A Computational Model of Figurative Language Understanding

Justine T. Kao

Stanford University

Leon Bergen

Stanford University

Noah D. Goodman

Stanford University

Author Note

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

Leon Bergen, Department of Psychology, Stanford University.

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 use language to convey information that goes far beyond an utterance's literal meaning. In particular, figurative language such as hyperbole, irony, and metaphor showcases people's ability to infer relevant and true information from utterances that are false under their literal semantics. Although figurative language is prevalent in human communication and has been studied across many fields, how people arrive at appropriate interpretations remains unclear. 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 question under discussion (QUD) is able to predict people's interpretations of hyperbole, irony, and metaphor. We argue that despite apparent differences among subtypes of figurative language, the same computational framework flexibly produces fine-grained interpretations for a range of figurative 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

A Computational Model of Figurative Language Understanding

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 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. This ability to go beyond direct evidence (the words) to infer unobserved information (the meanings) is a hallmark of human intelligence that underlies many aspects of how we understand and interact with the world. How do our linguistic, cognitive, and social faculties work together to allow us to fluently and accurately understand the communicative intent behind figurative utterances?

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, psychologists have proposed various ways in which people align shared properties and analogous relations across different domains in order to understand metaphor, including the domain interaction model (Tourangeau & Sternberg, 1982), structure mapping model (Gentner & Wolff, 1997; Gentner, 1983), and category assertion model (Glucksberg, 2003). To explain other types

of figurative language such as verbal irony, separate explanations are posited such as pretense (Clark & Gerrig, 1984) or allusion (Sperber & Wilson, 1981) to address the specific factors that shape ironic interpretations. While these approaches significantly advance our understanding of the cognitive factors that guide metaphor and irony understanding, they tend to treat each type of figurative language as requiring a different mechanism distinct from or in addition to the process involved in standard language understanding. By relying on more specialized mechanisms for figurative language, these approaches leave open the question of how utterances such as "My surgeon is a butcher" or "Such lovely weather we're having" (uttered in the middle of a storm) trigger these specialized mechanisms in the first place.

A different approach to studying figurative language focuses on how people use general communicative principles to arrive at contextually appropriate interpretations (Grice, 1975; J. Searle, 1979; Sperber & Wilson, 2008; Ortony, 1993; Tendahl & Gibbs, 2008). Two main theories in the pragmatics literature have taken this approach to explain figurative language: the standard pragmatic theory and relevance theory. The standard pragmatic view analyzes figurative utterances using the cooperative principle and standard Gricean maxims, which state that speakers tend to produce utterances that adhere to principles of quality (truthfulness), quantity (informativeness), relevance, and manner (e.g. brevity, orderliness, clarity) (Grice, 1975; J. Searle, 1979). Under this view, figurative utterances are understood through a three-step process: (1) determine the literal meaning of the utterance (2) determine whether the literal meaning violates the quality maxim by being untruthful (3) reanalyze the utterance to identify implied or figurative meanings that would allow the utterance to adhere to the Gricean maxims. Although the standard pragmatic view is appealing in that it fits naturally within a more general theory of language understanding, it has met with several criticisms. One of the critiques is the fact that many figurative statements do not violate the quality maxim because their literal meanings can also be true. "No man is an island," for example, is a literally true statement

(there does not exist a man that is literally a piece of land surrounded by water) in addition to a metaphorically meaningful one (people do not exist in isolation) (Gibbs, 1992). By relying on the violation of the quality maxim, the standard pragmatic view does not provide a satisfying explanation for how figurative meanings arise from these types of utterances. An even more common criticism in the psycholinguistics literature is that the standard pragmatic view requires the listener to first access the literal meaning of the utterance, verify that the literal meaning is false, compute potential figurative interpretations, and then select the interpretation that best satisfies conversational maxims. Given the extra steps involved, this model would predict that people should take longer to interpret figurative utterances than literal utterances. However, many experiments have shown that the figurative meanings of irony and metaphor can be accessed as quickly or even more quickly than their literal meanings given supporting contexts (Glucksberg, 2003; Gildea & Glucksberg, 1983; Gibbs, 1992). These empirical findings suggest that literal meanings do not have to be explicitly computed and then rejected before appropriate figurative interpretations emerge.

Relevance theory, on the other hand, is a general theory of communication whose central claim is that human cognition is governed by a tendency to maximize relevance (Sperber, Wilson, & Ran, 1986). Instead of using Grice's four maxims as the guiding principle for figurative interpretation, relevance theory proposes that the principle of relevance is sufficient for explaining a range of phenomena in communication and cognition more generally. More specifically, the interpretation of all language involves maximizing the relevance of the interpretation to a contextually determined topic (Sperber & Wilson, 2008; Tendahl & Gibbs, 2008). As a result, interpretations of the same utterance can vary dramatically given different topics. Suppose two interlocutors, Ann and Bob, are discussing their friend Cam. Ann asks, "Does Cam have a fever?" and Bob replies, "He is boiling." Ann will interpret Bob's utterance to mean that Cam has a very high temperature. If, on the other hand, Ann asks, "Is Cam upset?" and Bob replies, "He's boiling," Ann should

interpret Bob's utterance to mean that Cam is very angry. The word "boiling" receives different figurative meanings in these two contexts: in the first, Cam is very hot but not literally at boiling point; in the second, Cam is experiencing intense anger. Ann arrives at the appropriate interpretation by assuming that Bob's utterance provides maximally relevant information regarding her question. Under the relevance theoretic view, figurative uses such as hyperbole and metaphor are not distinct from literal language, but rather lie on a continuum of "loose uses" that all require listeners to use the principle of relevance to recover the intended meanings. This view situates figurative language within a general theory of communication and has been argued to provide a complementary perspective to cognitive linguistics in the study of metaphor (Tendahl & Gibbs, 2008). However, one concern with relevance theory is that the concept of relevance, while intuitively appealing, has not been clearly operationalized or tested in a quantitative manner to determine its specific role in figurative language understanding.

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 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 formalizes communication as recursive social reasoning. We then show that natural but critical extensions can be made

to RSA models to account for figurative language. We adapt the extended model to three types of figurative language: hyperbole, verbal irony, and metaphor, and present behavioral data and modeling results. Finally, we discuss the insights this model reveals about figurative language understanding as well as future research that our modeling approach licenses. We argue that a general model of figurative language enables us to more precisely examine the ways in which semantics and principles of communication interact to generate rich linguistic meaning.

Figurative Language: The Phenomena

Figurative language is often defined as utterances whose intended meanings differ in various ways from their "literal" or standard meanings (Gibbs & Colston, 1999, 2012). At first glance, this definition seems straightforward and corresponds with our intuitions regarding which usages of language are literal and which are figurative; however, it grows murky upon closer inspection. Suppose a speaker Bob says, "I arrived late and the theater was full." Since it is implausible that the entire space of the theater was occupied from floor to ceiling, the sentence's strict literal meaning appears to be false. Instead, Bob most likely intends to communicate that he had difficulty finding an empty seat at the theater. This is an example of "loose talk," also described as pragmatic slack, where a speaker uses a proposition Q (e.g. the theater was full) in order to communicate a set of propositions that can be derived from Q (e.g. there were a lot of people at the theater, and Bob was unable to find a seat), without being committed to the truthfulness of Q (Sperber & Wilson, 1985; Lasersohn, 1999; Bach, 1994).

On the other hand, consider an utterance such as, "Bob is always late," produced by an annoyed speaker. In order for the literal meaning of the utterance to be true, for all cases in which Bob can be either late or on time, Bob must be late in 100% of the cases. However, one can easily interpret the utterance to mean that the speaker thinks Bob is very often (but not literally always) late, thus arriving at an interpretation that differs

from the literal meaning of the utterance. Under these analyses, the intended meanings of both of these utterances ("The theater was full" and "Bob is always late") differ from their "literal" meanings. Indeed, Sperber and Wilson (1985) claim that there is no discontinuity between loose and figurative uses of language; both exploit the principle of relevance in order to express what is derivable from the utterance without committing to the truth of the literal meaning of the utterance. At the same time, a sentence such as "Bob is always late" feels qualitatively different from a sentence such as "The theater was full" and is more easily recognized as hyperbole. In fact, it has been observed that some utterances are intuitively recognized as "figurative" while other are not (Coulson & Oakley, 2005), which suggests that figurative language may be a psychologically meaningful category distinct from most other loose uses.

What factors, if any, distinguish figurative uses from loose uses? One distinction is that figurative utterances are often used with the intention to produce particular effects and in order to accomplish discourse goals beyond relaying objective information about the world. R. Roberts and Kreuz (1994) examined the discourse goals that motivate people to use various figurative tropes and identified a taxonomy of 19 discourse goals, such as to convey emotion, to emphasize, to be humorous, or to be eloquent. Colston and Keller (1998) showed that hyperbole and irony are often used to express surprise. In addition, hyperbole and irony are more often used with friends and may signal social intimacy between speaker and listener (Gibbs, 2000; Pexman & Zvaigzne, 2004; Kreuz, 1996). Other work has shown that verbal irony can heighten or soften criticism (Colston, 1997), elicit emotional reactions (Leggitt & Gibbs, 2000), highlight group membership (Gibbs, 2000), and express affective attitudes (Colston & Keller, 1998). Still other researchers have suggested that metaphors are often used to express subjective attitudes towards the subject (Ortony, 1979), and that subjective sentences frequently contain figures of speech such as metaphor and hyperbole (Riloff, Wiebe, & Phillips, 2005). It is possible that the intuitive judgment of figurativeness involves recognizing that the speaker's intent is not to

communicate objective information about the world, but rather to produce one or more of these rhetorical effects. As a result, it may be important to consider the affective subtexts and social information that figurative language communicates above and beyond most loose uses of language.

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

Types of figurative language

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

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

Hyperbole. A hyperbole is an exaggerated statement that purposefully presents its subject as more striking or extreme than it actually is (R. 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 is often used in humorous and other affective contexts. Norrick (1982) proposed that hyperbole is characterized by three properties: its affective dimension, its pragmatic nature, and its function as a vertical-scale metaphor where the comparison is between different positions on a scale rather than between discrete concepts. Gibbs (1994) makes a distinction between hyperbole and overstatement, where the former is purposefully produced for rhetorical effect. For a hyperbolic statement to be interpreted successfully, the listener must recognize the non-veridicality of the statement, thus entering an activity of joint pretense (Clark, 1996). Hyperbolic statements often include extreme case formulations (e.g. “It was the biggest

storm in the history of the universe") or implausible descriptions (e.g. "It's a thousand degrees outside.") These demonstrations of non-veridicality require the listener to produce what Fogelin (2011) called a "corrective" response that is more in line with reality.

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

Metaphor. Metaphors are utterances that implicitly compare ideas or concepts from different domains. They are extremely prevalent in both literary and everyday language (Gibbs & Colston, 1999; R. Roberts & Kreuz, 1994). For example, "Juliet is the

sun" expresses Juliet's beauty; "My lawyer is a shark" communicates the lawyer's ruthlessness; and "Art washes away from the soul the dust of everyday life" allows Picasso to compare "art" to a cleansing fluid and "the soul" to a physical object that collects dust, which gracefully accomplishes two poetic metaphors at once. One can find traces of metaphoricity even in mundane utterances such as "I waited for a long time," where the spatial term "long" is used to describe the abstract domain of time (Lakoff, 1993). Due in part to its ubiquity and in part to the possibility that metaphor is intimately tied to our ability to create mappings between concrete experiences and abstract concepts (Lakoff & Johnson, 2008), metaphor is by far the most widely studied trope in cognitive science and related fields (Gibbs & Colston, 2012). Researchers have suggested that metaphor requires aligning analogical structures between two domains and can shape our reasoning and inferences (Gentner & Wolff, 1997; Thibodeau & Boroditsky, 2011). Evidence that metaphors are often processed as quickly as literal statements suggests that metaphor understanding does not require first accessing literal meanings or necessarily involve different processing mechanisms from literal language (Glucksberg, 2003; Gibbs & Colston, 2012).

Factors that shape figurative interpretation

In reviewing these three figurative tropes, some common features emerge. First, each example from these three tropes produces multiple interpretations that are distinct and highly different from each other (e.g. "It's a thousand degrees outside" (literal) v.s. "It's unexpectedly hot outside, like 90 degrees"; "The weather is amazing" (literal) v.s. "The weather is terrible"; "Juliet is made of hot plasma" (literal) v.s. "Juliet is beautiful"). Second, the intended meanings of these utterances are related to their "literal" meanings in non-arbitrary ways (e.g. a thousand degrees and 90 degrees are both unexpectedly high; "beautiful" and "terrible" both describe an extreme attitude towards the weather; the sun and Juliet are both very important and appealing to Romeo). Third, these utterances tend

to express speakers' subjective experiences and attitudes rather than objective information about the world. Finally, a great deal of common ground is required to successfully interpret these utterances. For example, interpretation of an utterance such as "Such beautiful weather we are having," depends upon the speaker and listener's mutual beliefs about the relevant state of the world (e.g. it is raining), their shared background knowledge (e.g. sunshine is usually preferable to rain), and mutual awareness of potential discourse goals (e.g. the speaker wants to convey her opinion about the weather). Because the interpretation of such utterances depends upon these different flavors of common ground, it tends to be highly sensitive to changes in context. Here we will examine the various factors that together shape the interpretation of a figurative utterance in more detail.

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

Encyclopedic knowledge. One way in which literal meaning gives rise to intended meaning is through encyclopedic knowledge, which includes a network of background knowledge shared among people in a community (Taylor, 2003; Langacker, 1987). Indeed, some researchers propose that the meaning of a word itself includes encyclopedic knowledge. J. R. Searle (1978) argued that literal meaning is not entirely independent of extra-linguistic information and instead relies heavily on this kind of

add a note
about how
literal mean
we're lookin
at are rathe
simple and
controversia

encyclopedic knowledge. For example, the literal meanings of "Sally cut the cake" and "Sally cut the grass" depend on the manners in which cake and grass are usually cut, which is encoded in background encyclopedic knowledge (Gibbs, 1984). However, other linguists and philosophers argue that Searle "demands too much from literal meaning" and conflates the literal meaning of a sentence with its intended speaker meaning (Dascal, 1981; J. J. Katz, 1981; Gibbs, 1984).

Despite the fact that the distinction between literal meaning and encyclopedic knowledge is not always clear, encyclopedic knowledge often goes beyond the strict literal meanings of utterances to include stereotypes, conventions, and a community's beliefs and practices, which in turn shape the interpretation of language. For example, suppose Ann asks Bob, "Is Cam an honest person?" and Bob replies, "He's a politician." Ann will likely interpret Bob's utterance to mean no, he does not believe that Cam is an honest person. This interpretation arises because while the dictionary meaning of "politician" is "a person who is professionally involved in politics, especially as a holder of or a candidate for an elected office," the encyclopedic meaning of the word can encompass many more features and connotations, such as dishonesty and corruption. Bob's utterance not only asserts Cam's profession (the literal, dictionary meaning of "politician"), but also attributes features associated with that profession to Cam (e.g. dishonesty, corruption). Ann is able to successfully interpret Bob's utterance, and Bob is able to successfully use this utterance, because they both have access to the relevant encyclopedic meaning of "politician." Naturally, Ann's interpretation is sensitive to the contents of the background knowledge they share. If Ann and Bob belong to a community where politicians are associated with honesty, then Ann would interpret Bob's reply to mean that yes, he believes Cam is an honest person. Similarly, "It's a thousand degrees outside" is interpreted as "It's unbearably hot outside" partly based on the encyclopedic knowledge that "a thousand degrees" is exceedingly hot, and that one is unlikely to survive under that temperature. As a result, the encyclopedic knowledge that interlocutors share can heavily influence the

interpretation of figurative utterances.

Prior beliefs. In addition to encyclopedic knowledge, interpretation of language depends upon the listener's prior beliefs and expectations about various states of the world (Clark, 1991). Hörmann (1983) showed that people's interpretation of quantifiers such as "several" and "few" vary based on the kinds of objects to which they refer. For example, "several crumbs" is interpreted to mean around 10 crumbs, while "several mountains" is interpreted to mean around 5 mountains. Clark (1991) explains this phenomena using the "principle of possibilities:" to interpret language, people make use of their prior expectations about what situations or worlds are possible, as well as how likely those worlds are. To interpret "several crumbs" and "several mountains," people consider the number of crumbs and mountains that typically inhabit a scene or situation. Since a typical situation involving crumbs is likely to contain more crumbs than a typical situation involving mountains, the interpretation of "several" results in a higher number for "several crumbs" than in "several mountains." Given that prior beliefs affect the interpretation of superficially straightforward terms such as "several," it is unsurprising that prior beliefs factor into the interpretation of figurative language as well. In the dialogue between Ann and Bob, Ann's interpretation of the utterance "He's a politician" is sensitive to her prior beliefs about Cam. Suppose prior to her conversation with Bob, Ann did not know what Cam does for a living. She will have learned two facts about Cam from Bob's utterance: Cam is a politician, and Cam is not an honest person. Suppose, on the other hand, that Ann knew beforehand that Cam is a politician, and knows that Bob knows that she knows that Cam is a politician. She will not have learned anything new about Cam's profession from Bob; however, even though she already knows that politicians in general are believed to be dishonest, Bob's utterance makes her more certain that Bob thinks Cam *in particular* is dishonest, because that is the most informative and relevant interpretation given her question about Cam's honesty and given that she already knows Cam's profession. Finally, suppose Ann knows that Cam is not professionally involved in politics at all. How will Ann

interpret Bob's utterance? Instead of updating her beliefs about Cam using the dictionary meaning of "politician," she will rely on its encyclopedic meaning to conclude that Cam is dishonest (but not professionally involved in politics), resulting in a metaphorical interpretation. These examples show that interpretation of the same utterance in the same local context can vary in a rich and subtle manner based on the speaker and listener's prior beliefs.

examples

the litera-

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

Pragmatic reasoning. A critical insight in communication is that a speaker does not produce utterances in a social vacuum; he considers the listener's beliefs, goals, and disposition to determine which utterance is most effective in a given situation, which Clark and Murphy (1982) termed "audience design." In turn, a listener considers the speaker's beliefs, goals, and disposition as well as the speaker's representation of *her* beliefs, goals, and disposition in order to select the most likely meaning of an utterance (Clark, 1996; Levinson, 2000; Grice, 1975). Furthermore, listeners tend to assume speakers to be rational and cooperative agents who aim to be informative, known as the Cooperative Principle (Grice, 1975; Clark, 1996; Levinson, 2000). When interpreting an utterance, a listener uses these assumptions of rationality and informativeness to reason about what meaning a speaker could want to convey that would lead him to choose a particular utterance. This recursive social reasoning between listener and speaker is responsible for many phenomena in pragmatics and language understanding, such as various types of conversational implicatures (Horn, 2006; Levinson, 2000).

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

chose the metaphor "He's a politician" instead of "He's a salesman," both of which convey persuasiveness, suggests that Bob wants to communicate specific features about Cam such as deceptiveness and cunning, rather than pushiness. Reasoning about the speaker's choice of utterance and available alternatives allows the listener to derive rich figurative meanings as well as their subtleties using basic principles of communication. A theory of figurative language as a communicative act should thus incorporate the speaker's intent as well as how the listener reasons about this intent in various communicative contexts.

Putting it all together. While many researchers have suggested that the construction of meaning involves an interplay of the components outlined above (Coulson & Oakley, 2005; Gibbs, 1984; Clark, 1996; Stalnaker, 2002), to our knowledge there is no formal model that explicitly describes the relationships among these components and integrates them to produce concrete, quantitative, and fine-grained predictions that can be evaluated against empirical data. Here we propose a formal modeling framework for figurative language understanding that incorporates these components and captures the recursive social nature of communication. We show that these components are sufficient for producing appropriate interpretations of figurative utterances as well as rich affective and social subtexts.

rate on
t's useful
ve a formal
l

ote that
compo-
are not
e to figu-
language
rstanding

Probabilistic Models of Language Understanding

In recent years, a family of computational models have emerged that use probabilistic tools to formalize principles of communication, called Rational Speech 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, 2015), effects in syllogistic reasoning (Tessler & Goodman, under review), 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

[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.

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 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. Since L_0 interprets utterances literally, it is never optimal for S_1 to choose an utterance whose literal meaning directly contradicts the intended meaning. For example, suppose S_1 wants to communicate that the weather is terrible. According to the basic RSA model, he reasons about the literal listener L_0 to choose the utterance that will most likely convey this information. Because L_0 is a literal listener, she would interpret the utterance "The weather is amazing" to mean that S_1 believes the weather is literally amazing. She would thus *not* arrive at the interpretation that the speaker believes the weather is terrible. As a result, S_1 has no reason to say "The weather is amazing" to communicate that the weather is terrible (because L_0 would not receive the intended meaning). Consequently, a pragmatic listener who reasons about why the speaker chose various utterances will not interpret "The weather is amazing" to mean that the weather is terrible. The RSA model in its basic form is unable to explain many cases of figurative language use.

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 aspects of background knowledge and prior beliefs. For example, consider the utterance: "Cam is a wolf." To compute the probability that Cam is a wolf given this utterance, the pragmatic listener considers the prior probability of Cam being a wolf. However, believing that Cam is a wolf is more than believing that Cam is a large wild animal that often hunts in groups. Once you believe that Cam is a wolf, you are more likely to believe that Cam is furry, fierce, loyal, fast, hungry, etc. These beliefs are graded; one may have a strong belief that any given wolf is fierce, but only a weak belief that any given wolf is loyal. This network of encyclopedic knowledge forms a rich multi-dimensional representation of what it means to be a wolf. Note that while these other dimensions of

meaning may not be part of the core "literal" meaning of the word "wolf," they are easily accessible through association and are closely tied to the literal meaning. As a result, we assume that the literal listener L_0 also has access to these dimensions of meaning. Given a literal meaning l, associated encyclopedic meanings $\vec{E}$, and an utterance u, the literal listener's interpretation of u is now given by the following:

$$L_0(l, \vec{E}|u) = \begin{cases} P(\vec{E}|l) & \text{if } l \text{ is compatible with the literal meaning of } u \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

We thus provide a formal way of enriching literal meaning with encyclopedic knowledge. However, incorporating encyclopedic knowledge alone is insufficient for explaining figurative language understanding. Although the literal listener now has access to the associated encyclopedic meanings, she still assigns 0 probability to all interpretations that are incompatible with the literal meaning of the utterance. Given the utterance "Bob is a wolf," the literal listener will believe that Bob is a fierce, furry, and loyal wolf with some probability ($P(\vec{E}|l)$); however, she does *not* believe that Bob is a fierce person or any kind of person at all, because she believes that he is a wolf with probability 100%.

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

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

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

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

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

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

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

In what follows, we will describe three domains in which we empirically tested the extended RSA model—termed qRSA—and show that they predict people's interpretation with high accuracy. In particular, we will show that the model captures several desired effects in the interpretation of *hyperbole*, *irony*, and *metaphor*: (1) figurative interpretation (2) sensitivity to encyclopedic knowledge (3) sensitivity to prior beliefs (4) sensitivity to

utterance cost (5) sensitivity to local context (6) sensitivity to alternative utterances.

Modeling Figurative Language

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

In Kao, Wu, Bergen, and Goodman (2014), we examined how people arrive at the appropriate interpretations and affective subtexts of numeric utterances about prices. To empirically measure people's prior beliefs, we asked participants to rate the probabilities that different items (electric kettles, watches, and laptops) cost various amounts of money (e.g. \$50, \$51, \$1,000, \$10,000). To measure people's encyclopedic knowledge in this domain, we asked participants to rate the probability that someone would think an item that costs $\$x$ is expensive (e.g., a watch that costs \$1,000). We chose expensiveness as the associated dimension of interest, because utterances about cost seem to naturally evoke judgments of expensiveness . Using the empirically measured prior beliefs and background knowledge, we used the qRSA model to obtain predicted interpretations for each utterance. The model reasons about different types of QUDs that the speaker may wish to address, including a QUD concerning affect about the price of the item. By reasoning about relevance to a set of QUDs, the model captures a basic feature of hyperbole: utterances whose literal meanings are less likely given the price prior are more likely to be interpreted hyperbolically. For example, "The watch cost 1000 dollars" is more likely to be interpreted

explain this more

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

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

Verbal Irony

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

In what follows, we will examine interpretations of potentially ironic utterances in an innocuous domain—the weather. We chose the weather as the victim of irony for several reasons. First, people are quite familiar with talking (and complaining) about the weather. Second, we can visually represent the weather to participants with minimal linguistic description in order to obtain measures of nonlinguistic contextual knowledge, for example, showing participants a picture of a blue, cloudless sky and asking them to judge how likely

it is that someone would perceive the weather to be amazing or terrible. Finally, given the critical role that context plays in understanding irony, we can vary the context (a picture of a gray, cloudy sky instead of blue sky) to observe how the same utterance is interpreted differently given different contextual knowledge.

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

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

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

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 choice of utterance u given state s, his affect A towards the state, and the QUD is then described by the following: $S_1(u|s, A, Q) \propto e^{\lambda U(u|s,A,Q)}$, where λ is the rationality parameter. A pragmatic listener L_1 takes into account prior knowledge and his internal model of the speaker to determine the

state of the world as well as the speaker's affect, marginalizing over the possible QUDs under consideration:

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

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

We performed the following simulations to examine the model's behavior using affect spaces, A, that differ in complexity and structure. We assume that s has five possible ordered values: *terrible*, *bad*, *neutral*, *good*, and *amazing*. We consider two different weather contexts: apparently bad weather and apparently amazing weather, which are each specified by a prior distribution over these states (see gray dotted lines in Figure 3). 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. This is because a pragmatic listener who only considers emotional valence does not believe that the speaker has any reason to choose a negative utterance to express positive affect (because the utterance communicates no true information). As a result, a pragmatic listener that only considers one dimension of affect–emotional valence—is unlikely to infer a positive world state from a negative utterance (and vice versa), thus failing to evidence verbal irony.

This model simulation reveals a critical puzzle in the interpretation of verbal irony. What true information *could* a speaker communicate about a positive world state using a negative utterance? Affective science identifies two dimensions, termed valence and

arousal, that underly the slew of emotions people experience (Russell, 1980). For example, *anger* is a negative valence and high arousal emotion, while *contentment* is a positive valence and low arousal emotion. Could speakers leverage the arousal dimension to convey high arousal and positive affect (e.g. excitement) using utterances whose literal meanings are associated with high arousal but negative affect (e.g. "The weather is terrible!")? We test the consequences of incorporating a dimension of emotional arousal in the space of affects a listener considers. The orange lines in Figure 3 show simulations of the qRSA model with a two-dimensional affect space: whether the speaker feels negative/positive valence and low/high arousal towards the weather state. Given strong prior belief that the weather state is *bad*, the model interprets "The weather is terrible" to mean that the weather is likely to be *bad*, again producing a hyperbolic interpretation. However, given strong prior belief that the weather is *amazing*, the model now places much greater probability on the ironic interpretation of "The weather is terrible," meaning that the weather is likely *amazing*. This is because, with the enriched two-dimensional affect space, the pragmatic listener realizes that the speaker may be using "terrible" to communicate high emotional arousal. Note that this result is not simply due to the model falling back on the prior: given the same priors, the model interprets the neutral utterance "The weather is ok" as the weather state being *neutral* and not *amazing*. These simulations suggest that a more psychologically realistic, two-dimensional affect space enables the qRSA model to interpret ironic utterances in addition to hyperbolic ones.

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

feels about the weather given what she says in a weather context (e.g. "The weather is terrible!" when the context clearly depicts sunny weather).

Experiment 1a: Background knowledge for verbal irony. We selected nine images from Google Images that depict the weather. To cover a range of weather states, three of the images were of sunny weather, three of cloudy weather, and three of rainy or snowy weather. Each of these images represent what we will call a "weather context," as shown in Figure 4.

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*. Participants also used slider bars (end points labeled "Impossible" and "Absolutely certain") to rate how likely Ann is to feel each of the following seven emotions about the weather: *excited*, *happy*, *content*, *neutral*, *sad*, *disgusted*, and *angry*, which are common emotion categories (Ekman, 1992)[2]. The order of the emotions was randomized for each participant but remained consistent across trials [3].

For each of the nine weather contexts, we obtained the number of participants who gave each of the weather state ratings. We performed add-one Laplace smoothing on the counts to compute a smoothed prior distribution over weather states given each context, namely $P(s)$ (Figure 4). 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

[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 1a: `http://stanford.edu/~justinek/irony_exp/priors/priors.html`

done in studies of emotion ratings (Russell, 1980). We found that the first two principal components corresponded to the dimensions of emotional valence and emotional arousal, accounting for 69.14% and 13.86% of the variance in the data, respectively. As Figure 5 shows, the first two principle components successfully distinguish positively valenced emotions (*excited, happy, content* from negatively valenced emotions (*disgusted, angry, sad)*, as well as high arousal emotions (*excited, disgusted*) from low arousal emotions (*content, neutral, sad*).

The PCA represents emotion ratings for each trial as real values between negative and positive infinity on each of the dimensions. To map these values onto probability space, we first standardized the scores on each dimension to have zero mean and unit variance. We then used the cumulative distribution function to convert the standardized scores into values between 0 and 1. This gives us the probabilities of Ann feeling positive (vs. negative) valence and high (vs. low) arousal for each trial, which is a two-dimensional probabilistic representation of her affect. By calculating the average probabilities of positive valence and high arousal given each weather state rating, we obtain the probability of positive valence and high arousal associated with each weather state, namely $P(A|s)$ (Figure 6). We observe that the probability of positive valence given a weather state increases monotonically across the ordered set of states: *terrible, bad, neutral, good*, and *amazing*, where the probability of positive valence given a *terrible* state is significantly lower than the probability given an *amazing* state. However, the probability of high arousal given each weather state follows a U-shape curve, where the probability of high arousal given a *terrible* state is approximately equivalent to the probability of high arousal given an *amazing* state. In other words, while the valences associated with *terrible* and *amazing* differ significantly, the arousals evoked by these states are very similar.

Experiment 1b: Interpreting verbal irony. We conducted Experiment 1b to elicit people's interpretations of utterances, which we then use to evaluate model predictions. 59 native English speakers with IP addresses in the United States were

recruited on Amazon's Mechanical Turk. Each participant saw all nine images from Figure **??** in random order. In each trial, participants were told that a person (e.g. Ann) and her friend are in a room looking out the window together and see the view depicted by the image. Ann says, "The weather is _____!" where the adjective is randomly selected at each trial from the following set: "terrible," "bad," "ok," "good," and "amazing." Participants first rated how likely it is that Ann's statement is ironic using a slider with end points labeled "Definitely NOT ironic" and "Definitely ironic." They then indicated how Ann would actually rate the weather using a labeled 5-point Likert scale, ranging from *terrible*, *bad*, *neutral*, *good*, to *amazing*. Finally, participants used sliders to rate how likely Ann is to feel each of seven emotions about the weather [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 that participants can identify verbal irony based on its inconsistency with context, how do they then use context to determine the speaker's intended meaning? We examined participants' interpretations of utterances given different contexts. For each of

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

the 45 weather context (9) × 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 utterances, such that the weather is most likely to be *amazing* given that the speaker said "The weather is terrible" in weather context 1. Participants also produce hyperbolic interpretations, such that the weather is most likely to be *bad* given that the speaker said "The weather is terrible" in weather context 7. This confirms the intuition that people are highly sensitive to context and use it both to determine when an utterance is not meant literally and to appropriately recover the intended meaning. Finally, we examine participants' inferences about the speaker's affect given utterances in context. We used the loadings from the PCA on emotion ratings from Experiment 1a to project the emotion ratings from Experiment 1b onto the same dimensions. We then standardized and converted the scores into 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.

[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.

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 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.

Do we need a figure showing correlations for irony ratings? Not as good as straight-up lm with weather context and utterance as predictors

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

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

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

Discussion. We formalized intuitions about verbal irony understanding and clarified the role of shared prior knowledge in ironic interpretations. We explored the consequences of expanding the space of affect considered by RSA to account for verbal irony. By making a minimal extension to Kao, Wu, et al. (2014)'s hyperbole model, we were able to capture people's fine-grained interpretations of ironic utterances in addition to hyperbole. This provides evidence that hyperbole and irony may operate using similar underlying principles of communication, namely reasoning about shared background knowledge as well as the speaker's affective goals.

There remain important qualities of verbal irony to account for. For example, speakers often use verbal irony to remind the listener of previous utterances that turned out to be false, or of positive norms that were violated (Sperber & Wilson, 1981; ?, ?). On the other hand, pretense theory argues that when a speaker produces an ironic utterance, she is only pretending to be someone who would make such an utterance (Clark & Gerrig, 1984). While our model is able to capture the main characteristics of verbal irony, it does not account for the intuitions behind echoic mention or pretense theories. We hope to enrich our model's understanding of the social aspects of irony by addressing these intuitions in future research. In addition, we aim to further examine how people identify the particular dimensions of meaning that may be under discussion in a given context. For

example, affective dimensions such as valence and arousal may be particularly relevant in domains that involve evaluation (e.g. "good" or "terrible" weather), while non-affective dimensions may be more salient in other domains (Kao, Bergen, & Goodman, 2014).

Metaphor

Todo: Better segue into metaphor that highlights the differences

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

In the work described above, we assumed that the set of QUDs under consideration included the speaker's affect, which is supported by previous research on the rhetorical effect of hyperbole and verbal irony. In what follows, we will explore ways to systematically elicit the set of QUDs that listeners consider, as well as to manipulate prior probabilities over QUDs using discourse context. We show that considering these additional, non-affective QUDs allows the model to capture interesting effects of metaphor interpretation. In particular, we will focus on three aspects of the pragmatics of metaphor understanding that the qRSA model naturally captures. First, interpretation of the same metaphor differs systematically given different discourse contexts, which can be modeled as different prior probabilities over QUDs (Experiment 2). Second, metaphors are able to communicate information efficiently along several dimensions and address multiple QUDs at once, which may serve as an advantage over literal statements (Experiment 2). Finally, the specific interpretations of a metaphor are sensitive to the alternative utterances that a speaker could have chosen to address the QUDs under consideration (Experiment 3).

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

that the speaker chooses an utterance to maximize informativeness about a subject along dimensions that are relevant to the QUD. Unlike hyperbole and irony, however, these dimensions are not affective in nature. Rather, they are features associated with the metaphorical source, in this case "shark."

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

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

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

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

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

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

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

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

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

In our first exploration of the model's behavior, we made a number of simplifying assumptions. First, we restricted the number of possible categories to which a member may belong to c_a and c_p, denoting an animal category (in this case *shark*) or a person category, respectively. We also restricted the possible features of Cam under consideration to a vector of size three: $\vec{f} = [f_1, f_2, f_3]$, where f_i is either 0 or 1. Finally, we assumed a small and rather impoverished set of alternative utterances that the speaker could have said: the utterance she did say (e.g. "Cam is a shark"), and a grammatically similar and literally true utterance (e.g. "Cam is a person.").[7]

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

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

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

(3) $P(g)$: the probability a speaker has goal g. This prior can change based on the conversational context that a question sets up. For example, if the speaker is

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

responding to a vague question about Cam, e.g. "What is Cam like?", the prior over goals is uniform. If the question targets a specific features, such as "Is Cam scary?", then she is much more likely to have the goal of communicating John's scariness. However, she may still want to communicate other features about Cam that were not asked about. We assume that when the question is specific, the prior probability that S_n's goal is to answer the specific question is greater than 0.5, fitting the value to data below.

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

Experiment 2a: Feature Elicitation. We selected 32 common non-human animal categories from an online resource for learning English (`www.englishclub.com`). The full list is shown in Table 1. 100 native English speakers with IP addresses in the United States were recruited on Amazon's Mechanical Turk. Each participant read 32 animal category names presented in random order, e.g. "whale", "ant", "sheep". For each animal category, participants were asked to type the first adjective that came to mind in a text box. Using participants' responses, we constructed a list of adjectives for each animal category and ordered them by the number of times they were given by a different participant (i.e. their popularity). We removed all color adjectives, such as "white" and "black," to eliminate the possibility of interpreting these adjectives as racial descriptions. To avoid redundancy in the feature set, we used WordNet (Miller, 1995) to identify synonymous adjectives and only kept the most popular adjective among a set of synonyms. We then took the three most popular adjectives for each animal category and used them as the set of features. In what follows, f_1 is the most popular adjective, f_2 the second, and f_3 the third. Table 1 shows the animal categories and their respective features.

Experiment 2b: Feature Prior Elicitation. Using the features collected from Experiment 1a, we elicit the prior probability of a feature vector given an animal or person category (i.e. $P(\vec{f}|c)$). We assume that the adjective corresponding to a feature (e.g. *scary*) indicates that the value of that feature is 1 (present), while the adjective's antonym indicates that the value of that feature is 0 (not present). We used WordNet to construct antonyms for each of the adjective features produced in Experiment 1a. When multiple antonyms existed or when no antonym could be found on WordNet, the first author used her judgment to choose the appropriate antonym. Table 1 shows the resulting list of antonyms. For each animal category, eight possible feature combinations were constructed from the three features and their antonyms. For example, the possible feature combinations for a member of the category "ant" are {small, strong, busy}, {small, strong, idle}, {small, weak, busy}, and so on.

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

We normalized each participant's ratings for the eight feature combinations in a trial to sum up to 1 based on the assumption that the feature combinations exhaustively describe a member of a particular category. Using the Spearman-Brown prediction equation, reliability of the ratings was 0.941 (95% CI = $[0.9408, 0.9414]$). Averaging across participants' normalized ratings, we obtained feature priors $P(\vec{f}|c)$ for $c = c_a$ (animal) and $c = c_p$ (person). Since the features were created using the animal categories in Experiment 1a, by construction features are rated as significantly more likely to be present in the

animal category than in the person category ($F(1, 190) = 207.1$, $p < 0.0001$). These results confirm that participants are fairly confident that each animal category has certain distinguishing features (mean= 0.61, sd= 0.06), while those same features are rated as appearing in people less often (mean= 0.48, sd= 0.06).

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

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

For each condition of each scenario, we obtained the average probability ratings for the three features. Figure 10 shows the average ratings for each feature across animal categories given a vague or specific goal and a literal or metaphorical utterance. When the speaker gives a literal statement directly affirming the presence of f_1, participants rate f_1

as significantly more likely than when the speaker gives a metaphorical statement ($F(1, 126) = 52.6$, $p < 0.00001$). However, participants rate f_2 and f_3 as significantly more likely when the speaker produces a metaphorical utterance than when the utterance is literal ($F(1, 126) = 23.7$, $p < 0.0001$; $F(1, 126) = 13.66$, $p < 0.0005$). Comparing feature probability ratings in Experiment 2 to the feature priors obtained in Experiment 1b, we can measure how literal and metaphorical utterances change listeners' inferences about a person's features. Given a literal utterance that directly confirms the existence of f_1, probability ratings for f_1 are significantly higher than the prior probabilities of f_1 for a person ($t(63) = 59.19, p < 0.00001$). However, probability ratings for f_2 and f_3 are not significantly different from their prior probabilities ($t(63) = -0.13, p = 0.89$; $t(63) = 0.03, p = 0.97$). Given a metaphorical utterance, probability ratings for all three features are significantly higher than the prior probabilities ($t(63) = 15.74, p < 0.0001$; $t(63) = 7.29, p < 0.0001$; $t(63) = 5.91, p < 0.0001$). This analysis suggests that metaphorical utterances may convey richer information and update listeners' beliefs along more dimensions than literal utterances.

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

Metaphor model evaluation. We used the feature priors obtained in Experiment 2b to compute model interpretations of the 32 metaphors. As discussed in the previous section, the behavioral results in Experiment 2c show evidence that the context set up by a question changes participants' interpretation of a metaphor. Our model naturally accounts for this using the speaker's prior over communicative goals $P(g)$. When a speaker is responding to a vague question, we set the prior distribution for $P(g)$ as uniform. When the speaker is responding to a question specifically about f_1, we assume that $P(g_1) > 0.5$ and equal between $P(g_2) = P(g_3)$. Fitting the goal prior parameter to data yields a prior of $P(g_1) = 0.6$ when responding to a specific question about f_1. We fit the category prior $P(c_a) = 0.01$ and the speaker optimality parameter $\lambda = 3$.

Using these parameters, we obtained interpretation probabilities for each of the 32 metaphors under both vague and specific goal conditions. For each metaphor and goal condition, the model produces a joint posterior distribution $P(c, \vec{f}|u)$. We first show a basic but important qualitative result, which is that the model is able to interpret utterances metaphorically. Marginalized over values of $\vec{f}$, the probability of the person category given the utterance is close to one ($P(c_p|u) = 0.994$), indicating that the pragmatic listener successfully infers that the person described as an animal is actually a person and not an animal. This shows that the model is able to combine prior knowledge and reason about the speaker's communicative goal to arrive at nonliteral interpretations of utterances.

We now turn to the second component of the interpretation, $P(\vec{f}|u)$. To quantitatively evaluate the model's performance, we correlated model predictions with human interpretations of the metaphorical utterances. Given a metaphorical utterance and a vague or specific goal condition, we computed the model's marginal posterior probabilities for f_1, f_2, and f_3. We then correlate these posterior probabilities with participants' probability ratings from Experiment 2c. Figure 11 plots model interpretations for all metaphors, features, and goal conditions against human judgments. Correlation across the 192 items (32 metaphors $\times$ 3 features $\times$ 2 goal conditions) is 0.6 ($p < 0.001$). The

predicted reliability of participants' ratings using the Spearman-Brown prediction formula is 0.828 (95% CI = $[0.827, 0.829]$), suggesting first that people do not agree perfectly on metaphorical interpretations, and second that our model captures a significant amount of the reliable variance in the behavioral data. In particular, our model does especially well at predicting participants' judgments of f_1, which are the most salient features of the animal categories and were targeted by specific questions in Experiment 2. Correlation between model predictions and human judgments for f_1 is 0.7 ($p < 0.0001$), while the predicted reliability of participants' ratings for f_1 is 0.82 (95% CI = $[0.818, 0.823]$).

We now compare our model's performance to a baseline model that also considers the feature priors and the conversational context. We constructed a linear regression model that takes the marginal feature priors for the animal category, the marginal feature priors for the person category, and the vague or specific goal as predictors of participants' ratings. With four parameters, this model produced a fit of $r = 0.45$, which is significantly worse than our model ($p < 0.0001$ on a Cox test). This suggests that our computational model adequately combines people's prior knowledge as well as principles of pragmatics to produce metaphorical interpretations that closely fit behavioral data.

While our model predictions provide a reasonable fit to behavioral data and outperforms a linear regression model using fewer parameters, we observed residual variance that can be further addressed. Previous work has shown that alternative utterances—what the speaker could have said—can strongly affect listeners' interpretation of what the speaker *did* say (Bergen et al., 2012). In this experiment, we did not take into account the range of alternative utterances (both literal and metaphorical) that a listener considers when interpreting a speaker's utterance. We posit that other plausible alternative utterances may account for some of the variance in the data that our model does not capture. Consider the metaphor "He is an ant" and the corresponding features *small, strong*, and *busy*. Our model currently assigns a high probability to the feature *strong* given the metaphor, while participants assign it a lower probability. Indeed, this data point

has the highest residual in our model fit. To test the idea that alternative utterances may account for this discrepancy, we construct a model that has "He is an ox" as an alternative utterance. "Ox" has features that roughly align with the features of "ant": *strong, big*, and *slow*. Since *strong* is a higher probability feature for "ox" than for "ant," the listener reasons that if the speaker had intended to communicate the feature *strong*, she would have said "He is an ox" since it optimally satisfies that goal. Since the speaker did *not* produce the utterance "He is an ox," the listener infers that *strong* is a less probable feature. Adding this alternative utterance to the model indeed lowers the marginal posterior probability of *strong* given the utterance "He is an ant." Based on this simple error analysis, we posit that adding alternative utterances across all animal categories may significantly improve our model's performance.

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

Experiment 3a: Feature Elicitation. We selected 12 common animals that have various distinguishing features: *ant, whale, bird, elephant, panda, monkey, penguin, giraffe, cheetah, turtle, lion*, and *rabbit*, which we will call the core animals. 50 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk. Each participant read 12 animal category names presented in random order. Following the same paradigm as Experiment 2a, participants were asked to type the first adjective that came to mind in a text box. We removed color adjectives and combined synonymous adjectives by consulting WordNet (Miller, 1995). We then identified the two most popular adjectives

for each animal category and used those as features [8].

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

Experiment 3a yielded 15 unique features elicited from the 12 core animals: 2 features for each core animal, with 9 overlapping features across the 12 animals such as "small," "big," and "strong." 50 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk to produce animals associated with each of the 15 features. Each participant read the 15 feature adjectives presented in random order, with the prompt: "Write down the first animal you can think of that is _____," where _____ is a feature adjective. We tallied the number of times an animal was produced for each of the features and identified the animals most strongly associated with each feature.

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

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

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

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

Experiment 3d: Metaphor Interpretation. 45 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk.

Describe results.

Model evaluation.

Model comparison with and without alternative animals.

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

Hyperbolic Metaphor

Metaphors often elicit hyperbolic interpretations and result in more extreme beliefs about the world. In this section, we examine people's interpretations of metaphors such as "Cam is a giraffe," where the salient feature being communicated (e.g. height) lies on a continuous scale. We find that the interpretation of these types of metaphors demonstrates two effects: First, given the utterance "Cam is a giraffe," listeners believe that Cam is taller than they do when given the literal description "Cam is tall." Second, listeners believe that Cam is taller than the average person, but much shorter than the average

giraffe. We show that the qRSA model naturally captures these effects.

Model.

Experiment 4a: Prior Elicitation.

Experiment 4b: Metaphor Interpretation.

Model Evaluation.

Discussion.

General Discussion

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

We believe that the qRSA model makes several critical advancements to formal models of human language understanding. The model captures key intuitions about communication, including the role of common ground between listener and speaker, the assumption that speakers produce utterances that maximize informativeness, and the idea that informativeness should be considered with respect to the question under discussion and the speakers' communicative goals. By formalizing these intuitions, the model is able to go beyond the literal meanings of utterances and predict subtleties in interpretation that are sensitive to background knowledge, communicative efficiency, local context, and

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

In addition to advancing general models of language understanding, our work may provide explanations for phenomena that particularly interest researchers of figurative language. One distinctive aspect of figurative utterances is that they communicate multiple dimensions of meaning at once and are often difficult to paraphrase . By considering cite encyclopedic knowledge such as affects associated with different states and features associated with different categories, our model naturally accounts for the multi-dimensional quality of figurative meaning. In addition, many researchers have observed that the interpretation of figurative language is especially sensitive to common ground (Pexman & Zvaigzne, 2004; Gibbs, 2000). Given that the literal meanings of figurative utterances are often implausible or false, listeners rely on background knowledge in order to reason about and recover speakers' intended meanings. As a result, it is important for listeners and speakers to share background knowledge (both encyclopedic knowledge and specific prior

beliefs) in order to communicate successfully using figurative utterances. Indeed, researchers have found that willingness to use ironic utterances is positively correlated with social intimacy between interlocutors (Kreuz, 1996), and interlocutors who use metaphorical speech with each other are perceived as having a closer relationship(Horton, 2007). Kreuz (1996) attributes the effects of social closeness to what he calls a principle of inferability: speakers are more likely to use a non-literal utterance when they are more certain that it will be understood appropriately, which in turn is more likely when the speakers and listeners share similar background knowledge and beliefs. For example, suppose Ann saw a watch that cost $1000 and wants to tell Bob that she thinks it is very expensive. In order to communicate her meaning effectively using a hyperbolic utterance: "That watch cost a million dollars!", Ann must be fairly confident that Bob's prior beliefs would not lead him to believe that the watch literally cost a million dollars. And since Bob reasons about Ann's motivation for choosing an utterance, given such a hyperbolic utterance, Bob arrives at the inference that Ann must be fairly confident about his prior beliefs. Our model provides a natural way to incorporate inferences about common ground in order to predict and model effects of social closeness. In its current form, our model assumes that the listener is certain that the speaker has perfect knowledge of the listener's background knowledge and prior beliefs. If we relax this assumption to incorporate uncertainty about the speaker's knowledge of the listener, the pragmatic listener is able to derive information about the speaker's knowledge of the listener. Preliminary explorations of our model show that it naturally affords these types of social inferences, which we plan to examine more thoroughly in future work.

One of the most important contributions of our work is that it unifies the interpretation of diverse types of figurative language in a single computational model. The generality of this model suggests that separate processing mechanisms may not be necessary to derive different types of figurative interpretations, or to derive figurative interpretations at all. The qRSA model does not distinguish *a priori* between figurative

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

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

Future directions

The work we described invites many directions for future research, both for computational models of pragmatics and figurative language understanding. While our model explicitly reasons over a set of QUDs and alternative utterances, we have not precisely defined how listeners select the particular set of QUDs and alternatives over which to reason. For example, when should listeners consider speakers' affects and subjective attitudes to be potential QUDs? How do listeners decide that certain features of the metaphorical source are likely QUDs (e.g. *scary* in the case of "Cam is a shark"), while others are not (e.g. *has teeth* or *swims*)? In the work described in this paper, we made the simplifying assumption that QUDs arise from the literal meanings of utterances in an associative manner, and relied on intuition and previous research to focus on affective

QUDs in the case of hyperbole and irony and feature QUDs in the case of metaphor. We then defined a set of QUDs based on the affects and features associated with various states and categories. Future research should examine whether QUDs can be more systematically inferred from prior context or the utterance itself, thus removing the need to predefine a set of QUDs that the model should consider for a given type of figurative utterance. A more flexible and general way to define a set of QUDs could also enable us to determine which types of QUDs are most effectively addressed by different kinds of figurative utterances.

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

nd (4) may
ut to be
me thing.

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

While our model provides a unified explanation for three diverse uses language: hyperbole, irony, and metaphor, future work should examine whether the model extends to

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

idioms?

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

politeness?

Conclusion

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

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. (2015). The strategic use of noise in pragmatic reasoning. *Topics in Cognitive Science*, *7*(2), 336–350.

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

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. *Journal of Experimental Psychology: General.*

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

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

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

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

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

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

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

169–200.

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.

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

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

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

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

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

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

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

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

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

Giora, R. (2002). Literal vs. figurative language: Different or equal? *Journal of Pragmatics*, *34*(4), 487–506.

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

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

Glucksberg, S. (2003). The psycholinguistics of metaphor. *Trends in Cognitive Sciences*,

7(2), 92–96.

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

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

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*.

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

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

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

Katz, A. N., & 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. (1996). The use of verbal irony: Cues and constraints. *Metaphor:*

Implications and Applications, 23–38.

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

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

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

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

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

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

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

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

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

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

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). There's millions of them: hyperbole in everyday

conversation. *Journal of pragmatics*, *36*(2), 149–184.

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

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

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.

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

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

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.

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

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

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

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

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

pragmatics, *49*.

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

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

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

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

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. (under review). Some arguments are probably valid: Syllogistic reasoning as communication. *under review*.

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

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

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).*

Animal	$f_1 = 1$	$f_2 = 1$	$f_3 = 1$	$f_1 = 0$	$f_2 = 0$	$f_3 = 0$
ant	small	strong	busy	large	weak	idle
bat	scary	blind	nocturnal	unalarming	sighted	diurnal
bear	scary	big	fierce	unalarming	small	nonviolent
bee	busy	small	angry	idle	large	unangry
bird	free	graceful	small	unfree	awkward	large
buffalo	big	strong	wild	small	weak	tame
cat	independent	lazy	soft	dependent	fast	hard
cow	fat	dumb	lazy	thin	smart	fast
dog	loyal	friendly	happy	disloyal	unfriendly	unhappy
dolphin	smart	friendly	playful	stupid	unfriendly	unplayful
duck	loud	cute	quacking	quiet	unattractive	non-quacking
elephant	huge	smart	heavy	small	stupid	light
fish	scaly	wet	smelly	smooth	dry	fragrant
fox	sly	smart	pretty	artless	stupid	ugly
frog	slimy	noisy	jumpy	nonslippery	quiet	relaxed
goat	funny	hungry	loud	humorless	full	quiet
goose	loud	mean	annoying	quiet	nice	agreeable
horse	fast	strong	beautiful	slow	weak	ugly
kangaroo	jumpy	bouncy	cute	relaxed	inelastic	unattractive
lion	ferocious	scary	strong	nonviolent	unalarming	weak
monkey	funny	smart	playful	humorless	stupid	unplayful
owl	wise	quiet	nocturnal	foolish	loud	diurnal
ox	strong	big	slow	weak	small	fast
penguin	cold	cute	funny	hot	unattractive	humorless
pig	dirty	fat	smelly	clean	thin	fragrant
rabbit	fast	furry	cute	slow	hairless	unattractive
shark	scary	dangerous	mean	unalarming	safe	nice
sheep	wooly	fluffy	dumb	hairless	hard	smart
tiger	striped	fierce	scary	unpatterned	nonviolent	unalarming
whale	large	graceful	majestic	small	awkward	inferior
wolf	scary	mean	angry	unalarming	nice	unangry
zebra	striped	exotic	fast	unpatterned	native	slow

Table 2

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

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

Table 3

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

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

Table 4

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

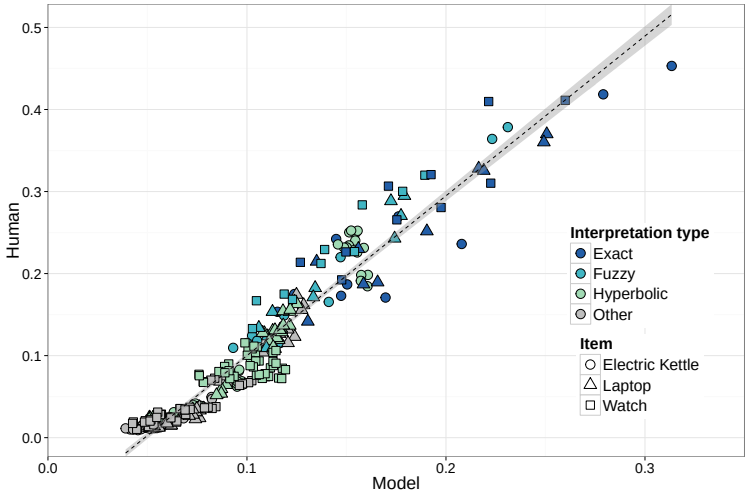

Figure 1. Model predictions v.s. average human interpretations. Each point represents an utterance and price state pair (u, s). The x-coordinate of each point is the probability of the model interpreting utterance u as meaning price state s; the y-coordinate is the empirical probability. Correlation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

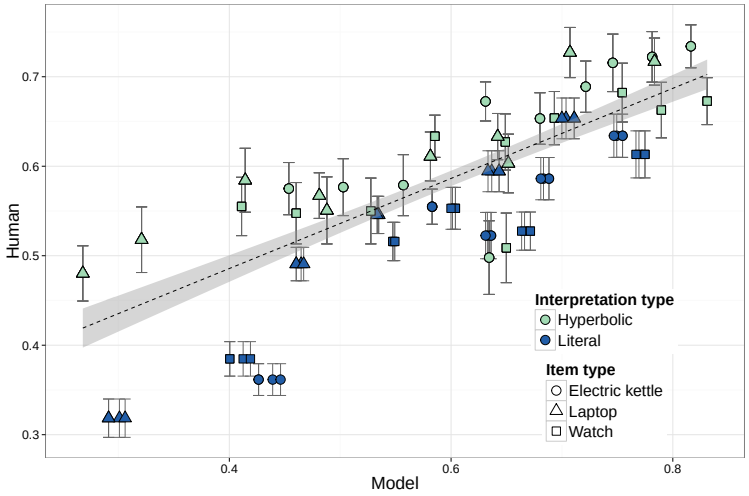

Figure 2. Model predictions of affect v.s. human ratings. Each point represents an utterance and price state pair (u, s). For pairs where $u = s$, the utterance is literal; for $u > s$, the utterance is hyperbolic. The x-coordinate of each point is the model's prediction of the probability that the utterance/price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

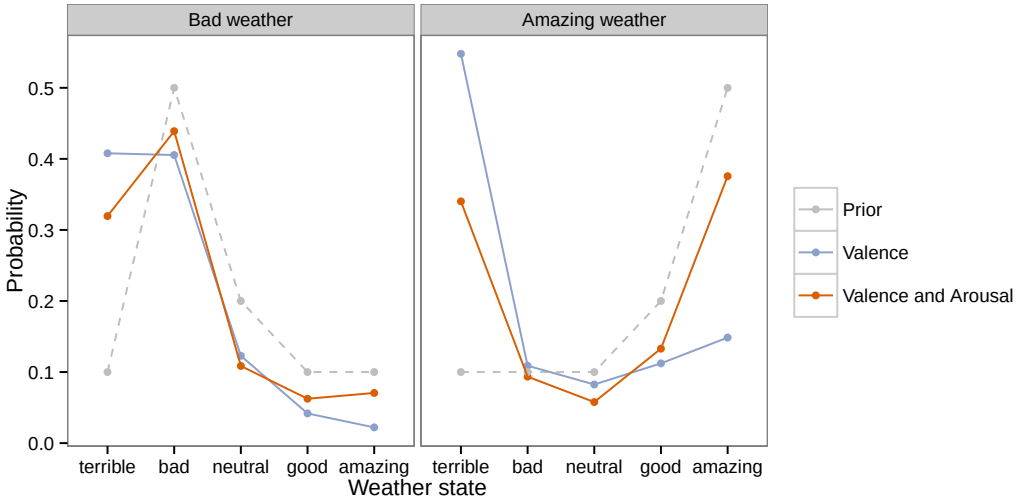

Figure 3. Model interpretations of "The weather is terrible" given different prior beliefs about the weather state and affect dimensions. Gray dotted lines indicate prior beliefs about weather states given a weather context; blue lines indicate interpretations when reasoning only about the speaker's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

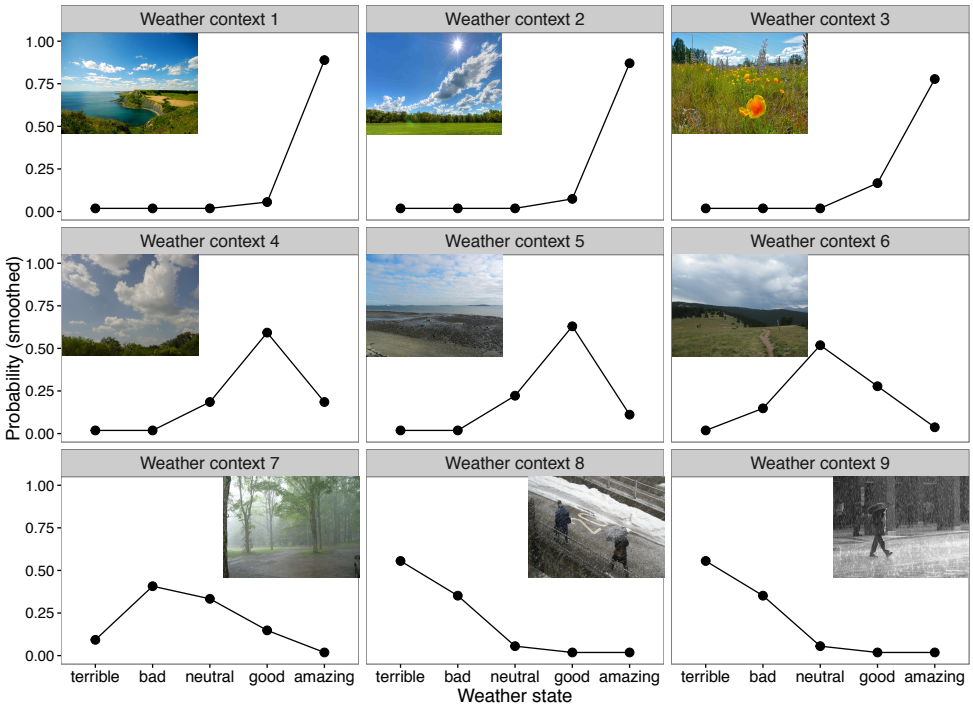

Figure 4. Smoothed prior probability distributions over weather states for each of the nine weather contexts. Participants saw each image and chose a state label from the set: *terrible, bad, neutral, good, amazing*. Probability distributions over weather states were computed by performing Laplace smoothing on the counts for each state label given a weather context and normalizing the counts to sum up to 1.

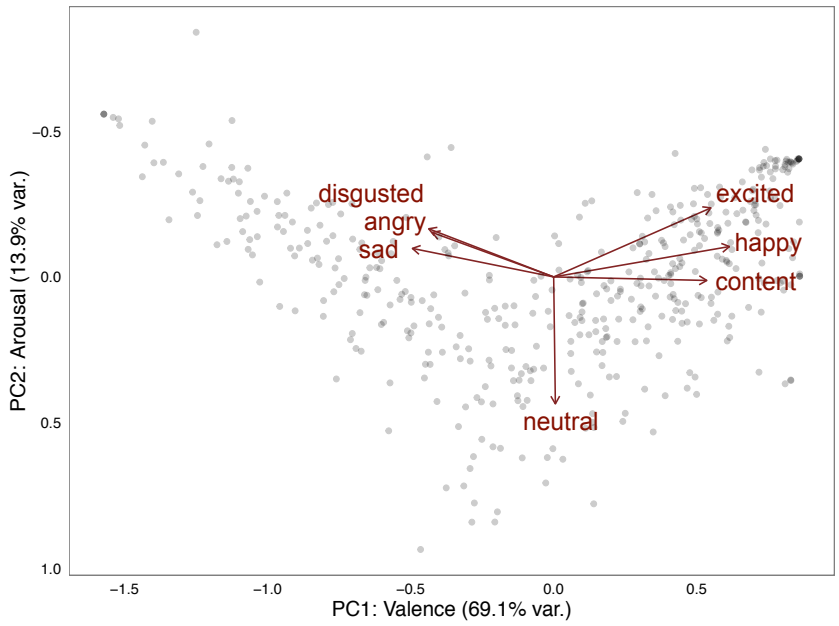

Figure 5. Biplot of the first two principle components of the seven emotion ratings. The first two PCs correspond roughly to valence and arousal, with positively valenced emotions (*excited, happy, content*)clustering on the right, and more high arousal emotions (*disgusted, excited*) appearing appearing at the top.

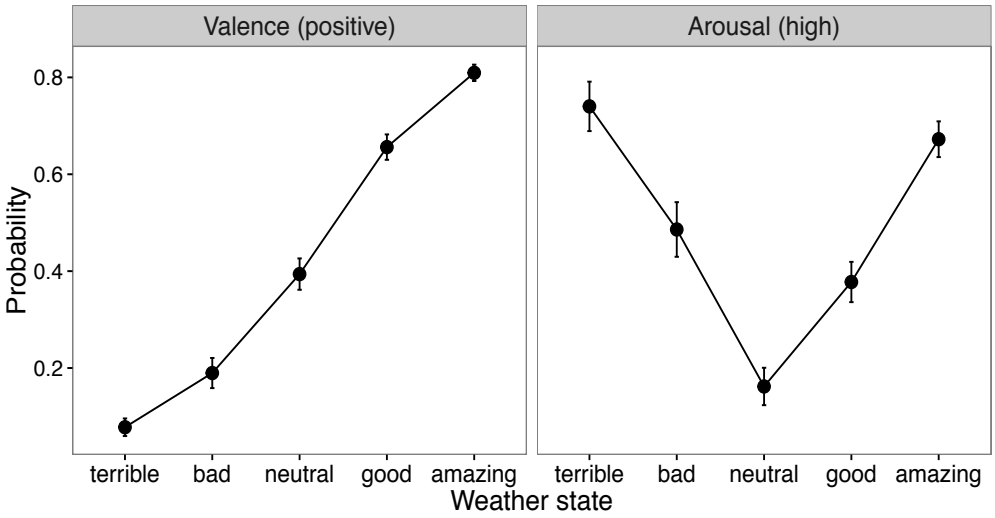

Figure 6. Average probabilities of positive valence and high arousal given each weather state. Error bars are 95% confidence intervals. Probability of positive valence increases monotonically over the five weather states; probability of high arousal follows a symmetric U-shaped curve and does not differ significantly for the *terrible* and *amazing* weather states.

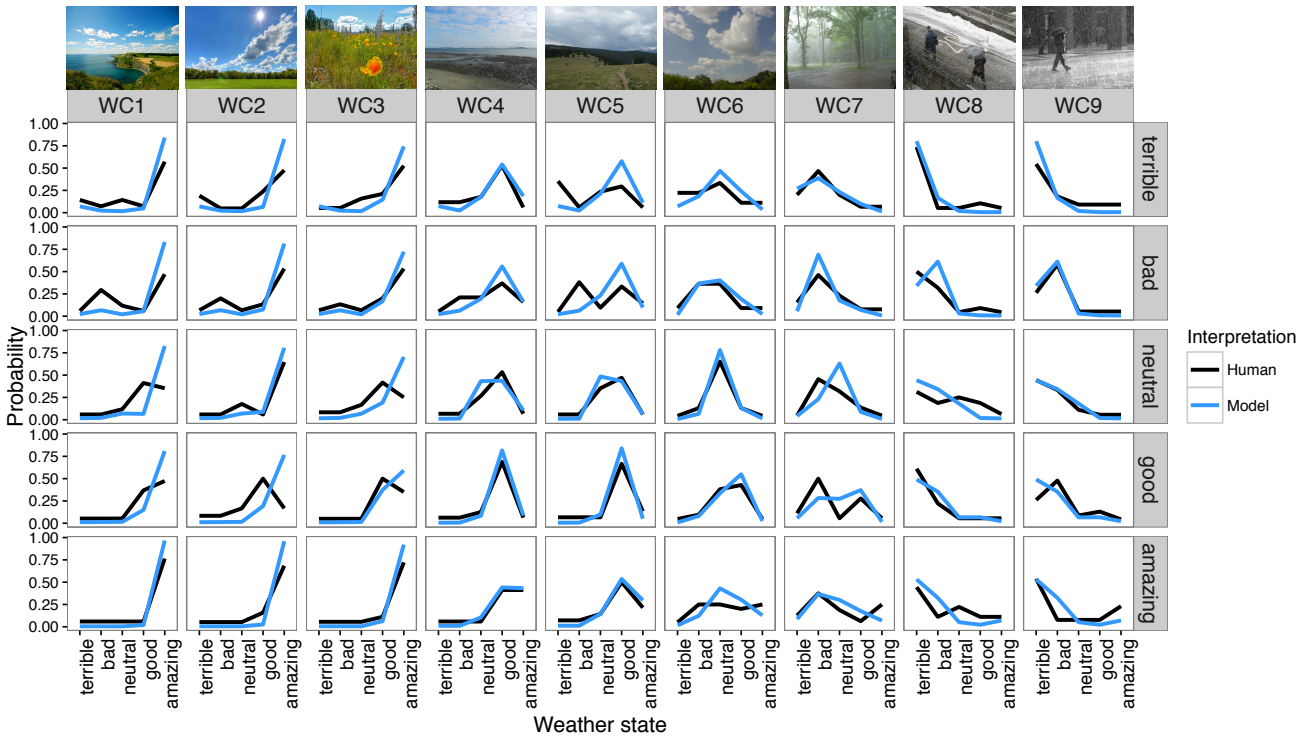

Figure 7. Model's and participants' inferences about the weather state (x-axis) given a weather context (column) and an utterance (row). Each panel represents an interpretation given an utterance in a weather context. The dark lines are participants' ratings; the light lines are the model's posterior distributions over weather states.

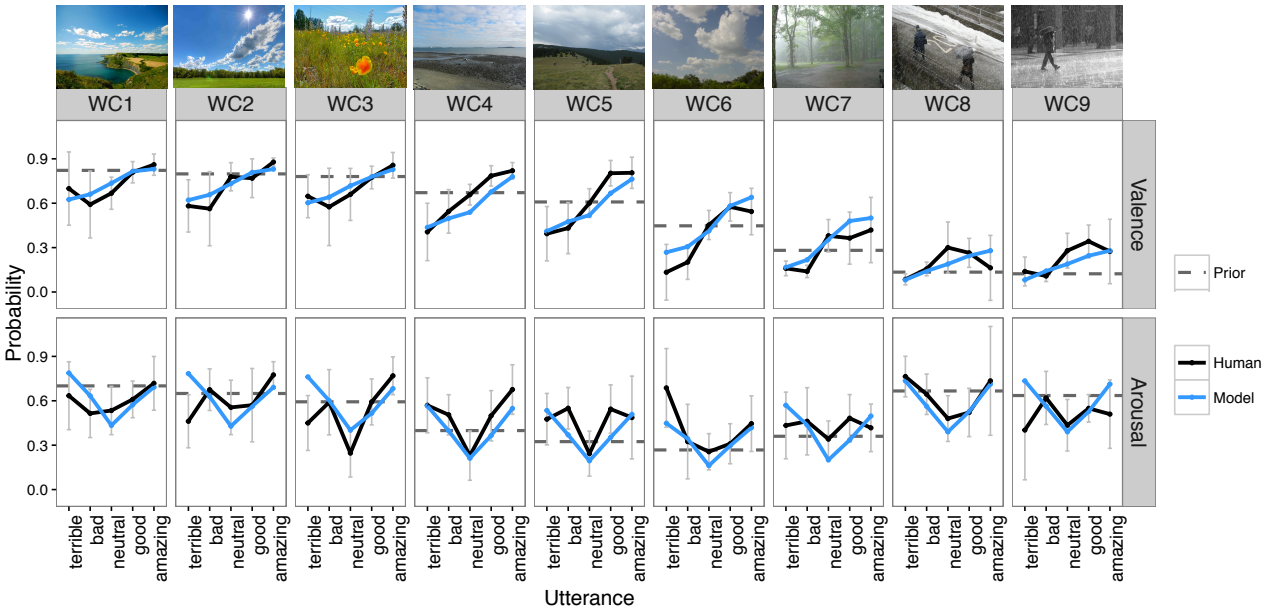

Figure 8. Model's and participants' inferences about the probability of valence and arousal (row) given a weather context (column) and an utterance (x-axis). The dark lines are participants' ratings; the light lines are the model's posterior probabilities of positive valence and high arousal given an utterance in a weather context. The dotted lines are prior probabilities of positive valence and high arousal for each weather context. Error bars are 95% confidence intervals on the participants' ratings.

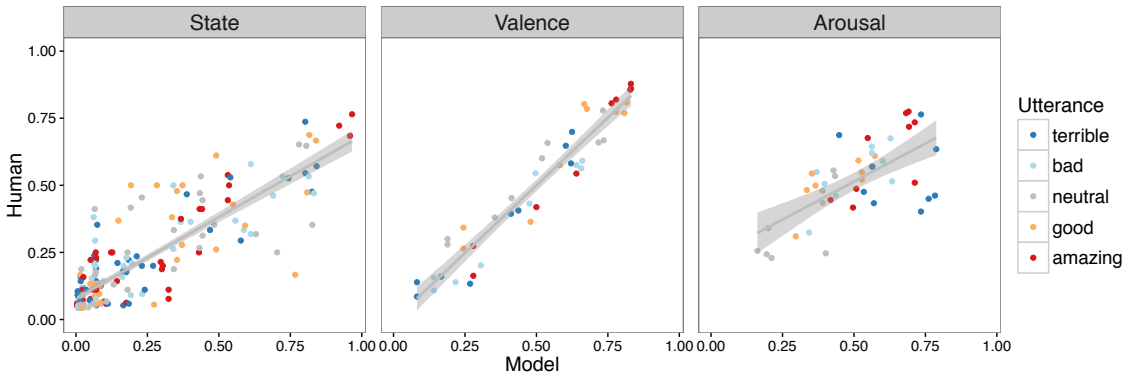

Figure 9. Scatter plot showing correlations between model predictions and human ratings for weather state, speaker valence, and speaker affect. Colors indicate utterances.

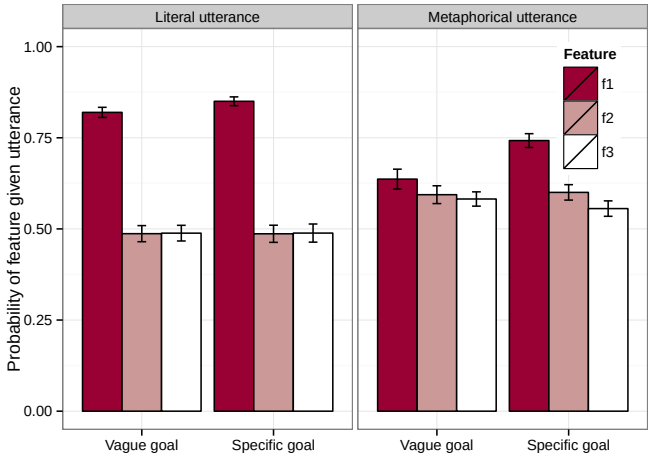

Figure 10. Average probability ratings for the three features given a vague/specific goal and a literal/metaphorical utterance. Error bars are standard error over the 32 items.

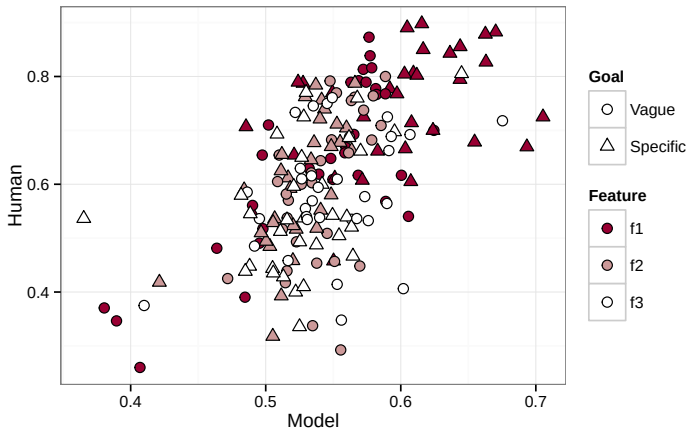

Figure 11. Model predictions (x axis) vs participants' probability ratings (y axis) for 192 items (32 metaphors $\times$ 3 features $\times$ 2 goal conditions). Shape of points indicates goal condition and color indicates feature number.