger, 1996), increases interpersonal attraction (Lundy, Tan, & Cunningham, 1998), helps resolve intergroup conflicts (W. J. Smith, Vernard Harrington, & Neck, 2000), and improves psychological wellbeing (Martin, Kuiper, Olinger, & Dance, 1993). However, little is known about the cognitive basis of such a pervasive and enjoyable experience. By providing a formal model of linguistic humor, we aim to solve part of the mystery of what makes us laugh.

Theories of humor have existed since the time of Plato and Aristotle (see Attardo (1994) for review). A leading theory in modern research posits that incongruity, loosely characterized as the presence of multiple incompatible meanings in the same input, may be critical for humor (Veale, 2006; Forabosco, 1992; Hurley, Dennett, & Adams, 2011; MacGhee & Pistolesi, 1979; Vaid & Ramachandran, 2001). However, despite relative consensus on the importance of incongruity, definitions of incongruity vary across informal analyses of jokes. As Ritchie (2009) wrote, "There is still not a rigorously precise definition that would allow an experimenter to objectively determine whether or not incongruity was present in a given situation or stimulus" (p. 331). This lack of precision makes it difficult to empirically test the role of incongruity in humor or extend these ideas to a concrete computational understanding. On the other hand, most work on computational humor focuses either on joke-specific templates and schemata (Binsted, 1996; J. Taylor & Mazlack, 2004) or surface features and properties of individual words (Mihalcea & Strapparava, 2006; Kiddon & Brun, 2011; Reyes, Rosso, & Buscaldi, 2012). One exception is Mihalcea, Strapparava, and Pulman (2010), which used features inspired by incongruity theory to detect humorous punch lines; however, the incongruity features proposed did not significantly outperform a random baseline, leading the authors to conclude that joke-specific features may be preferable. While these dominant approaches in computational humor are able to identify humorous stimuli within certain constraints, they fall short of testing a more general cognitive theory of humor.

In this work, we suggest that true measures of incongruity in linguistic humor may require a model that infers meaning from words in a principled manner. We build upon theories of humor

and language processing to formally measure the multiplicity of meaning in puns -- sentences "in which two different sets of ideas are expressed, and we are confronted with only one series of words," as described by Philosopher Henri Bergson (Bergson, 1914). Puns provide an ideal test bed for our purposes because they are simple, humorous sentences with multiple meanings. Here we focus on phonetic puns, defined as puns containing words that sound identical or similar to other words in English. The following is an example:

(1) "The magician got so mad he pulled his hare out."

Although the sentence's written form unambiguously contains the word "hare," previous work has suggested that phonetic representations play a central role in language comprehension even during reading (Niznikiewicz & Squires, 1996; Pexman, Lupker, & Jared, 2001; Pollatsek, Lesch, Morris, & Rayner, 1992). Taking the lexical ambiguity of its phonetic form into account, this sentence thus implicitly expresses two "ideas," or meanings*:

(1a) The magician got so mad he performed the trick of pulling a rabbit out of his hat.

(1b) The magician got so mad he pulled out the hair on his head.

At the most basic level, the humor in this pun relies on the fact that it contains the word "hare," which is phonetically confusable with "hair." However, the following sentence also contains a phonetically ambiguous word, but is clearly not a pun:

(2) "The hare ran rapidly across the field."

A critical difference between (1) and (2) is that *hare* and *hair* are both probable meanings in the context of sentence (1), whereas *hare* is much more likely than *hair* in sentence (2). From this informal analysis, it seems that both meanings are compatible with context in a phonetic pun, suggesting that a sentence must contain ambiguity to be funny. However, another example shows that ambiguity alone is insufficient. Consider the sentence:

(3) "Look at that hare."

This sentence is also ambiguous between *hare* and *hair*, but is unlikely to elicit chuckles. A critical difference between (1) and (3) is that while each meaning is strongly supported by distinct groups of words in (1) (*hare* is supported by "magician" and "hare"; *hair* is supported by "mad" and "pulled"), both meanings are weakly supported by all words in (3). This comparison suggests that in addition to ambiguity, distinctiveness of support may also be an important criterion for humor.

*In this work we focus on written sentences that contain phonetic ambiguity. In the future, it would be interesting to examine humorous effects in spoken sentences, where ambiguity cannot be partially resolved by the orthographic form.

Observations on the putative roles of ambiguity of sentence meaning and distinctiveness of support will motivate our formal measures of humor. *

How should we represent the meaning of a sentence in order to measure its ambiguity and distinctiveness? While formally representing sentence meanings is a complex and largely unsolved problem (Grefenstette, Sadrzadeh, Clark, Coecke, & Pulman, 2014; Socher, Huval, Manning, & Ng, 2012; Liang, Jordan, & Klein, 2013), we can utilize certain properties of phonetically ambiguous sentences to simplify the problem. We notice that in sentence (1), meaning (1a) arises if the word "hare" is interpreted as *hare*, while meaning (1b) arises if "hare" is interpreted as its homophone *hair*. Each sentence-level meaning directly corresponds to the meaning of a phonetically ambiguous word. As a result, we can represent sentence meaning (1a) with *hare* and (1b) with *hair*. This approximation is coarse and captures only the "gist" of a sentence rather than its full meaning. However, we will show that it is sufficiently powerful for modeling the interpretation of sentences with only a phonetic ambiguity.

Given the space of candidate sentence meanings, a comprehender's task is to infer a distribution over these meanings from the words she observes. Formally, a phonetically ambiguous sentence such as (1) is composed of a vector of words $\overrightarrow{w} = \{w_1, \ldots, w_i, h, w_{i+1}, \ldots, w_n\}$, where h is phonetically confusable with its homophone *h'*. The sentence meaning is a latent variable m, which we assume has two possible values $m_a$and m_b. These sentence meanings can be identified with h and *h'*, respectively. Consistent with a noisy channel approach (Levy, 2008; Levy, Bicknell, Slattery, & Rayner, 2009; Gibson et al., 2013), we construe the task of understanding a sentence as inferring m using probabilistic integration of noisy evidence given by $\overrightarrow{w}$. We construct a simple probabilistic generative model that captures the relationship between the meaning of a sentence and the words that compose it (Figure 5.1). If a word is semantically relevant ($f_i = 1$), we assume that it is sampled based on semantic relatedness to the sentence meaning; if the word is irrelevant, or "noise," it only reflects general language statistics and is sampled from an n-gram model. Because the comprehender maintains uncertainty about which words are relevant, it is possible for her to arrive at multiple interpretations of a sentence that are each coherent but incongruous with one another, a situation that we hypothesize gives rise to humor. To capture this intuition, we introduce two measures of

*Note that it is not necessary for both meanings to be completely compatible with the full context, as illustrated by puns such as *I used to be addicted to soap, but I'm clean now*, in which the most common meaning of *clean* is actually ruled out, rather than supported, by full compositional interpretation of the context. What instead seems necessary is that the support derived from the subset of context for each meaning is balanced.

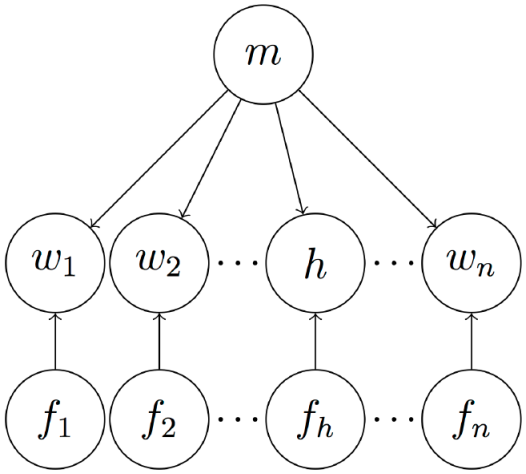

Figure 5.1: Graphical representation of a generative model of a sentence. If the indicator variable f_i has value 1, w_i is generated based on semantic relatedness to the sentence meaning m; otherwise, w_i is sampled from a trigram language model based on the immediately preceding two words.

humor derived from the distribution over sentence meanings (details in Methods section).

Given words in a sentence, we infer the joint probability distribution over sentence meanings and semantically relevant words, which can be factorized into the following:

$$P(m, \vec{f}\,|\vec{w}) = P(m\,|\vec{w})P(\vec{f}|m, \vec{w}) \tag{5.1}$$

We compute a measure of humor from each of the two terms on the right-hand side. Ambiguity is quantified by the entropy of the distribution $P(m|\vec{w})$. If entropy is high, then the sentence is ambiguous because both meanings are near-equally likely. Distinctiveness captures the degree to which the relevant words differ given different sentence meanings. Given one meaning m_a, we compute $F_a = P(\vec{f}\,|\,m_a, \vec{w})$. Given another meaning m_b, we compute $F_b = P(\vec{f}\,|\,m_b, \vec{w})$. Distinctiveness is quantified by the symmetrized Kullback-Leibler divergence between these two distributions, $D_{\mathrm{KL}}(F_a\,||F_b) + D_{\mathrm{KL}}(F_b|||F_a)$. If the symmetrized KL distance is high, it suggests that the two sentence meanings are supported by distinct subsets of words in the sentence. Derivation details of these two measures are in the Methods section below. We empirically evaluate ambiguity and distinctiveness as predictors of humor in a set of phonetically ambiguous sentences.

5.2 Methods

5.2.1 Computing model predictions

We define the ambiguity of a sentence as the entropy of $P(m|\vec{w})$, where $\vec{w}$ is a vector of observed content words in a sentence (which contains a phonetically ambiguous word h and m is the latent sentence meaning. Given the simplifying assumption that the distribution over sentence meanings is not affected by function words, each w_i in $\vec{w}$ is a content word. The distribution over sentence meanings given words can be derived using Bayes' rule:

$$\begin{aligned} P(m|\vec{w}) &= \sum_{\vec{f}} P(m, \vec{f}|\vec{w}) \\ &\propto \sum_{\vec{f}} P(\vec{w}|m, \vec{f}) P(m) P(\vec{f}) \\ &= \sum_{\vec{f}} \left(P(m) P(\vec{f}) \prod_i P(w_i|m, f_i) \right) \end{aligned} \tag{5.2}$$

Each value of m is approximated by either the meaning of the observed phonetically ambiguous word h (e.g. "hare" in sentence (1)) or its unobserved homophone h' (e.g. "hair"). We can thus represent $P(m)$ as the unigram frequency of h or h'. For example, $P(m = hare)$ is approximated as proportional to $P(\text{"hare"})$. We assume equal prior probability that each subset of the words is semantically relevant, hence $P(\vec{f})$ is a constant. $P(w_i|m,\ f_i)$ depends on the value of the semantic relevance indicator variable f_i. If $f_i = 1$, w_i is semantically relevant and is sampled in proportion to its relatedness with the sentence meaning m. If $f_i = 0$, then w_i is generated from a noise process and sampled in proportion to its probability given the previous two words in the sentence. Formally,

$$P(w_i|m, f_i) = \begin{cases} P(w_i|m) & \text{if} f_i = 1 \\ P(w_i|\text{bigram}_i) & \text{if} f_i = 0 \end{cases} \tag{5.3}$$

We estimate $P(w_i|m)$ using empirical association measures described in the Experiment section and compute $P(w_i|\text{bigram}_i)$ using the Google N-grams corpus (Brants & Franz, 2006). Once we

derive $M = P(m|\vec{w})$, we compute its information-theoretic entropy as a measure of ambiguity:

$$\text{Ambiguity}(M) = -\sum_{k\in\{a,b\}} P(m_k|\vec{w}) \log P(m_k|\vec{w}) \tag{5.4}$$

We next compute the distinctiveness of words supporting each sentence meaning. Using Bayes' Rule:

$$P(\vec{f}|m, \vec{w}) \propto P(\vec{w}|m, \vec{f})P(\vec{f}|m) \tag{5.5}$$

Since $\vec{f}$ and m are independent, $P(\vec{f}|m) = P(\vec{f})$, which is a constant. Let $F_a = P(\vec{f}|m_a, \vec{w})$ and $F_b = P(\vec{f}|m_b, \vec{w})$. We compute the symmetrized Kullback-Leibler divergence score $D_{\text{KL}}(F_a||F_b) + D_{\text{KL}}(F_b||F_a)$, which measures the difference between the distribution of supporting words given one sentence meaning and the distribution of supporting words given another sentence meaning. This results in the distinctiveness measure*:

$$\text{Distinctiveness}(F_a, F_b) = \sum_i \left(\ln\left(\frac{F_a(i)}{F_b(i)}\right) F_a(i) + \ln\left(\frac{F_b(i)}{F_a(i)}\right) F_b(i) \right) \tag{5.6}$$

Given these derivations, we conducted the following experiment to implement and test the ambiguity and distinctiveness measures.

5.2.2 Experiment

We collected 435 sentences consisting of phonetic puns and regular sentences that contain phonetically ambiguous words. We obtained the puns from a website called "Pun of the Day" (`http://www.punoftheday.com`), which at the time of collection contained over a thousand puns submitted by users. We collected 40 puns where the phonetically ambiguous word has an identical homophone, for example "hare." Since only a limited number of puns satisfied this criterion, a research assistant generated an additional 25 pun sentences based on a separate list of homophone words, resulting in a total of 65 identical-homophone puns. We selected 130 corresponding non-pun sentences from an online version of Heinle's Newbury House Dictionary of American English (`http://nhd.heinle.com`). 65 of the non-pun sentences contain the ambiguous words observed in the pun sentences (e.g. "hare");

*In addition to the symmetrized KL divergence of Eq. 6, we also experimented with non-symmetrized KL divergence in both directions and found qualitatively identical results.

Homophone	Type	Example
Identical	Pun	The magician was so mad he pulled his hare out.
Identical	Non-pun	The hare ran rapidly across the field.
Identical	Non-pun	Some people have lots of hair on their heads.
Near	Pun	A dentist has to tell a patient the whole tooth.
Near	Non-pun	A dentist examines one tooth at a time.
Near	Non-pun	She always speaks the truth.

Table 5.2: Example sentence from each category. Identical homophone sentences contain phonetically ambiguous words that have identical homophones; near homophone sentences contain phonetically ambiguous words that have near homophones. Pun sentences were selected from a pun website; non-pun sentences were selected from an online dictionary (see main text for details).

the other 65 contain the unobserved homophone words (e.g. "hair")*. To test whether our measures generalize to sentences containing phonetically ambiguous words that do not have identical homophones, we collected 80 puns where the phonetically ambiguous word sounds similar (but not identical) to other words in English (e.g. "tooth" sounds similar to "truth"). We also collected 160 corresponding non-pun sentences. Table 5.2 shows an example sentence from each category†.

We obtained funniness ratings for each of the 435 sentences. 100 participants on Amazon's Mechanical Turk‡ rated the 195 sentences that contain identical homophones. Each participant read roughly 60 sentences in random order, counterbalanced for the sentence types, and rated each sentence on funniness ("How funny is this sentence?") on a scale from 1 (not at all) to 7 (extremely). We removed 7 participants who reported a native language other than English and z-scored the ratings within each participant. A separate group of 160 participants on Mechanical Turk rated the 240 near homophone sentences. Each participant read 40 sentences in random order, counterbalanced for the sentence types, and rated each sentence on funniness on a scale from 1 to 7. We removed 4 participants who reported a native language other than English and z-scored the ratings within each participant. We used the average z-scored ratings across participants as human judgments of funniness for all 435 sentences.

As described in the measure derivations, computing ambiguity and distinctiveness requires the

*Results for the 195 identical homophone sentences were reported in J. T. Kao, Levy, and Goodman (2013), which was published in the proceedings of the 35th Annual Meeting of the Cognitive Science Society (a non-archival publication).

†The full set of sentences can be found here: `http://web.stanford.edu/~justinek/pun-paper/results.html`

‡The sample sizes were chosen such that each sentence would receive roughly 20-30 funniness ratings, in order for the uncertainty in funniness measurement to be reasonably low, while keeping the number of sentences rated by each participant manageably small.

conditional probabilities of each word given a sentence meaning, i.e. $P(w_i \mid m)$. In practice, this value is difficult to obtain reliably and accurately in an automated way, such as through WordNet distances or semantic vector space models (Gabrilovich & Markovitch, 2007; Zhang, Gentile, & Ciravegna, 2011; Mihalcea et al., 2010)*. Instead of tackling the challenging problem of automatically learning $P(w_i \mid m)$ from large corpora, we observe that $P(w_i \mid m)$ is related to point wise mutual information (PMI) between w_i and m, an information-theoretic measure defined mathematically as the following:

$$\log \frac{P(w_i,\ m)}{P(w_i)\,P(m)} = \log P(w_i|m) - \log P(w_i) \tag{5.7}$$

Intuitively, PMI captures the relatedness between w_i and m, which can be measured empirically by asking people to judge the semantic relatedness between two words. This allows us to harness people's rich knowledge of the relationships between word meanings without relying solely on co-occurrence statistics in corpora. We assume that the z-scored human ratings of relatedness between two words, denoted $\mathrm{R}(w_i,\ m)$, approximates true PMI. With the proper substitutions and transformations† from Eq. 7, we derive the following:

$$P(w_i \mid m) = \ e^{R(w_i,\ m)} P(w_i) \tag{5.8}$$

To obtain $R(w_i,\ m)$ for each of the words in the stimuli sentences, function words were removed from each of the sentences in our dataset, and the remaining words were paired with the phonetically ambiguous word h and its homophone h' (e.g., for the pun in Table, ["magician", "hare"] is a legitimate word pair, as well as ["magician", "hair"]). This resulted in 1460 distinct word pairs for identical homophone sentences and 2056 word pairs for near homophone sentences. 200 participants on Amazon's Mechanical Turk rated the semantic relatedness of word pairs for identical homophone sentences. Each participant saw 146 pairs of words in random order and were asked to rate how related each word pair is using a scale from 1 to 10. We removed 5 participants who reported a native language other than English. A separate group of 120 participants rated word pairs for near homophone sentences. We removed 2 participants who reported a native language other than

*We experimented with computing these values from corpora in early stages of this work. However, we found that it is difficult to obtain reliable co-occurrence statistics for many word pairs of interest (such as "hare" and "magician"), due to the sparsity of these topics in most corpora. Future work could further explore methods for extracting these types of commonsense-based semantic relationships from corpus statistics.

†By assuming $R(w_i,\ m) = \log \frac{P(w_i,\ m)}{P(w_i)P(m)}$, we get $R(w_i,\ m) = \log P(w_i|m) - \log P(w_i)$ from Eq. 7; exponentiating both sides gives us Eq. 8.

	Estimate	Std. Error	p-value
Intercept	-2.139	0.306	< 0.0001
Ambiguity	1.915	0.221	< 0.0001
Distinctiveness	0.264	0.040	< 0.0001

Table 5.4: Regression coefficients using ambiguity and distinctiveness to predict funniness ratings for all 435 sentences; p-values are computed assuming that the t statistic is approximately normally distributed.

English. Since it is difficult to measure the relatedness of a word with itself, we assume that it is constant for all words and treat it as a free parameter, r. After computing our measures, we fit this parameter to people's funniness judgments (resulting in $r = 13$). We used the average z-scored relatedness measure for each word pair to obtain $R(w_i,\ m)$ and Google Web unigrams to obtain $P(w_i)$. This allowed us to compute $P(w_i \mid m)$ for all word and meaning pairs.

5.3 Results

We computed an ambiguity and distinctiveness score for each of the 435 sentences (see Methods). We found no significant differences between identical and near homophone puns in terms of funniness ratings ($t(130.91) = 0.13, p = 0.896$), ambiguity scores ($t(137.80) = 1.13, p = 0.261$), and distinctiveness scores ($t(134.91) = -0.61, p = 0.543$), suggesting that ambiguity and distinctiveness are fairly robust to the differences between puns that involve identical or near homophone words.

As a result, we collapsed across identical and near homophone sentences for the remaining analyses. We found that ambiguity was significantly higher for pun sentences than non-pun sentences ($t(159.48) = 7.89, p < 0.0001$), which suggests that the ambiguity measure successfully captures characteristics distinguishing puns from other phonetically ambiguous sentences. Distinctiveness was also significantly higher for pun sentences than non-pun sentences ($t(248.99) = 6.11, p < 0.0001$). Figure 5.2 shows the standard error ellipses for the two sentence types in a two-dimensional space of ambiguity and distinctiveness. Although there is a fair amount of noise in the predictors (likely due to simplifying assumptions, the need to use empirical measures of relatedness, and the inherent complexity of humor), pun sentences (both identical and near homophone) tend to cluster at a space with higher ambiguity and distinctiveness, while non-pun sentences score lower on both measures.

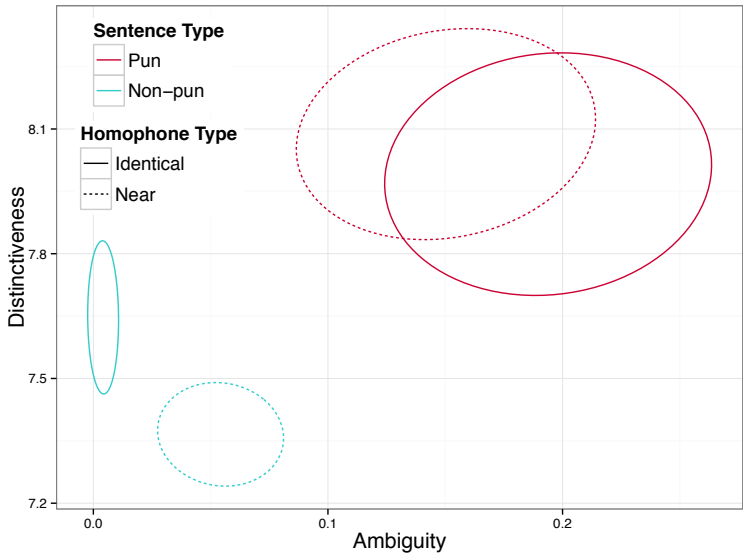

Figure 5.2: Standard error ellipses of ambiguity and distinctiveness for each sentence type. Puns (both identical and near homophone) score higher on ambiguity and distinctiveness; non-pun sentences score lower.

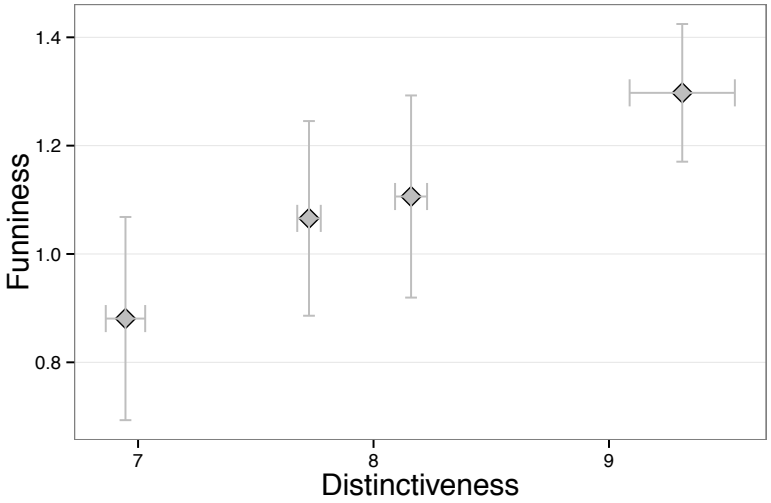

Figure 5.3: Average funniness ratings and distinctiveness of 145 pun sentences binned according to distinctiveness quartiles. Error bars are confidence intervals.

m_a	m_b	Type	Sentence	Amb.	Dist.	Funni.
hare	*hair*	Pun	The **magician** got so mad he *pulled* his **hare** out.	0.15	7.87	1.71
		Non	The **hare ran rapidly** through the **fields**.	1.43E^{-5}	7.25	-0.40
tooth	*truth*	Pun	A **dentist** has to *tell* a **patient** the *whole* **tooth**.	0.1	8.48	1.41
		Non	A **dentist *examines*** one **tooth** at a time.	8.92E^{-5}	7.65	-0.45

Table 5.6: Semantically relevant words, ambiguity/distinctiveness scores, and funniness ratings for sentences from each category. Words in boldface are semantically relevant to m_a; words in italics are semantically relevant to m_b.

We constructed a linear mixed-effects model of funniness judgments with ambiguity and distinctiveness as fixed effects, a by-item random intercept, and by-subject random slopes for entropy and distinctiveness. We found that ambiguity and distinctiveness were both highly significant predictors, with funniness increasing as each of ambiguity and distinctiveness increases (Table 5.4). Furthermore, the two measures capture a substantial amount of the reliable variance in funniness ratings averaged across subjects ($F(2, 432) = 74.07, R^2 = 0.25, p < 0.0001$). A linear mixed effects model including a term for the interaction between ambiguity and distinctiveness (both as fixed effect and by-subjects random slope) showed no significant interaction between the two ($t = 1.39, p > 0.15$).

We then examined whether the measures are able to go beyond distinguishing puns from non-puns to predicting fine-grained levels of funniness within puns. We found that ambiguity does not correlate with human ratings of funniness within the 145 pun sentences ($r = 0.03, p = 0.697$). However, distinctiveness ratings correlate significantly with human ratings of funniness within pun sentences ($r = 0.28, p < 0.001$). By separating the puns into four equal bins based on their distinctiveness, we found that puns with distinctiveness measures in the top-most quartile were significantly funnier than puns with distinctiveness measures in the lower quartiles ($t(90.15) = 3.41, p < 0.001$) (Figure 5.3). This suggests that while ambiguity helps distinguish puns from non-puns, high distinctiveness characterizes exceptionally humorous puns. To our knowledge, our model provides the first quantitative measure that predicts fine-grained levels of funniness within humorous stimuli.

In addition to predicting the funniness of a sentence, the model can also be used to reveal critical features of each pun that make it amusing. For each sentence, we identified the set of words that is most likely to be semantically relevant given $\vec{w}$ and each sentence meaning m. Formally, we computed $P(\vec{f}|m_a, \vec{w})$ and $P(\vec{f}|m_b, \vec{w})$. Table 5.6 shows a group of identical-homophone sentences and a group of near-homophone sentences. Sentences in each group contain the same pair of candidate meanings

for the homophone; however, they differ on ambiguity, distinctiveness, and funniness. Words that are most likely to be relevant given sentence meaning m_a are in boldface; words that are most likely to be relevant given m_b are in italics. Qualitatively, we observe that the two pun sentences (which are significantly funnier) have more distinct and balanced sets of meaningful words for each sentence meaning than other sentences in their groups. Non-pun sentences tend to have no words in support of the meaning that was not observed. Furthermore, the boldfaced and italicized words in each pun sentence are what one might intuitively use to explain why the sentence is funny—for example, the fact that magicians tend to perform magic tricks with hares, and people tend to be described as pulling out their hair when angry.

5.4 Discussion

In this chapter, we presented a simple model of gist-level sentence processing and used it to derive formal measures that predict human judgments of humor in puns. We showed that a noisy-channel model of sentence processing facilitates flexible context selection, which enables a single series of words to express multiple meanings. Our work is one of the first to integrate a computational model of sentence processing to analyze humor in a manner that is both intuitive and quantitative. In addition, it is the first computational work to our knowledge to go beyond classifying humorous versus regular sentences to predict fine-grained funniness judgments within humorous stimuli.

The idea of deriving measures of humor from a model of general language understanding is closely related to previous approaches, where humor is analyzed within a framework of semantic theory and language comprehension. Raskin (2012) Semantic Script Theory of Humor (SSTH) builds upon a theory of language comprehension in which language is understood in terms of scripts. Under this analysis, a text is funny when it activates two scripts that are incompatible with each other. This theory explains a number of classic jokes where the punch line introduces a script that is incongruous with the script activated by the joke's setup. Attardo and Raskin (1991) proposed a revision to SBST in the General Theory of Verbal Humor (GTVH), which details six hierarchically organized knowledge resources that inform the understanding of texts as well as the detection of humor. Nirenburg and Raskin (2004) further formalized the ideas proposed in SBST and GTVH by developing a system for computational semantics termed Ontological Semantics, which includes a large concept ontology, a repository of facts, and an analyzer that translates texts into an ontology-based knowledge

representation. This system provides rich ontological knowledge to support in-depth language comprehension and has been applied productively to a variety of domains (Nirenburg & Raskin, 2004; Beale, Lavoie, McShane, Nirenburg, & Korelsky, 2004; J. M. Taylor, Raskin, & Hempelmann, 2011). C. Hempelmann, Raskin, and Triezenberg (2006) used a classic joke to show that an extension to the Ontological Semantics system can in principle detect as well as generate humorous texts. However, to our knowledge the system has not yet been tested on a larger body of texts to demonstrate its performance in a quantitative manner (Raskin, 2008; J. M. Taylor, 2010). While providing detailed analyses that reveal many important characteristics of humor, much of the work on formalizing humor theories falls short of predicting people's fine-grained judgments of funniness for a large number of texts (Raskin & Attardo, 1994; Ritchie, 2001, 2001; Attardo, Hempelmann, & Di Maio, 2002; C. F. Hempelmann, 2004; Veale, 2006). In this regard, we believe that our work advances the current state of formal approaches to humor theory. By implementing a simple but psychologically motivated computational model of sentence processing, we derived measures that distinguish puns from regular sentences and correlate significantly with fine-grained humor ratings within puns. Our approach also provides an intuitive but automatic way to identify features that make a pun funny. This suggests that a probabilistic model of general sentence processing (even without the support of rich ontological semantics) may enable powerful explanatory measures of humor.

In addition to advancing computational approaches, our work contributes to cognitive theories of humor by providing evidence that different factors may account for separate aspects of humor appreciation. Some humor theorists argue that while incongruity is necessary for humor, resolving incongruity—discovering a cognitive rule that explains the incongruity in a logical manner—is also key (Ritchie, 1999, 2009; Suls, 1972). We can construe our measures as corresponding roughly to incongruity and resolution in this sense, where ambiguity represents the presence of incongruous sentence meanings, and distinctiveness represents the degree to which each meaning is strongly supported by different parts of the stimulus. Our results would then suggest that incongruity distinguishes humorous input from regular sentences, while the intensity of humor may depend on the degree to which incongruity is resolved by focusing on two different supporting sets. Future work could more specifically examine the relationship between incongruity resolution and the measures presented in our framework.

Although our task here was limited in scope, it is a step towards developing computational models

that explain higher-order linguistic phenomena such as humor. To address more complex jokes, future work may incorporate more sophisticated models of language understanding to consider the time course of sentence processing (Kamide, Altmann, & Haywood, 2003; McRae, Spivey-Knowlton, & Tanenhaus, 1998), effects of pragmatic reasoning and background knowledge (J. T. Kao, Bergen, & Goodman, 2014; J. T. Kao, Wu, et al., 2014; J. Kao & Goodman, 2015), and multi-sentence discourse (Polanyi, 1988; Chambers & Jurafsky, 2008). Our approach could also benefit greatly from the rich commonsense knowledge encoded in the Ontological Semantics system and may be combined with it to measure ambiguity and distinctiveness at the script level rather than at the level of the sentence.

Previous research on creative language use such as metaphor, idioms, and irony has contributed a great deal to our understanding of the cognitive mechanisms that enable people to infer rich meanings from sparse and often ambiguous linguistic input (Lakoff & Turner, 2009; Nunberg, Sag, & Wasow, 1994; Gibbs & O'Brien, 1991). We hope that our work on humor contributes to theories of language understanding to account for a wider range of linguistic behaviors and the social and affective functions they serve. By deriving the precise properties of sentences that make us laugh, our work brings us one step closer to understanding that funny thing called humor (pun intended).

Chapter 6

Conclusions

This thesis set out to answer two questions: how do people understand creative uses of language, and how do these types of uses produce the range of social and emotional effects that we experience in everyday communication? To answer these questions, we took a modeling approach that builds upon general theories of language comprehension and produces graded, quantitative results that capture the subtleties of human interpretation. In Chapter 2, we introduced the idea of communicative goals to the RSA framework and formalized a notion of the relevance principle. We showed that basic principles of communication, combined with background knowledge and social reasoning, allows the model to appropriately interpret hyperbolic utterances as well as their associated affects. In Chapter 3, we further extended the space of communicative goals to include a richer representation of emotions as well as contextually determined topics. We showed that this extension allowed RSA models to interpret a range of figurative uses such as verbal irony and metaphor, again using the same basic principles of communication. In Chapter 4, we relaxed the assumption that speaker and listener share the same background knowledge, and allowed the listener to reason about common ground given various utterances. We showed that the model infers a higher probability of common ground given figurative utterances than literal ones, suggesting that figurative language may strengthen social bonds via inferences about common ground. Finally, in Chapter 5, we used a simple model of sentence comprehension to derive quantitative measures of humor that are motivated by both general theories of language comprehension under noise and the incongruity theory of humor. With the Rational Speech-acts framework, we were able to introduce increasingly rich and psychologically

realistic components of communication, and to show how they are responsible for different effects. Taken together, this work helps shed light on scientific theories of creative language use and advances formal approaches to language, such that computational models can explain a broader range of phenomena that enrich our linguistic and social lives.

6.1 Limitations and future directions

"The past is always tense, the future perfect." — Zadie Smith

The models we described capture rich aspects of language use that were previously beyond the reach of quantitative study. However, it is only the beginning of a long road ahead. Each chapter opens up new directions for future research. While Chapter 2 focused on the nonliteral interpretation of number words, many hyperbolic uses involve non-numeric utterances, such as "Andrew *never* does the dishes," or "Bob is the best cook in the world." Because the semantics of these types of utterances are more complex than number words, there is more opportunity to explore how semantic and pragmatic information interact to produce various types of interpretations. Some preliminary work on universal quantifiers (e.g. "Cam ate *all* of the cookies") shows how different contexts may shift the interpretation from literal (Cam ate literally all of the cookies), to imprecise (Cam ate almost all of the cookies), to domain restricted (Cam ate all of the *chocolate chip* cookies) to hyperbolic (Cam ate way too many cookies, and I am upset!) (J. T. Kao, Degen, & Goodman, 2015). In Chapter 3, we showed how ironic interpretations could arise through reasoning about speakers' emotional valence and arousal. However, our model did not demonstrate a classic asymmetry effect in verbal irony, where ironic insults (e.g. "What a wonderful idea") are more common and rated as more appropriate than ironic compliments (e.g. "What a terrible idea"). This asymmetry is not explained by inferences about speakers' emotional goals alone, and may require additional assumptions in the model to capture. And while the extended RSA model produces appropriate interpretations of metaphor, so far it has only been tested on simple nominal metaphors with a limited set of features, and may require significant extensions to account for the more complex metaphors that we see in literature and poetry. Furthermore, more work is needed to test the boundaries of the RSA framework when applied to figurative language, for example whether it can predict other types of figurative use such as understatement and metonymy where the relevant goals and questions under discussion may be less clear.

Regarding the social implications of figurative language discussed in Chapter 4, our model currently does not account for individual or cultural differences in norms of nonliteral use. Some individuals may be more or less prone to speak sarcastically, and some cultures may be more or less tolerant of nonliteral communication. These types of high-level knowledge of personality traits and social norms may also play into people's inferences regarding nonliteral uses of language and are useful to incorporate in our models (A. N. Katz, Blasko, & Kazmerski, 2004). Finally, in Chapter 5, we used puns as a case study for formalizing the concept of incongruity and deriving quantitative measures of humor. Through preliminary analysis, we find that the same measures of incongruity—ambiguity and distinctiveness—can also be used to quantify the humor in figurative uses such as hyperbole and irony. Future work could examine this generalization in more detail, with the goal of predicting humor in a broader class of creative language.

Beyond the need for extensions to help cover a wider set of creative language use, a deeper question regarding this work is: how "creative" are these uses of language, really? Many of the examples of hyperbole, verbal irony, and metaphor we examined in this thesis occur in everyday language in fairly predictable and conventionalized ways. While adding interest and color to communication, they are not the most ingenious demonstrations of linguistic creativity. Could the models described here also interpret and appreciate the "true" creativity that we see in the works of Shakespeare, Emily Dickinson, and Pablo Neruda?

Fully understanding figurative uses such as "Hope is the thing with feathers" and "I want to do with you what spring does with the cherry trees" requires an amount of world knowledge that is not yet fully amenable to quantification. Many of the most evocative metaphors communicate information along many dimensions at once, and it may be challenging to identify all of the features of the source and target domains that are relevant for producing figurative interpretations and well as associated poetic effects*. Overall, we believe that the models we described take us one step closer towards understanding how people interpret creative and figurative language. However, it remains to be seen whether the principles of communication that we use to understand simple metaphors such as "John is a shark" are sufficient for poetic metaphors as well. It is possible that in order to truly understand and appreciate the best examples of linguistic creativity, more research is needed to identify predictors of what makes a metaphor "good." For example, is the aptness of

*For example, what the poet wants to do with the addressee is metaphorically similar to what spring does with the cherry trees, but it is difficult to put into words precisely what the action could be, and in what dimensions are the two actions similar (except, perhaps, that they both change the receiver of the action in positive ways).

a metaphor predicted by the number of dimensions of meaning it can communicate? Could aptness and aesthetic value be related to other types of measures, such as ambiguity (existence of multiple interpretations) and distinctiveness (difference in contextual support for one interpretation versus the other)? How do features such as concreteness and imageability factor into measures of aesthetic quality, or phonological features such as rhyme and alliteration? While we have done some work in this direction identifying the linguistic features of poetic value (J. Kao & Jurafsky, 2012; J. T. Kao & Jurafsky, 2015), these questions are currently beyond the reach of the models described in this thesis. In future work, an approach that combines figurative interpretation with features of aesthetic value could be the most promising for understanding poetic language in a quantitative and psychologically motivated manner.

Finally, the models presented in Chapters 2 and 3 were tested on specific domains (e.g. prices, weather, animals), where we elicited people's background knowledge through behavioral experiments. While this approach directly taps into people's knowledge and is relatively precise, it is difficult to scale up these elicitation experiments to cover a wide range of domains. By using data mining and machine learning methods (e.g. scraping websites for distributions of prices, or automatically learning features of animals from online encyclopedias), we may be able to automate the knowledge elicitation process. This would allow us to test the models on a larger set of utterances and domains, including naturally occurring examples that we could mine from social media corpora. Ultimately, a combination of methods—targeted Bayesian cognitive models, behavioral experiments, and large-scale corpus analyses—could help us make the most progress towards understanding the many aspects of creative language use.

6.2 Final remarks

Modeling psychological phenomena is often a process of peeling away layers of skin to take apart and reconstruct the skeleton underneath. In this thesis, we dissected several cases of creative language use and discovered that it is held up by the same structure that supports many other aspects of cognition, such as reasoning about other people and reasoning under uncertainty more generally. By discovering this common structure, we showed how people harness basic principles of language comprehension to derive socially meaningful information from a range of creative use, suggesting that our ability to understand and appreciate creative uses of language arises organically from our

general capacity to understand language and to simulate other minds. In addition, we identified and formalized new structures that may be critical for standard language understanding, such as reasoning about speakers' communicative goals, emotional attitudes, and uncertainty over shared background knowledge.

One of the goals of this thesis was to use a novel computational framework to formalize ideas established by great thinkers in the field, and in the process discover new insights that add to the scientific discourse. I hope that we have shown how this framework could be useful for establishing the connection between rational, information-processing accounts of language and the creative uses that help us express, make us laugh, and bring us closer together.

References

Ariel, M. (2002). The demise of a unique concept of literal meaning. *Journal of Pragmatics*, *34*(4), 361–402.

Aristotle. (1911). Poetics. *The Basic Works of Aristotle, ed. Richard McKeon (New York: Random House, 1941)*, *10*.

Attardo, S. (1994). *Linguistic theories of humor* (Vol. 1). Walter de Gruyter.

Attardo, S., Hempelmann, C. F., & Di Maio, S. (2002). Script oppositions and logical mechanisms: Modeling incongruities and their resolutions. *Humor*, *15*(1), 3–46.

Attardo, S., & Raskin, V. (1991). Script theory revis (it) ed: Joke similarity and joke representation model. *Humor-International Journal of Humor Research*, *4*(3-4), 293–348.

Bach, K. (1994). Semantic slack: What is said and more. *Foundations of Speech Act theory: Philosophical and Linguistic Perspectives*, 267–291.

Bamman, D., & Smith, N. A. (2015). Contextualized sarcasm detection on twitter. In *Proceedings of the 9th international conference on web and social media* (pp. 574–77).

Bastiaanse, H. (2009). The rationality of round interpretation. In *Vagueness in communication* (pp. 37–50). Springer.

Beale, S., Lavoie, B., McShane, M., Nirenburg, S., & Korelsky, T. (2004). Question answering using ontological semantics. In *Proceedings of the 2nd workshop on text meaning and interpretation* (pp. 41–48).

Bergen, L., & Goodman, N. D. (2015). The strategic use of noise in pragmatic reasoning. *Topics in Cognitive Science*, *7*(2), 336–350.

Bergen, L., Goodman, N. D., & Levy, R. (2012). That's what she (could have) said: How alternative utterances affect language use. In *Proceedings of the Thirty-Fourth Annual Conference of the*

Cognitive Science Society.

Bergson, H. (1914). *Laughter: An essay on the meaning of the comic*. Macmillan.

Berry, D. M. (2012). *Understanding digital humanities*. Palgrave Macmillan.

Binsted, K. (1996). Machine humour: An implemented model of puns.

Boden, M. A. (2009). Computer models of creativity. *AI Magazine*, *30*(3), 23.

Brown, P. (1995). Politeness strategies and the attribution of intentions: the case of tzeltal irony. *Social Intelligence and Interaction, Cambridge University Press, Cambridge*, 153–174.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

Carter, R. (1999). Common language: corpus, creativity and cognition. *Language and Literature*, *8*(3), 195–216.

Carter, R., & McCarthy, M. (2004). Talking, creating: interactional language, creativity, and context. *Applied Linguistics*, *25*(1), 62–88.

Caucci, G. M., & Kreuz, R. J. (2012). Social and paralinguistic cues to sarcasm. *Humor*, *25*(1), 1–22.

Cazden, C. B. (1974). Play and metalinguistic awareness: One dimension of language experience. *The Urban Review*, *7*(1), 28–39.

Chambers, N., & Jurafsky, D. (2008). Unsupervised learning of narrative event chains. In *Acl* (Vol. 94305, pp. 789–797).

Clark, H. H. (1991). Words, the world, and their possibilities. *The Perception of Structure*, 263–278.

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

Clark, H. H. (1997). Dogmas of understanding. *Discourse Processes*, *23*(3), 567–598.

Clark, H. H., & Gerrig, R. J. (1984). On the pretense theory of irony. *Journal of Experimental Psychology: General*.

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. *Advances in Psychology*, *9*, 287–299.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809.

Cohen, P. R., & Levesque, H. J. (1985). Speech acts and rationality. In *Proceedings of the 23rd annual meeting on association for computational linguistics* (pp. 49–60).

Colston, H. L. (1997). Salting a wound or sugaring a pill: The pragmatic functions of ironic criticism. *Discourse Processes*, *23*(1), 25–45.

Colston, H. L., & Keller, S. B. (1998). You'll never believe this: Irony and hyperbole in expressing surprise. *Journal of Psycholinguistic Research*, *27*(4), 499–513.

Cook, G. (1996). Language play 1n engl1sh. *Using English from Conversation to Canon*, *2*, 198.

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

Dascal, M. (1981). Contextualism. *Possibilities and Limitations of Pragmatics*.

Degen, J., & Tanenhaus, M. K. (2015). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive science*.

Der Henst, V., Carles, L., Sperber, D., et al. (2002). Truthfulness and relevance in telling the time. *Mind & Language*, *17*(5), 457–466.

Dews, S., Kaplan, J., & Winner, E. (1995). Why not say it directly? the social functions of irony. *Discourse processes*, *19*(3), 347–367.

Dews, S., & Winner, E. (1999). Obligatory processing of literal and nonliteral meanings in verbal irony. *Journal of pragmatics*, *31*(12), 1579–1599.

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

Ely, R., & McCabe, A. (1994). The language play of kindergarten children. *First Language*, *14*(40), 019–35.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological review*, *116*(4), 752.

Fogelin, R. J. (2011). *Figuratively speaking: Revised edition*. Oxford University Press.

Forabosco, G. (1992). Cognitive aspects of the humor process: The concept of incongruity. *Humor-International Journal of Humor Research*, *5*(1-2), 45–68.

Fowler, H. W., et al. (1926). dictionary of modern english usage.

Frank, M. (2013). Throwing out the bayesian baby with the optimal bathwater: Response to. *Cognition*, *128*(3), 417–423.

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

Friedrich, P. (1979). Poetic language and the imagination: A reformulation of the sapir hypothesis. *Language, context, and the imagination*, 441–512.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Ijcai* (Vol. 7, pp. 1606–1611).

García-Carpintero, M. (2001). Gricean rational reconstructions and the semantics/pragmatics distinction. *Synthese*, *128*(1-2), 93–131.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170.

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

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

Gibbs, R. W. (1986). On the psycholinguistics of sarcasm. *Journal of Experimental Psychology: General*, *115*(1), 3.

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

Gibbs, R. W. (1994). *The poetics of mind.* Cambridge: Cambridge University Press.

Gibbs, R. W. (2000). Irony in talk among friends. *Metaphor and Symbol*, *15*(1-2), 5–27.

Gibbs, R. W., & Colston, H. (1999). Figurative language. *The MIT Encyclopedia of the Cognitive Sciences*, 314–315.

Gibbs, R. W., & Colston, H. L. (2012). *Interpreting figurative meaning.* Cambridge University Press.

Gibbs, R. W., & O'Brien, J. (1991). Psychological aspects of irony understanding. *Journal of pragmatics*, *16*(6), 523–530.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.

Giesen, H. (2000). Non-photorealistic rendering.

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. (2002). Literal vs. figurative language: Different or equal? *Journal of Pragmatics*, *34*(4),

487–506.

Giora, R. (2003). *On our mind: Salience, context, and figurative language.* Oxford University Press New York.

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.

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.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Grefenstette, E., Sadrzadeh, M., Clark, S., Coecke, B., & Pulman, S. (2014). Concrete sentence spaces for compositional distributional models of meaning. In *Computing meaning* (pp. 71–86). Springer.

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

Hall, G. (2001). The poetics of everyday language. *CAUCE, Revista de Filología y su Didáctica*(24), 69–86.

Halliday, M. A. (1971). Linguistic function and literary style: an inquiry into the language of william golding?s the inheritors. In *Literary style: A symposium* (pp. 330–368).

Hempelmann, C., Raskin, V., & Triezenberg, K. E. (2006). Computer, tell me a joke... but please make it funny: Computational humor with ontological semantics. In *Flairs conference* (Vol. 13, pp. 746–751).

Hempelmann, C. F. (2004). Script opposition and logical mechanism in punning. *Humor-International Journal of Humor Research*, *17*(4), 381–392.

Holtgraves, T. M. (2013). *Language as social action: Social psychology and language use.* Psychology Press.

Honeck, R. P. (1986). Verbal materials in research on figurative language. *Metaphor and Symbol*,

1(1), 25–41.

Hörmann, H. (1983). *Was tun die wörter miteinander im satz?, oder, wieviele sind einige, mehrere und ein paar?* Verlag für Psychologie.

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

Horton, W. S. (2007). Metaphor and readers? attributions of intimacy. *Memory & cognition*, *35*(1), 87–94.

Hurley, M. M., Dennett, D. C., & Adams, R. B. (2011). *Inside jokes: Using humor to reverse-engineer the mind*. MIT press.

Jäger, G., & Ebert, C. (2009). Pragmatic rationalizability. In *Proceedings of sinn und bedeutung* (Vol. 13, pp. 1–15).

Jakobson, R. (1960). Closing statement: Linguistics and poetics. *Style in language*, *350*, 377.

Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Joos, M. (1950). Description of language design. *The Journal of the Acoustical Society of America*, *22*(6), 701–707.

Jorgensen, J. (1996). The functions of sarcastic irony in speech. *Journal of Pragmatics*, *26*(5), 613–634.

Jorgensen, J., Miller, G. A., & Sperber, D. (1984). Test of the mention theory of irony. *Journal of Experimental Psychology: General*, *113*(1), 112.

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and language*, *49*(1), 133–156.

Kao, J., & Goodman, N. (2015). Let's talk (ironically) about the weather: Modeling verbal irony. In *Proceedings of the 37th annual conference of the cognitive science society.*

Kao, J., & Jurafsky, D. (2012). A computational analysis of style, affect, and imagery in contemporary poetry. In *Naacl workshop on computational linguistics for literature* (pp. 8–17).

Kao, J. T., Bergen, L., & Goodman, N. D. (2014). Formalizing the pragmatics of metaphor understanding. In *Proceedings of the 36th annual meeting of the cognitive science society.*

Kao, J. T., Degen, J., & Goodman, N. D. (2015). When "all" means not all: Nonliteral interpretations of universal quantifiers. In *Xprag conference.*

Kao, J. T., & Jurafsky, D. (2015). A computational analysis of poetic style. *LiLT (Linguistic Issues in Language Technology)*, *12*.

Kao, J. T., Levy, R., & Goodman, N. D. (2013). The funny thing about incongruity: A computational model of humor in puns. In *Proceedings of the 35th annual conference of the cognitive science society* (Vol. 728733).

Kao, J. T., Levy, R., & Goodman, N. D. (2015). A computational model of linguistic humor in puns. *Cognitive science*.

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.

Katz, A. N., Blasko, D. G., & Kazmerski, V. A. (2004). Saying what you don't mean social influences on sarcastic language processing. *Current Directions in Psychological Science*, *13*(5), 186–189.

Katz, A. N., & Ferretti, T. R. (2001). Moment-by-moment reading of proverbs in literal and nonliteral contexts. *Metaphor and Symbol*, *16*(3-4), 193–221.

Katz, J. J. (1981). Literal meaning and logical theory. *The Journal of Philosophy*, 203–233.

Kiddon, C., & Brun, Y. (2011). That's what she said: double entendre identification. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies: short papers-volume 2* (pp. 89–94).

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148.

Kreuz, R. J. (1996). The use of verbal irony: Cues and constraints. *Metaphor: Implications and Applications*, 23–38.

Kreuz, R. J., Kassler, M. A., Coppenrath, L., & Allen, B. M. (1999). Tag questions and common ground effects in the perception of verbal irony. *Journal of Pragmatics*, *31*(12), 1685–1700.

Kreuz, R. J., & Link, K. E. (2002). Asymmetries in the use of verbal irony. *Journal of Language and Social Psychology*, *21*(2), 127–143.

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

Krifka, M. (2007). Approximate interpretation of number words.

Kruger, A. (1996). The nature of humor in human nature: Cross-cultural commonalities. *Counselling Psychology Quarterly*, *9*(3), 235–241.

Lakoff, G. (1986). The meanings of literal. *Metaphor and Symbol*, *1*(4), 291–296.

Lakoff, G. (1993). The contemporary theory of metaphor.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

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

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

Lanham, R. A. (1991). *A handlist of rhetorical terms*. University of California Press.

Lasersohn, P. (1999). Pragmatic halos. *Language*, 522–551.

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).

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... others (2009). Life in the network: the coming age of computational social science. *Science (New York, NY)*, *323*(5915), 721.

Leggitt, J. S., & Gibbs, R. W. (2000). Emotional reactions to verbal irony. *Discourse Processes*, *29*(1), 1–24.

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

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 234–243).

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *106*(50), 21086–21090.

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).

Liang, P., Jordan, M. I., & Klein, D. (2013). Learning dependency-based compositional semantics. *Computational Linguistics*, *39*(2), 389–446.

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

Lundy, D. E., Tan, J., & Cunningham, M. R. (1998). Heterosexual romantic preferences: The importance of humor and physical attractiveness for different types of relationships. *Personal Relationships*, *5*(3), 311–325.

MacGhee, P. E., & Pistolesi, E. (1979). *Humor: Its origin and development*. WH Freeman and Company.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. *Debates in the digital humanities*, *2*, 460–475.

Martin, R. A. (2010). *The psychology of humor: An integrative approach*. Academic press.

Martin, R. A., Kuiper, N. A., Olinger, L., & Dance, K. A. (1993). Humor, coping with stress, self-concept, and psychological well-being. *Humor: International Journal of Humor Research*.

Maybin, J., & Swann, J. (2007). Everyday creativity in language: Textuality, contextuality, and critique. *Applied Linguistics*, *28*(4), 497–517.

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

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283–312.

Mihalcea, R., & Strapparava, C. (2006). Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, *22*(2), 126–142.

Mihalcea, R., Strapparava, C., & Pulman, S. (2010). Computational models for incongruity detection in humour. In *Computational linguistics and intelligent text processing* (pp. 364–374). Springer.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Mumford, M. D. (2003). Where have we been, where are we going? taking stock in creativity research. *Creativity Research Journal*, *15*(2-3), 107–120.

Nirenburg, S., & Raskin, V. (2004). *Ontological semantics*. MIT Press.

Niznikiewicz, M., & Squires, N. K. (1996). Phonological processing and the role of strategy in silent

reading: behavioral and electrophysiological evidence. *Brain and language*, *52*(2), 342–364.

Norrick, N. R. (1982). On the semantics of overstatement. *Akten des*, *16*, 168–176.

Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. *Language*, 491–538.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in cognitive sciences*, *5*(8), 349–357.

Ortony, A. (1979). Beyond literal similarity. *Psychological Review*, *86*(3), 161.

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.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Pexman, P. M., Lupker, S. J., & Jared, D. (2001). Homophone effects in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 139.

Pexman, P. M., & Zvaigzne, M. T. (2004). Does irony go better with friends? *Metaphor and Symbol*, *19*(2), 143–163.

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

Pinker, S. (2007). The evolutionary social psychology of off-record indirect speech acts. *Intercultural pragmatics*, *4*(4), 437–461.

Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of pragmatics*, *12*(5), 601–638.

Pollatsek, A., Lesch, M., Morris, R. K., & Rayner, K. (1992). Phonological codes are used in integrating information across saccades in word identification and reading. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(1), 148.

Potts, C. (2007). The expressive dimension. *Theoretical Linguistics*, *33*(2), 165–197.

Raskin, V. (2008). *The primer of humor research* (Vol. 8). Walter de Gruyter.

Raskin, V. (2012). *Semantic mechanisms of humor* (Vol. 24). Springer Science & Business Media.

Raskin, V., & Attardo, S. (1994). Non-literalness and non-bona-fide in language: An approach to formal and computational treatments of humor. *Pragmatics & Cognition*, *2*(1), 31–69.

Reyes, A., Rosso, P., & Buscaldi, D. (2012). From humor recognition to irony detection: The

figurative language of social media. *Data & Knowledge Engineering*, *74*, 1–12.

Ricoeur, P. (1973). Creativity in language. *Philosophy Today*, *17*(2), 97–111.

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).

Ritchie, G. (1999). Developing the incongruity-resolution theory.

Ritchie, G. (2001). Current directions in computational humour. *Artificial Intelligence Review*, *16*(2), 119–135.

Ritchie, G. (2009). Variants of incongruity resolution. *Journal of Literary Theory*, *3*(2), 313–332.

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

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

Russell, J. (1980). A circumplex of affect. *Journal of Personality and Social Psychology*, *36*, 1152–1168.

Searle, J. (1979). Metaphor. In A. Ortony (Ed.), *Metaphor and Thought.* Cambridge University Press.

Searle, J. R. (1978). Literal meaning. *Erkenntnis*, *13*(1), 207–224.

Smith, J. (1969). *Mystery of rhetoric unveiled, 1657.* Scolar P.

Smith, W. J., Vernard Harrington, K., & Neck, C. P. (2000). Resolving conflict with humor in a diversity context. *Journal of Managerial Psychology*, *15*(6), 606–625.

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 1201–1211).

Sperber, D., & Wilson, D. (1981). Irony and the use-mention distinction. *Radical pragmatics*, *49*.

Sperber, D., & Wilson, D. (1985). Loose talk. In *Proceedings of the Aristotelian Society* (pp. 153–171).

Sperber, D., & Wilson, D. (2008). A deflationary account of metaphors. *The Cambridge Handbook of Metaphor and Thought*, 84–105.

Sperber, D., Wilson, Z. H., Deirdre, & Ran, Y. (1986). *Relevance: Communication and cognition*. Harvard University Press Cambridge, MA.

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

Suls, J. M. (1972). A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. *The psychology of humor: Theoretical perspectives and empirical issues*, *1*, 81–100.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.

Tannen, D. (2007). *Talking voices: Repetition, dialogue, and imagery in conversational discourse* (Vol. 26). Cambridge University Press.

Taylor, J., & Mazlack, L. (2004). Computationally recognizing wordplay in jokes. In *Proceedings of cogsci* (Vol. 2004).

Taylor, J. M. (2010). Ontology-based view of natural language meaning: the case of humor detection. *Journal of Ambient Intelligence and Humanized Computing*, *1*(3), 221–234.

Taylor, J. M., Raskin, V., & Hempelmann, C. F. (2011). From disambiguation failures to common-sense knowledge acquisition: A day in the life of an ontological semantic system. In *Proceedings of the 2011 ieee/wic/acm international conferences on web intelligence and intelligent agent technology-volume 01* (pp. 186–190).

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.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279–1285.

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

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. (1982). Understanding and appreciating metaphors. *Cognition*, *11*(3), 203–244.

Vaid, J., & Ramachandran, V. S. (2001). Laughter and humor. *Oxford companion to the body*, 426–427.

Veale, T. (2006). Computability as a test on linguistics theories. *Applications of Cognitive Linguistics*, *1*, 461.

Veale, T. (2012). *Exploding the creativity myth: The computational foundations of linguistic creativity*. A&C Black.

Wallach, H. (2016). *Computational social science*. Cambridge University Press.

Wilson, D., & Carston, R. (2006). Metaphor, relevance and the "emergent property" issue. *Mind & Language*, *21*(3), 404–433.

Yuille, A., & Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, *10*(7), 301–308.

Zhang, Z., Gentile, A. L., & Ciravegna, F. (2011). Harnessing different knowledge sources to measure semantic relatedness under a uniform model. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 991–1002).