Formalizing the Pragmatics of Figurative Language

Justine T. Kao

Stanford University

Leon Bergen

Massachusetts Institute of Technology

Noah D. Goodman

Stanford University

Author Note

Justine T. Kao, Department of Psychology, Stanford University.

Leon Bergen, Department of Brain and Cognitive Sciences, MIT.

Noah D. Goodman, Department of Psychology, Stanford University.

Correspondence concerning this article should be addressed to Justine T. Kao, Department of Psychology, Stanford University, 450 Serra Mall, Stanford, CA 94025. E-mail: justinek@stanford.edu

Abstract

People often use language to convey information that goes far beyond an utterance's literal meaning. In particular, figurative language such as hyperbole, irony, and metaphor showcase people's ability to use background knowledge and social reasoning to interpret utterances in context. While figurative utterances are often literally false, people are highly adept at inferring relevant and true information from these utterances. What is the cognitive and linguistic basis of figurative language understanding? Here we describe a computational model that formalizes figurative communication as recursive social reasoning between speaker and listener. Our model extends and overcomes limitations of basic Rational Speech Act (RSA) models, a family of computational models that capture many phenomena in human pragmatic reasoning but are restricted to the interpretation of literally true utterances. We show that an RSA model extended to accommodate inferences about the topic of conversation is able to predict people's interpretations of hyperbole, irony, and metaphor. We argue that despite apparent differences among these subtypes of figurative language, the same computational framework flexibly produces fine-grained interpretations for a range of figurative uses. 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.

Keywords: pragmatics, figurative language, computational modeling

Formalizing the Pragmatics of Figurative Language

The ability to understand figurative language is necessary in a world where people do not always mean what they say. We implicate, exaggerate, make 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 and interest to linguistic behavior (Glucksberg, 2001; Pilkington, 2000; Lakoff & Turner, 2009; R. 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, 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?

An ocean of ink has been spilled across many disciplines to answer this question, 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, researchers 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 a different type of figurative language such as verbal irony, researchers then posit separate

mechanisms such as pretense (Clark & Gerrig, 1984) or echoic mention (Sperber & Wilson, 1981) to address the specific factors that shape ironic interpretations. While these approaches significantly advance our understanding of the conceptual factors that guide metaphor and irony understanding, they tend to treat each type of figurative language as requiring a separate mechanism distinct from or in addition to standard language understanding. By relying on a more specialized mechanism for figurative language, these approaches leaves 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 these particular 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 views in the pragmatics literature have taken this view to examine figurative language: the standard pragmatic view and the relevance theoretic view. 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 naturally within a more general theory of pragmatics, it has met with several criticisms. One of them 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 Jr, 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. Another 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 Jr, 1992). These empirical findings suggest that literal meanings do not have to be explicitly computed and then rejected before appropriate figurative interpretations emerge.

A second view on the pragmatics of figurative language stems from relevance theory, a general theory of communication whose central claim is that human cognition is governed by a tendency to maximize relevance (Sperber, Wilson, He, & Ran, 1986). Sperber and Wilson (2008) introduced the communicative principle of relevance, which is that "every act of inferential communication conveys a presumption of its own optimal relevance." 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.

Taking the approach of analyzing general communicative principles that shape interpretations of figurative language, our goal in this paper 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. In what follows, we first review core empirical phenomena in figurative language and highlight existing research and open questions regarding factors that shape figurative language understanding. Next, we will describe the ways in which many of these factors can be integrated through a framework of pragmatic reasoning. We introduce Rational Speech Act (RSA) models, a family of computational models that formalize 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 (R. Gibbs & Colston, 1999; Gibbs Jr & Colston, 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 quickly 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 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.

One characterization of this distinction is that figurative utterances are often used with the intention of producing particular effects and in order to accomplish discourse goals beyond relaying objective information about the world. R. 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 (?, ?) . Other work has shown that verbal irony can heighten or soften criticism (?, ?), elicit emotional reactions (Leggitt & Gibbs, 2000), highlight group membership (?, ?), 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 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 (R. 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.

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, many of which do not seem intuitively figurative (e.g. *apodioxis*, which means to indignantly reject an argument as false) (R. Roberts & Kreuz, 1994). Kreuz and Roberts (1993) 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", when 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?"). R. 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, in this paper we will focus on *hyperbole*, *irony*, and *metaphor*. Here we briefly describe each of the three tropes and highlight 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. 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 (Smith, 1969). In a modern analysis of a corpus of spoken English, (McCarthy & Carter, 2004) found that hyperbole occurs frequently in everyday conversations and are 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. R. 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 then 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 (R. Roberts & Kreuz, 1994; R. Gibbs & Colston, 1999): for example, saying "Such lovely weather we are having" in the middle of a storm. 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 ("lovely") is quite 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 "lovely" 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. Kreuz and Glucksberg (1989) 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 (Sperber & Wilson, 1981). 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 (R. Gibbs & Colston, 1999; R. 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 et al., 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 Jr & 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 Jr & 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 lovely" (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; 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 lovely 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 lovely 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. As a result, it is important to consider the various ways in which we can formulate literal meaning and analyze the ways in which it contributes to figurative meaning.

According to Frege (1984), the literal meaning of a sentence is its interpretation given no information about who said the sentence, when, or why. This view led to assumptions about the literal meaning of a sentence as being determined only by the meanings of its component words and how they are composed, independent of any extra-linguistic information. As a response against this traditional view, J. R. Searle (1978) argued that literal meaning is not entirely independent of extra-linguistic information and instead relies heavily on background 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 knowledge (R. W. Gibbs, 1984). Searle draws a distinction between this kind of background knowledge, which he believes is needed to determine literal meanings, and the context in which sentences are uttered, which helps determine contextual meanings. 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; R. W. Gibbs, 1984).

Another way literal meaning has been conceptualized is through conventionality (Davies, 1996). Perhaps literal meanings are those that are more conventional, such that without additional context, one chooses the sentence's most conventional meaning by default (Recanati, 2002). However, equating literalness with conventionality is problematic when we consider idioms. For example, the conventional meaning of "kick the bucket" (to die) is certainly different from our intuitive sense of its "literal" meaning (to come into contact with a bucket with one's foot). Giora (1997) proposed that instead of focusing on the literal/figurative distinction, a more useful dimension along which to analyze meaning is salience. She introduces the graded salience hypothesis, where more salient meanings—defined as more frequent, context-independent, and prominent—are accessed first and are "default," regardless of whether or not they are literal. While the idea of salience many explain when and why certain meanings are accessed automatically independent of context and literalness (Giora, 1999), it is not always clear precisely how salience should be measured or operationalized (Gibbs Jr & Colston, 2012). Despite vibrant discussions of these and related matters in the past few decades, the problem of literal meaning is still largely unresolved (Recanati, 2004; Cruse, 2004; Carston, 2008; Coulson & Oakley, 2005)

note here
how the
meanings
looking
rather
e and un-

Encyclopedic knowledge. One of the most important insights in the study of language use is that interlocutors make heavy use of common ground in order to

communicate (Clark, 1996). This common ground includes the extra-linguistic information and world knowledge shared among interlocutors, which can be intimately tied to linguistic meaning. Indeed, some researchers propose that the meaning of a word itself includes a network of background knowledge shared among people in a community, known as an encyclopedic approach (Taylor, 2003; Langacker, 1987). This encyclopedic knowledge often reflects 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 all politicians are believed to be honest, 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 a "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 fine-grained 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 it 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 thiss 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 to introduce. Given the importance of local context in shaping interpretation, a model of figurative language understanding should flexibly integrate contextual information.

more examp from the lit ture?

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 fine-grained and rich figurative meanings 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 it 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; R. W. Gibbs, 1984; Clark, 1996; Stalnaker, 2002), to our knowledge there is no formal model that explicitly describes the relationships among these components and integrates them to produce concrete, fine-grained predictions that can be evaluated against empirical data. 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 o
why it's use
to have a fo
model

add note th
these comp
nents are n
unique to fi
rative langu
understand

Probabilistic Models of Pragmatics

In recent years, a family of computational models have emerged that use probabilistic tools to formalize principles of communication, called Rational Speech Act (RSA) models (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, Goodman, & Levy, 2012), vagueness and context-sensitivity in gradable adjectives (Lassiter & Goodman, 2014), the pragmatic use and interpretation of prosody (Bergen & Goodman, n.d.), effects in syllogistic reasoning (Tessler & Goodman, n.d.), and more (Goodman & Lassiter, 2014).

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:

$$S_1(u|m) \propto L_0(m|u) \cdot e^{-Cost(u)} \tag{1}$$

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:

$$L_1(m|u) \propto P(m)S_1(u|m) \tag{2}$$

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 [1].

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. 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

[1] 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.

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 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 (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:

$$S_1(u|s,a) = \sum_o S_1(u|o,a)P(o|a,s) \tag{3}$$

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:

$$L_1(s|u,a) \propto S_1(u|s,a)P(s) \tag{4}$$

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, and 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. Since 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.

Rational Speech Act 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 certain aspects of background knowledge and prior beliefs. For example, to compute the probability that Cam is a wolf given the utterance "Cam is a wolf," the pragmatic listener consider the prior probability that Cam is 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:

$$L_0(l, \vec{E}|u) = \begin{cases} P(\vec{E}|l) & \text{if } l = u \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

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 interprets utterances literally. Given the utterance "Bob is a wolf," the literal listener will believe that Bob is a fierce, furry, and loyal wolf with some probability ($P(\vec{E}|l)$); however, she does *not* believe that Bob is a fierce person or any kind of person at all, because she believes that he is literally a wolf with probability 1. For figurative meaning to arise, the speaker and pragmatic listener must reason about which dimension of meaning is relevant to the QUD. We formalize relevance to the QUD by introducing a function Q, which projects the meaning that a literal listener derives from an utterance onto only the dimension that is under discussion. This leads to the following utility function for speaker S_1:

$$U(u|l, \vec{E}, Q) = \log \sum_{l,\vec{E}} \delta_{Q(l,\vec{E})=Q(l',\vec{E}')} L_0(l', \vec{E}'|u) \tag{6}$$

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

$$S_1(u|l, \vec{E}, Q) \propto e^{\lambda U(u|l,\vec{E},Q)}, \tag{7}$$

where λ is a speaker rationality parameter (Luce, 2005). As before, 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 question under discussion, 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. Since the literal listener is likely to believe that Bob is fierce if she believes that Bob is a wolf, the speaker is motivated to say "Bob is a wolf" to get her to believe that Bob is a wolf and thus fierce. Furthermore, since the speaker only cares to communicate Bob's fierceness and not which species Bob belongs to, he does not mind that the literal listener will believe that Bob is actually a wolf. The pragmatic listener simulates the speaker's choice of utterance given different QUDs. Combining this simulation with the prior belief that Bob is very unlikely to actually be a wolf, the pragmatic listener ultimately believes that Bob is a fierce person, which is the intuitive interpretation of the utterance "Bob is a wolf." This simple example suggests that by incorporating QUD inference, the RSA model is able to produce 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.

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: numerals. 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 much less than 1000 degrees.

In ? (?), 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 too 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 communicative goals that the speaker may have, including the goal to communicate affect about the price of the item. By reasoning about relevance to these communicate goals, the model captures a basic feature of hyperbole: utterances whose literal meanings are less likely given the price prior are more likely to be interpreted hyperbolically. For example, "The watch cost 1000 dollars" is more likely to be interpreted hyperbolically than "The laptop cost 1000 dollars."

To quantitatively evaluate the model's predictions, we asked participants to interpret potentially hyperbolic utterances. For example, given that Sam said: "The watch cost 1000 dollars," how likely is it that the watch cost x dollars? For all utterances, we then compared the model's and participants' interpretations. The model predictions are highly correlated with people's interpretations ($r = 0.968, p < 0.0001$) (Figure 1), suggesting that the qRSA model is able to combine linguistic information, background knowledge, and reasoning about the speaker's goals to interpret hyperbolic utterances.

In addition to producing the appropriate corrective response to hyperbolic utterances, the model also captures the affective subtext of hyperbole. We conducted a separate experiment to examine peoples' interpretation of affect in hyperbolic versus literal utterances. Participants read scenarios in which a speaker bought an item that cost s

dollars and says it cost u dollars, where $u \geq s$. They then rated how likely it is that the buyer thinks the item was too expensive. Results showed that utterances u where $u > s$ (hyperbolic utterances) are rated as significantly more likely to convey affect than utterances where $u{=}s$ ($F(1, 25) = 12.57, p < 0.005$). Moreover, if a watch actually cost 100 dollars and Sam says something hyperbolic such as "The watch cost 1000 dollars," people are more likely to believe that Sam thinks the watch is too expensive than if the watch actually cost 1000 dollars and Sam says "The watch cost 1000 dollars." This suggests that listeners infer affect from hyperbolic utterances above and beyond the affect associated with a given price state. Quantitatively, we compared model and human interpretations of affect for each of the 45 utterance and price state pairs (u, s) where $u \geq s$. While there is a significant amount of noise in the human judgments (average split-half correlation is 0.833), the model predicts human interpretations of the utterances' affective subtext significantly better than chance ($r = 0.775, p < 0.00001$), capturing most of the reliable variation in these data (Figure 2).

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

Verbal Irony

An ironic statement describes something as contrary to what it actually is (R. Roberts & Kreuz, 1994; R. Gibbs & Colston, 1999). For example, a speaker who says "Such lovely weather we are having" in the middle of a storm means that the weather is *not* lovely and expresses a negative attitude towards it. How do people appropriately

interpret these superficially positive or negative utterances? Can our model use QUD inference to interpret an utterance when its literal meaning is not just an exaggerated version of the intended meaning, but rather its opposite? In this section, we will examine the consequence of expanding the set of emotions we consider to an empirically derived affect space. We show that this minimal change enables the qRSA model to capture many of the rich inferences resulting from verbal irony.

In what follows, we will examine interpretations of potentially ironic utterances in an innocuous domain—the weather. We chose the weather as the victim of irony for several reasons. First, people are quite familiar with talking (and complaining) about the weather. Second, we can visually represent the weather to participants with minimal linguistic description in order to obtain measures of nonlinguistic contextual knowledge. Finally, given the critical role that context plays in understanding irony, we can vary the weather states to observe how the same utterance is interpreted differently given different contextual knowledge. This offers to our knowledge the first fine-grained manipulation and quantitative measure of context in studies of irony. We first explore how an enriched space of affect impacts the qRSA model and find that it produces ironic interpretations in some simple simulations. We then present two behavioral experiments that examine people's interpretations of utterances given different weather contexts. We show that by accounting for two types of affective dimensions, valence and arousal, our model produces interpretations that closely match humans'.

Model. Following the qRSA model described in (?, ?), a speaker chooses an utterance that most effectively communicates information regarding the question under discussion (QUD) to a literal listener. We consider a meaning space that consists of the variables s, A, where s is the state of the world, and A represents the speaker's (potentially multidimensional) affect towards the state. We formalize a QUD as a projection from the full meaning space to the subset of interest to the speaker, which could be s or any of the dimensions of A. We specify the speaker's utility as information gained by the listener

about the topic of interest—the negative surprisal of the true state under the listener's distribution given an utterance, u, along the QUD dimension, q. This leads to the following utility function:

$$U(u|s, A, q) = \log \sum_{s', A'} \delta_{q(s,A)=q(s',A')} L_0(s', A'|u) \tag{8}$$

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 chooses an utterance according to a softmax decision rule (?, ?): $S_1(u|s, A, q) \propto e^{\lambda U(u|s,A,q)}$, where λ is the rationality parameter. A pragmatic listener L_1 then takes into account prior knowledge and his internal model of the speaker to determine the state of the world as well as the speaker's affect. Because L_1 is uncertain about the QUD, he marginalizes over the possible QUDs under consideration:

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

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

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). 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 show the model's interpretation of "The weather is terrible" using this one-dimensional affect space. The model is capable of non-literal interpretation: it produces a hyperbolic interpretation (that the weather is merely `bad`)

given "The weather is terrible" in the bad weather situation. However, it produces a literal interpretation (that the weather is `terrible`) in the amazing weather situation. Since a pragmatic listener that only considers emotional valence does not believe that the speaker has any reason to choose a negative utterance to express positive affect (because the utterance communicates no true information), a model that only considers emotional valence is unlikely to infer a positive world state from a negative utterance (and vice versa), thus failing to evidence verbal irony.

What true information *could* a speaker communicate about a positive world state using a negative utterance? Affective science identifies two dimensions, termed valence and arousal, that underly the slew of emotions people experience (?, ?). For example, *anger* is a negative valence and high arousal emotion, while *contentment* is a positive valence and low arousal emotion. Could speakers leverage the arousal dimension to convey high arousal and positive affect (e.g. excitement) using utterances whose literal meanings are associated with high arousal but negative affect (e.g. "The weather is terrible!")? We test the consequences of incorporating the arousal dimension. The orange lines in Figure 3 show simulations of the qRSA model with a two-dimensional affect space: whether the speaker feels negative/positive valence and low/high arousal towards the weather state. Given strong prior belief that the weather state is `bad`, the model interprets "The weather is terrible" to mean that the weather is likely to be `bad`, again producing a hyperbolic interpretation. However, given strong prior belief that the weather is `amazing`, the model now places much greater probability on the ironical interpretation of "The weather is terrible," meaning that the weather is likely `amazing`. This is because, with the enriched two-dimensional affect space, the pragmatic listener realizes that the speaker may be using "terrible" to communicate high emotional arousal. Note that this result is not simply due to the model falling back on the prior: given the same priors, the model interprets the neutral utterance "The weather is ok" as the weather state being `neutral` and not `amazing`. These simulations suggest that a psychologically realistic, two-dimensional affect

space enables the qRSA model to interpret ironic utterances in addition to hyperbolic ones.

Experiment 1a: Background knowledge for verbal irony. 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. The latter 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).

We selected nine images from Google Images that depict the weather. To cover a range of weather states, three of the images were of sunny weather, three of cloudy weather, and three of rainy or snowy weather. We refer to these images as weather contexts. Figure 4 shows these nine images. 49 native English speakers with IP addresses in the United States were recruited on Amazon's Mechanical Turk. Each participant saw all nine images in random order. In each trial, participants were told that a person (e.g. Ann) looks out the window and sees the view depicted by the image. They then indicated how Ann would rate the weather using a labeled 5-point Likert scale, ranging from `terrible`, `bad`, `neutral`, `good`, to `amazing`. Finally, participants used slider bars (end points labeled "Impossible" and "Absolutely certain") to rate how likely Ann is to feel each of the following seven emotions about the weather: *excited*, *happy*, *content*, *neutral*, *sad*, *disgusted*, and *angry*, which are common emotion categories (?, ?)[2]. The order of the emotions was randomized for each participant but remained consistent across trials [3]. **Results**

For each of the nine weather contexts, we obtained the number of participants who

[2]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.

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

gave each of the weather state ratings. By performing add-one Laplace smoothing on the counts, we computed a smoothed prior distribution over weather states given each context, namely $P(s)$ (Figure 5). To examine participants' ratings of the affect associated with each context, we first performed Principal Component Analysis (PCA) on the seven emotion category ratings. This allowed us to compress the ratings onto a lower-dimensional space and reveal the main affective dimensions that are important in this domain, as is often done in affective science (?, ?). We found that the first two principal components corresponded to the dimensions of emotional valence and emotional arousal, accounting for 69.14% and 13.86% of the variance in the data, respectively. The PCA represents emotion ratings for each trial as real values between negative and positive infinity on each of the dimensions. To map these values onto probability space, we first standardized the scores on each dimension to have zero mean and unit variance. We then used the cumulative distribution function to convert the standardized scores into values between 0 and 1. This gives us the probabilities of Ann feeling positive (vs. negative) valence and high (vs. low) arousal for each trial, which is a two-dimensional probabilistic representation of her affect. By calculating the average probabilities of positive valence and high arousal given each weather state rating, we obtain the probability of positive valence and high arousal associated with each weather state, namely $P(A|s)$ (Figure 7).

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

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 4 in random order. In each trial, participants were told that a person (e.g. Ann) and her friend are in a room looking out the window together and see the view depicted by the image. Ann says, "The weather is _____!" where the adjective is randomly selected at each trial from the

following set: "terrible," "bad," "ok," "good," and "amazing." Participants first rated how likely it is that Ann's statement is ironic using a slider with end points labeled "Definitely NOT ironic" and "Definitely ironic." They then indicated how Ann would actually rate the weather using a labeled 5-point Likert scale, ranging from `terrible`, `bad`, `neutral`, `good`, to `amazing`. Finally, participants used sliders to rate how likely it is that Ann feels each of seven emotions about the weather [4].

We first examined participants' irony ratings for each of the weather context and utterance pairs. We found a basic irony effect, where utterances whose polarities are inconsistent with the polarity of the weather context are rated as significantly more ironic than utterances whose polarities are consistent with the weather context ($t(34.16) = -11.12, p < 0.0001$). For example, "The weather is terrible" (a negative utterance) is rated as more ironic in weather context 1 (positive context) ($M = 0.90$, $SD = 0.21$) than in weather context 7 (negative context) ($M = 0.15$, $SD = 0.27$). A linear regression model with the polarity of the utterance, the polarity of the weather context, and their interaction as predictors of irony ratings produced an adjusted R^2 of 0.91, capturing most of the variance in the data. This suggests that participants' lay judgments of irony align with its basic definition: utterances whose apparent meanings are opposite in polarity to the speaker's intended meaning. Given the fact that participants can identify verbal irony based on its inconsistency with context, how do they then use context to determine the speaker's intended meaning? We examined participants' interpretations of utterances given contexts. For each of the 45 weather context (9) $\times$ utterance (5) pairs, we obtained the number of participants who gave each of the five weather state ratings (`terrible, bad, neutral, good, amazing`). We performed add-one Laplace smoothing on the counts to obtain a smoothed distribution over weather states given each context and utterance (solid lines in Figure 8). Results show that participants produce ironic interpretations of

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

utterances, such that the weather is most likely to be `amazing` given that the speaker said "The weather is terrible" in weather context 1. Participants also produce hyperbolic interpretations, such that the weather is most likely to be `bad` given that the speaker said "The weather is terrible" in weather context 7. This confirms the intuition that people are highly sensitive to context and use it both to determine when an utterance is not meant literally and to appropriately recover the intended meaning. Finally, we examine participants' inferences about the speaker's affect given utterances in context. We used the loadings from the PCA on emotion ratings from Experiment 1 to project the emotion ratings from Experiment 2 onto the same dimensions. We then standardized and converted the scores into values between 0 and 1, as before, which gives us probability ratings of the speaker feeling positive valence and high arousal given an utterance and weather context.

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 ($P(q_{state}) = 0.3$, $P(q_{valence}) = 0.3$, $P(q_{arousal}) = 0.4$)[5]. 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 9 shows scatter plots correlating model predictions with human interpretation data for each of the dimensions: weather state, valence, and arousal. The model predictions of weather state given utterance match humans' interpretations, with a correlation of 0.86. Since the split-half correlation for the human data is $\rho = 0.898$ (95%CI $= [0.892, 0.903]$)[6] we find that our model captures much of the explainable variance in

[5]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.

[6]Split-half correlations ρ were calculated by repeatedly bootstrapping samples from the data (sample

human judgements. The model predicts humans' interpretations of valence extremely well, with a correlation of 0.96, capturing essentially all of the explainable variance in the data ($\rho = 0.948 \pm 0.001$). Importantly, the model infers the appropriate valence even when it is inconsistent with the valence of the utterance's literal meaning. The model's predictions for emotional arousal match humans' with a correlation of 0.66, capturing a substantial amount of the explainable variance ($\rho = 0.763 \pm 0.005$). Furthermore, the absolute difference between the model's inferred valence and the valence of the utterance's literal meaning correlates significantly with people's irony ratings ($r = 0.86$, $\rho = 0.94 \pm 0.005$), suggesting that the model is able to use inconsistencies between literal and interpreted meanings to identify ironic uses.

We considered a series of simpler models to show that the full model using a two-dimensional affect space best predicts human interpretations. We first examined a model that interprets utterances literally, such that "The weather is terrible" is always interpreted as the weather state being `terrible`, along with the valence and arousal associated with `terrible` weather. We then examined a model that simply ignores the speaker's utterance and takes into account only the state and affect priors associated with each weather context. Finally, we examined the performance of the qRSA model with a unidimensional affect space (valence only). Table 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

---

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.

successfully leverages richer knowledge of affect and uses pragmatic reasoning to produce the appropriate figurative interpretations.

Discussion. In this work, we formalized intuitions about verbal irony understanding and clarified the role of shared prior knowledge in ironic interpretations. We explored the consequences of expanding the space of affect considered by previous Rational Speech Act models to account for verbal irony. By making a minimal extension to (?, ?)'s hyperbole model, we were able to capture people's fine-grained interpretations of ironic utterances in addition to hyperbole. This provides evidence that hyperbole and irony may operate using similar underlying principles of communication—reasoning about shared background knowledge as well as the speaker's affective goals.

Metaphor

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

Why do people choose to use metaphors to communicate? What are some characteristics of metaphor that contribute to its popularity? R. Roberts and Kreuz (1994) examined the discourse goals that people have when they use various figurative tropes. They found that the most common goals for using metaphor were to clarify (82%), to add interest (71%), to compare similarities (35%), to provoke thought (35%), and to be

eloquent (35%). Interestingly, although metaphors are defined as implicit comparisons, "to compare similarities" is less frequently listed as a goal than "to clarify" and "to add interest." This suggests that beyond examining the cognitive processes for comparing and aligning concepts that may be involved in metaphor understanding, it is also important to consider other higher-level communicative functions that metaphors may serve.

In addition to (R. Roberts & Kreuz, 1994)'s exploration of discourse goals, other researchers have suggested that metaphorical utterances can be used to efficiently express complex meanings (Ortony, 1975; Boerger, 2005; Glucksberg, 1989). In communication tasks where pairs of participants are separated by a screen and asked to refer to abstract geometrical objects, participants often prefer to describe objects analogically in terms of other known objects rather than use literal analytical descriptions (Clark & Wilkes-Gibbs, 1986; Glucksberg, 1989). Fussell and Krauss (1989) found that these analogical and figurative descriptions tend to be shorter than literal descriptions. In addition, they found that figurative descriptions are used significantly more often when the intended audience is one's self, where presumably there is a great deal of common ground, than when the intended audience is a different person. These findings suggest that people may be balancing efficiency and clarity when choosing figurative versus literal descriptions. (Glucksberg & Keysar, 1990) wrote, "Metaphors are used to communicate a complex, patterned set of properties in a shorthand that is understood by the members of a speech community who share relevant mutual knowledge" (pp. 16). As a result, a potential benefit of speaking in metaphor may be to utilize common ground to communicate both clearly and efficiently.

In my dissertation, I plan to examine the following (hypothesized) characteristics of metaphorical utterances: (1) metaphors lead to more striking, or extreme, interpretations than literal descriptions (2) metaphors can communicate information along various dimensions with a minimal number of words (3) the interpretation of a metaphor is highly sensitive to the local context (4) the aptness of a metaphor is related to whether there are

alternative utterances that can communicate the same amount of information more efficiently. To examine these hypotheses, I will test the qRSA model on different types of metaphorical interpretations.

Model sketch

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

We again introduce a literal listener L_0, who interprets the utterances as meaning that Bob is literally a giraffe. Since L_0 believes Bob is a giraffe, she also believes that Bob is likely to have features associated with giraffes, for example, being tall. The following equation represents the literal listener's interpretation, where c is Bob's category (either a "person" or a "giraffe"), and $\vec{f}$ is a vector representation of Bob's features. $P(\vec{f}|c)$ is thus the prior probability that a member of category c has feature vector $\vec{f}$.

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

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

$$U(u|Q, c, \vec{f}) = \log \sum_{c,\vec{f}} \delta_{Q(c,\vec{f})=Q(c',\vec{f}')} L_0(c', \vec{f}'|u) \tag{9}$$

Given this utility function, the speaker chooses an utterance according to a softmax decision rule that describes an approximately rational planner (?, ?), where λ is an

optimality parameter:

$$S_1(u|Q, c, \vec{f}) \propto e^{\lambda U(u|Q,c,\vec{f})}, \tag{10}$$

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

Study 1: "Hyperbolic" effects

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

Study 2: Non-paraphrasability

In Study 2, we use a series of experiments to show that some metaphorical utterances such as "Bob is an ox" allow speakers to communicate about multiple dimensions of Bob at once—for example, that Bob is strong, big, and slow. In order to communicate the same amount of information using a literal description, a speaker would need to choose a much

longer and costlier utterance, such as: "Bob is strong, big, and slow." We show that listeners infer information along more dimensions given a metaphorical utterance than given a single literal description. We then show that the qRSA model is able to capture these differences between literal and metaphorical utterances.

Study 3: Context-sensitivity

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

Study 4: Aptness and alternatives

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

General Discussion

References

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

Bach, K. (1994). Semantic slack: What is said and more. *Foundations of speech act theory: Philosophical and linguistic perspectives*, 267–291.

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

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

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

Carston, R. (2008). Linguistic communication and the semantics/pragmatics distinction. *Synthese*, *165*(3), 321–345.

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., & Gerrig, R. J. (1984). On the pretense theory of irony.

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

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

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.

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

Cruse, D. A. (2004). Meaning in language: An introduction to semantics and pragmatics.

Dascal, M. (1981). Contextualism. *Herman Parret, Marina Sbisà & Jef Verschueren (eds.)*

Possibilities and Limitations of Pragmatics. John Benjamins, Amsterdam.

Davies, M. (1996). Philosophy of language. *The Blackwell Companion to Philosophy, Second Edition*, 90–146.

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

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

Frege, G. (1984). On sense and reference. *Translations from the philosophical writings of Gottlob Frege*, 126–151.

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

Gentner, D. (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. (1994). *The poetics of mind.* Cambridge: Cambridge University Press.

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

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

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

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

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

Giora, R. (1997). Understanding figurative and literal language: The graded salience

hypothesis. *Cognitive linguistics*, *8*, 183–206.

Giora, R. (1999). On the priority of salient meanings: Studies of literal and figurative language. *Journal of Pragmatics*, *31*(7), 919–929.

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

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

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

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

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

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

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

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*.

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

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.

Kreuz, R. J., & Glucksberg, S. (1989). How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General*, *118*(4), 374.

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

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

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

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

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.

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).

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.

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).

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

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

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

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

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.

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

Recanati, F. (2002). Literal/nonliteral. *Midwest Studies in Philosophy*, *25*, 264–274.

Recanati, F. (2004). *Literal meaning*. Cambridge University Press.

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).

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., & Kreuz, R. (1994). Why do people use figurative language? *Psychological*

Science, *5*(3), 159–163.

Searle, J. (1979). *Metaphor in ortony, a.(1979) 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.

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

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

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

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

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

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

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

Tourangeau, R., & Sternberg, R. J. (1982). Understanding and appreciating metaphors. *Cognition*, *11*(3), 203–244.

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 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 6).*

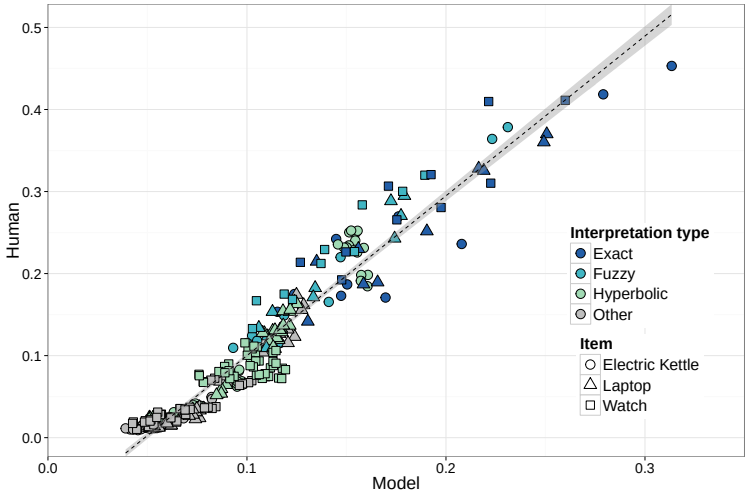

Figure 1. Model predictions v.s. average human interpretations. Each point represents an utterance and price state pair (u, s). The x-coordinate of each point is the probability of the model interpreting utterance u as meaning price state s; the y-coordinate is the empirical probability. Correlation between model and human interpretations is 0.968.

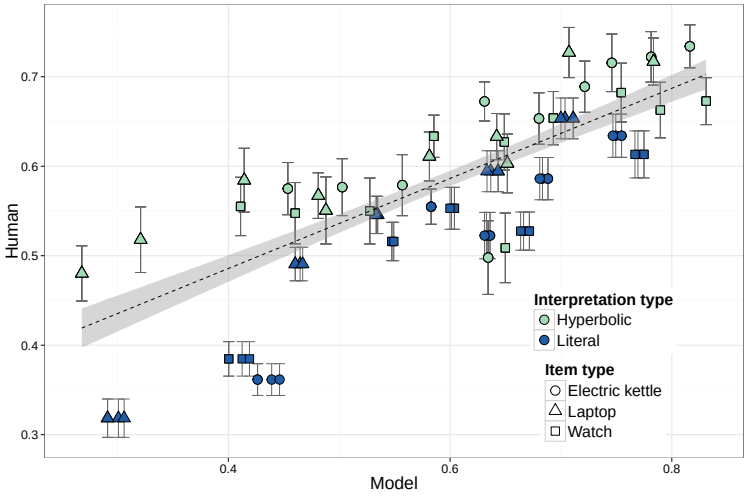

Figure 2. Model predictions of affect v.s. human ratings. Each point represents an utterance and price state pair (u, s). For pairs where $u = s$, the utterance is literal; for $u > s$, the utterance is hyperbolic. The x-coordinate of each point is the model's prediction of the probability that the utterance/price state pair conveys affect; the y-coordinate is participants' affect ratings (error bars are standard error). Correlation between model and humans is 0.775.

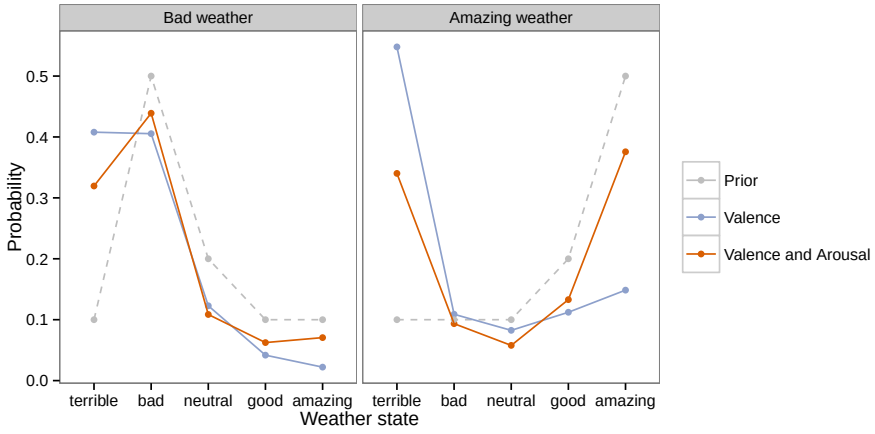

Figure 3. Model interpretations of "The weather is terrible" given different prior beliefs about the weather state and affect dimensions. Gray dotted lines indicate prior beliefs about weather states given a weather context; blue lines indicate interpretations when reasoning only about the speaker's valence; orange lines indicate interpretations when reasoning about both valence and arousal.

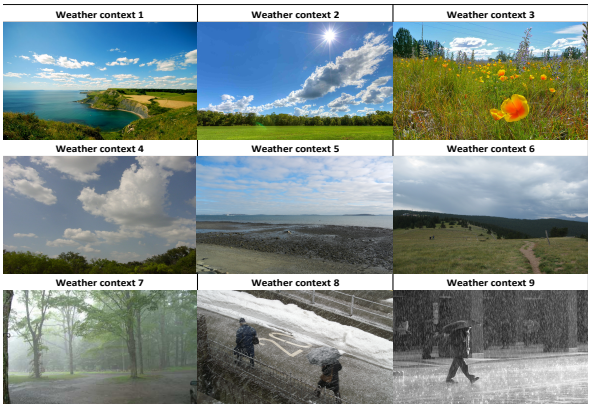

Figure 4. Weather images shown to participants in Experiments 1 and 2.

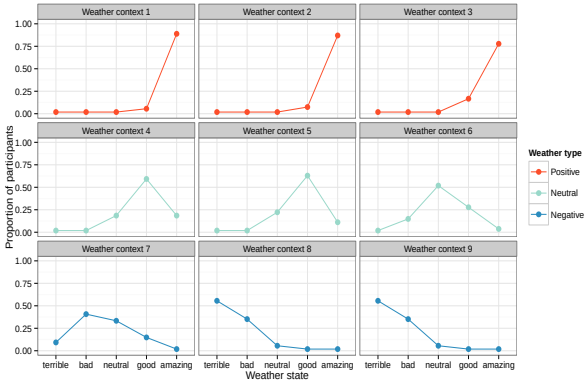

Figure 5. Proportion of participants who rated each weather context as each weather state.

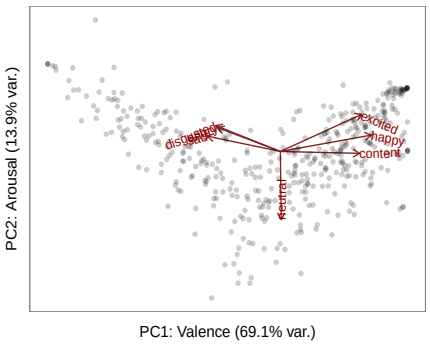

Figure 6. Biplot of the first two principle components of the seven emotion ratings, which roughly correspond to valence and arousal.

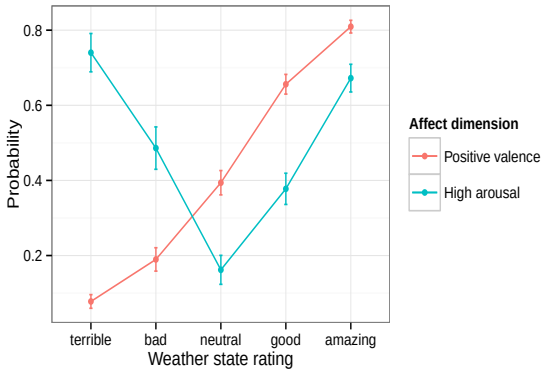

Figure 7. Average probabilities of positive valence and high arousal associated with each weather state; error bars are 95% confidence intervals.

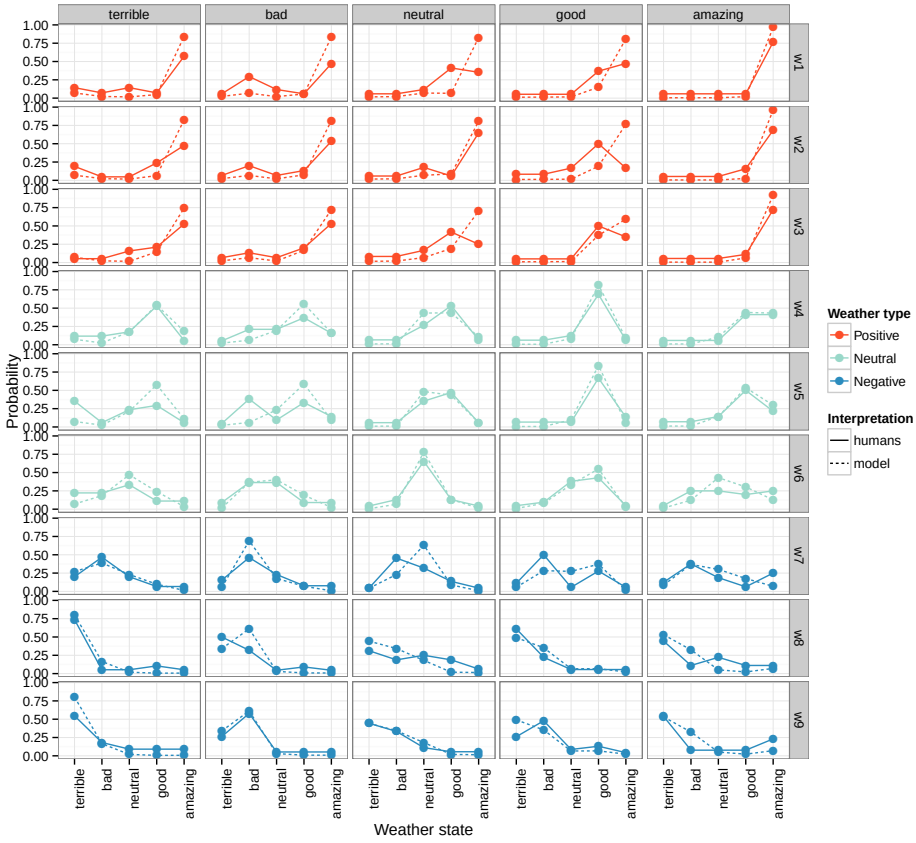

Figure 8. Model's and participants' inferences about the weather state (x-axis) given a weather context (row) and an utterance (column). Each panel represents interpretations of an utterance in a weather context. The solid lines are participants' ratings; the dotted lines are model's posterior distributions over weather states.

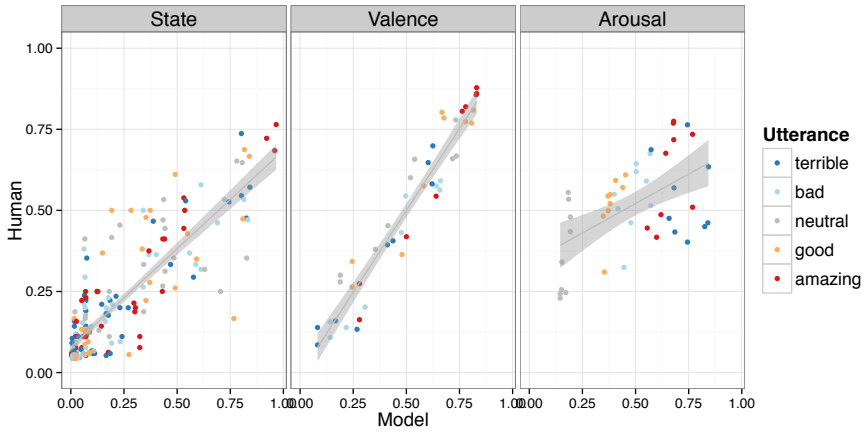

Figure 9. Scatter plot showing correlations between model predictions and human ratings for weather state, speaker valence, and speaker affect. Colors indicate utterances.