Let's talk (ironically) about the weather: A computational model of verbal irony

Justine T. Kao (justinek@stanford.edu)
Department of Psychology, 450 Serra Mall
Stanford, CA 94305 USA

Noah D. Goodman (ngoodman@stanford.edu)
Department of Psychology, 450 Serra Mall
Stanford, CA 94305 USA

Abstract

Verbal irony plays an important role in how we communicate and express our opinions about the world. While there exist many interesting theories and empirical findings about how people use and understand verbal irony, there is to our knowledge no formal model of how people incorporate shared background knowledge and linguistic information to communicate ironically. Here we present two behavioral experiments that examine people's interpretations of utterances given different contexts in the weather domain. We then describe a computational model that reasons about background knowledge, affect, and the speaker's communicate goals to interpret ironic utterances and their rich affective subtexts. We show that by accounting for two types of affect goals—valence and arousal—our model produces interpretations that closely match humans'. Finally, we discuss the implications of our model on irony and its relationship to other types of nonliteral language understanding.

Keywords: irony; computational modeling; pragmatics; nonliteral language understanding

Introduction

For better or for worse, verbal irony—roughly defined as utterances whose apparent meanings are opposite to the speaker's intended meaning (?, ?, ?)—is a major figurative trope of our time. From popular sitcoms to political satire to *#sarcasm* on Twitter and casual conversations among friends, verbal irony plays an important role in how we communicate and express opinions about the world. The prevalence of verbal irony poses a puzzle for theories of language understanding: Why would speakers ever use an utterance to communicate its opposite meaning, and how can listeners appropriately interpret such an utterance? Previous work has shown that verbal irony serves several important communicative goals, such as to heighten or soften criticism (?, ?), elicit emotional reactions (?, ?), highlight group membership (?, ?), and express affective attitudes (?, ?, ?). These findings suggest that while ironic statements are false under their literal meanings, they are often highly informative with respect to social and affective meanings. In this paper, we present a computational model and behavioral experiments to show that people may use inferences about these alternative dimensions of meaning and the speaker's communicative goals to understand ironic utterances.

Linguists and psychologists have proposed several informal theories of how people understand verbal irony. According to a classic Grician analysis, listeners first need to recognize that an ironic utterance blatantly violates the maxim of quality (to be truthful); they then arrive at a conversational implicature that the intended meaning is contrary to the utterance's literal meaning (?, ?, ?). While Grice's account is appealing in its treatment of verbal irony as arising naturally from conversational maxims, it does not provide a detailed or satisfactory explanation for how the appropriate implicature is derived from these maxims (?, ?). On the other hand, previous work suggests that it is possible and indeed desirable to consider figurative language understanding as a product of general principals of communication, such as reasoning about informativeness with respect to the speaker's communicative goals (?, ?, ?). In particular, a model that reasons about the speaker's affective goals is able to interpret hyperbolic utterances and infer the appropriate affective subtext (?, ?). Given the fact that many researchers believe hyperbole and irony to be closely related phenomena (cite), we suggest that similar principles may account for verbal irony understanding as well. Our goal in this paper is to identify these communicative principles and provide a precise formal account of how they interact to produce ironic interpretations.

Rational Speech Act (RSA) models are a family of computational models that formalize language understanding as recursive reasoning between speaker and listener, and have been shown to account for many phenomena in pragmatics (?, ?, ?). (?, ?) introduces a critical extension to basic RSA models by considering the idea that speakers may have different communicative goals when formulating an utterance, and that an important task for the listener is to jointly infer these goals as well as the intended meaning. For example, a speaker may want to communicate negative affect about a situation (e.g. unhappiness about the temperature outside) instead of the precise situation (e.g. the precise temperature outside), in which case choosing an exaggerated utterance (e.g. "It's freezing outside!") satisfies her goal of communicating negative affect. A listener who reasons about the speaker and her goals as well as background knowledge about temperatures is then able to correctly infer that the speaker is upset about the temperature, but that it is unlikely to be literally freezing outside (especially if she is in California). Such an extended model—which we will refer to as tRSA—is able to produce nonliteral interpretations of hyperbolic utterances that closely match humans'. However, (?, ?) considered only a fairly impoverished space of affects, namely the presence or absence of negative feeling. This overlooks the range of attitudes and emotions that speakers could express with nonliteral utterances. In particular, since verbal irony involves expressing

negative meanings with positive utterances and vice versa, a richer space of affect that includes both positive and negative emotions is necessary. Here we examine the consequence of extending the tRSA model to consider a wider range of emotions in an empirically derived two-dimensional affect space, and show that this minimal extension is able to capture verbal irony.

In what follows, we will examine interpretations of potentially ironic utterances in an innocuous domain—the weather. We chose the weather as the victim of irony for several reasons. First, people are quite familiar with talking (and complaining) about the weather. Second, we can visually represent the weather to participants with minimal linguistic description in order to obtain measures of nonlinguistic contextual knowledge. Finally, we can vary the weather states to observe how the same utterance is interpreted differently given different contextual knowledges. We first describe an extension to the tRSA model and show that an enriched space of affect enables the model to produce ironic interpretations. 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 affect goals, valence and arousal, our model produces interpretations that closely match humans'. Finally, we discuss the implications of our model on irony and its relationship to other types of nonliteral language understanding.

Computational Model

In RSA models, a speaker chooses an utterance that most effectively communicates her intended meaning to a literal listener; a pragmatic listener then reasons about the speaker to infer a pragmatically enriched meaning given the utterance (?, ?, ?). To model nonliteral language understanding, ? (?) extended the basic RSA model to incorporate an affective dimension of meaning and multiple communicative goals, opening up the possibility for a speaker to produce an utterance that is literally false but satisfies her goal to convey affect.

Here we further extend the interpreted meaning of an utterance to have the following dimensions: the weather state s, where $s \in \{$*terrible, bad, neutral, good, amazing*$\}$; the speaker's emotional valence v towards the weather state, where $v \in \{$*positive, negative*$\}$, and her emotional arousal a towards the weather state, where $a \in \{$*high, low*$\}$. We consider the interpreted meanings of utterances u in the following set: {"terrible," "bad," "ok," "good," "amazing"}.

JTK: But do we refer to these affect dimensions as valence and arousal? Without first introducing these dimensions with the experiment, how should we talk about them in the model?

let's try to compress this formal model description, given that it is the same as the PNAS paper. perhaps just say that we give a sketch, and the details can be found in the other paper?

We begin with a literal listener L_0, who interprets all utterances literally. For example, the literal listener will interpret the utterance "The weather is terrible" to mean that the weather state $s = terrible$, and that the speaker has the corresponding valence and arousal towards the weather. Formally, if u is the uttered adjective:

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

where $P(v,a|s)$ is the prior probability that a speaker would feel valence v and arousal a towards the weather state s.

We assume that the speaker's goal g is to communicate along any (or multiple) of the three dimensions of meaning, which results in $2^3 - 1 = 7$ possible goals (the speaker cannot communicate along none of the dimensions). For example, the speaker may want to communicate only the weather state s, or only her valence about s, or both. A goal is thus a projection from the full meaning space to the subset of interest to the speaker. Following the Rational Speech Act model, we define the speaker's utility as the negative surprisal of the true state under the listener's distribution, given an utterance. However, here we consider only the surprisal along the goal dimension(s). To do so we project along the goal dimension(s), which leads to the following utility function for speaker S_1:

$$U(u|s,v,a,g) = \log \sum_{s',v',a'} \delta_{g(s,v,a)=g(s',v',a')} L_0(s,v,a|u) \quad (1)$$

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

$$S_1(u|s,v,a,g) \propto e^{\lambda U(u|s,v,a,g)}, \quad (2)$$

where λ is an optimality parameter. Imagine that S_1 has the goal to convey her emotional arousal a about John. Based on S_1's understanding of L_0's prior knowledge, she knows that if she produces the utterance "The weather is terrible," L_0 will believe that the weather is terrible and that the speaker is likely to feel high arousal about it. Since S_1's goal is satisfied if the listener believes that she feels high arousal towards the weather state, S_1 is motivated to produce such an utterance.

A pragmatic listener L_1 takes into account prior knowledge and his internal model of the speaker to determine the intended meaning. Because the pragmatic listener is uncertain about which goal the speaker has, L_1 will marginalize over the possible speaker goals under consideration:

$$L_1(s,v,a|u) \propto P(s)P(v,a|s)\sum_g P(g)S_1(u|s,v,a,g)$$

If L_1 thinks it is likely that speaker S_1's goal is to convey high arousal but believes (e.g. based on visual evidence) that it is *a priori* very likely that the actual weather state is *amazing* instead of *terrible*, he will determine that S_1 is using her utterance "The weather is terrible!" ironically, and that the weather is, in fact, *amazing*, and the speaker feels positive

valence and high arousal about it. Common knowledge of a domain (e.g. the weather and associated affects) and reasoning about communicative goals thus allows the pragmatic listener to infer the appropriate nonliteral and affective meanings from an ironic utterance.

Behavioral Experiments

We conducted the following two experiments to elicit people's judgements of potentially ironic utterances and their relationship to context and background knowledge. In Experiment 1, we collected participants' ratings of various weather contexts as well as how a person (e.g. Ann) feels about the weather. The weather ratings measure knowledge of the current weather state. The affect ratings measure how people generally feel about different weather states, which captures people's intuitive theory of attitudes towards the weather. In Experiment 2, we collected people's ratings of Ann's judgment of and feelings towards the weather given what she says about it, for example, "The weather is terrible!" when the weather depicted is clearly sunny and nice.

Experiment 1: Prior elicitation

Materials and methods We selected nine images from Google Images that depict the weather. To cover a range of weather states, three of the images were of sunny weather, three of cloudy weather, and three of rainy or snowy weather. Each of these images represents a *weather context*, or w_i. The images can be found here:

49 native English speakers with IP addresses in the United States were recruited on Amazon's Mechanical Turk. Each participant saw all nine images in random order. In each trial, participants were told that a person (e.g. Ann) looks out the window and sees the view depicted by the image. They then indicated how Ann would rate the weather using a labeled 5-point Likert scale, ranging from *terrible*, *bad*, *neutral*, *good*, to *amazing*. Finally, participants used slider bars 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*. The order of the emotions were randomized for each participant but remained consistent across trials for the same participant. The end points of the slider bars were labeled as "Impossible" and "Absolutely certain." A link
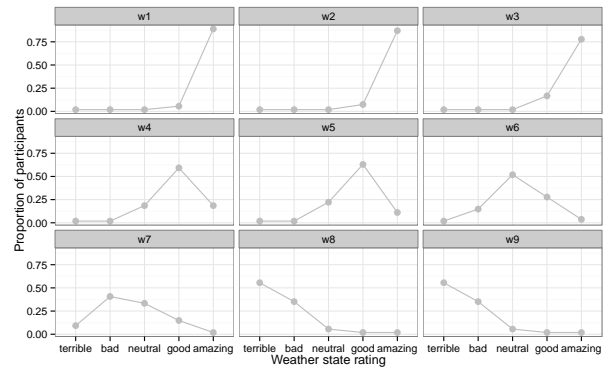
Figure 1: The proportion of participants who rated each of the weather contexts (w_i) as each weather state. Each panel represents a weather context; each line represents the distribution over weather states for a weather context prior to any linguistic input. For example, the vast majority of participants rated w_1 as *amazing*, suggesting that the prior probability of someone believing w_1 to be *amazing* is extremely high.

to the experiment is here:

Results For each of the nine weather contexts, we obtained the number of participants who gave each of the weather state ratings and performed add-one Laplace smoothing on the counts. This allowed us to compute a smoothed prior distribution over weather states given each context (see Figure **??**). To examine participants' ratings of the affects 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 also reveal the main affective dimensions that are important in this domain. We found that the first two principal components accounted for 69.14% and 13.86% of the variance in the data (see Figure **??**). In addition, they correspond roughly to the dimensions of emotional valence (positive or negative) and emotional arousal (high or low), respectively, which are consistent with theories of emotion in affective science (cite). In order to examine the probabilities of Ann feeling positive or negative affect and high or low arousal given different weather states, we converted the PCA scores into probabilities as follows. We first normalized the scores in each dimension to have zero mean and unit variance. Treating these normalized scores as quantiles of a standard normal distribution, we used the cumulative distribution function to convert the normalized scores into values between 0 and 1. Figure **??** shows the probability of positive valence and high arousal given each weather state.

Experiment 2: Irony understanding

Materials and methods We used the same nine weather images from Experiment 1. 59 native English speakers with IP addresses in the United States were recruited on Amazon's Mechanical Turk. Each participant saw all nine images in
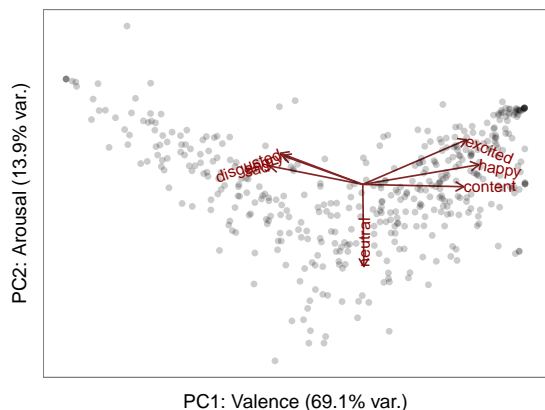
Figure 3:

Figure 2: A biplot of the two principal components of the emotion ratings from Experiment 1. PC1 corresponds roughly to positive valence, with higher values for *content, happy*, and *excited* and lower values for *sad, angry*, and *disgusted*. PC2 corresponds roughly to high arousal, with higher values for *excited* and disgusted than for *content* and *neutral*.

random order. In each trial, participants were told that Ann (characters' names were randomized) and her friend look out the window together and sees the view depicted by the image. Ann says, "The weather is _____!" where the adjective is randomly selected at each trial from the following set: "terrible," "bad," "ok," "good," and "amazing." Participants first rated how likely it is that Ann's statement is ironic using a slider with end points labeled "Definitely NOT ironic" and "Definitely ironic." They then indicated how Ann would actually rate the weather using a labeled 5-point Likert scale, ranging from *terrible*, *bad*, *neutral*, *good*, to *amazing*. Finally, participants used sliders to rate how likely it is that Ann feels each of the seven emotions about the weather. A link to the experiment is here:

Results We first examined participants' irony ratings for each of the weather context and utterance pairs. We coded "terrible" and "bad" as *negative* utterances, "ok" as a *neutral* utterance, and "good" and "amazing" as *positive* utterances. We also coded each image as *positive, neutral*, or *negative* based on the modal weather state rating from Experiment 1 (e.g. w_1 is *positive*, w_6 is *neutral*, and $w7$ is *negative*). We found a basic irony effect, where utterances whose polarities are inconsistent with the polarity of the weather context (e.g. using "terrible" to describe w_1) are rated as significantly more ironic than utterances whose polarities are consistent with the weather context (e.g. using "good" or "amazing" to describe w_1) (STATS) (Figure XX?). Furthermore, a linear regression model with the polarity of the utterance, the polarity of the
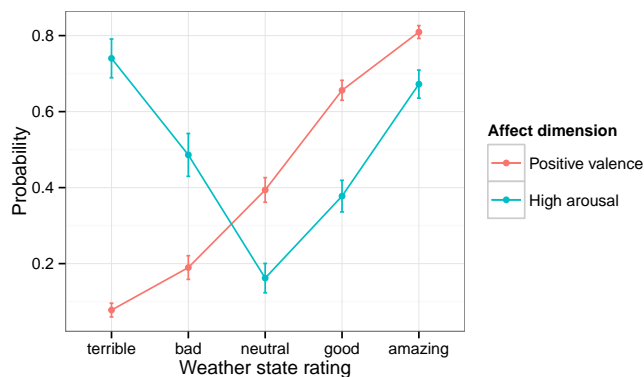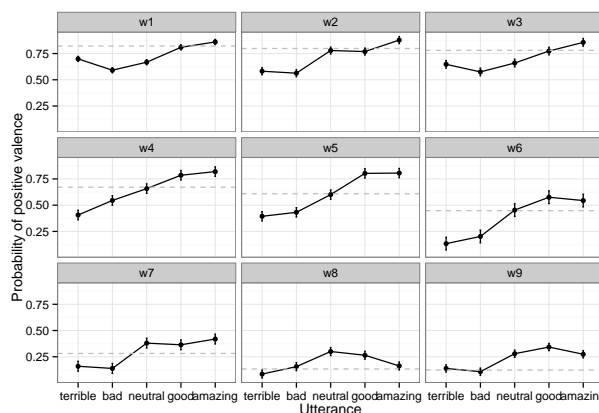
weather context, and their interaction as predictors of irony ratings produced an adjusted R-squared of 0.91, capturing most of the variance in the data.

We then examined participants' ratings of the weather context given an utterance. For each of the 45 weather context (9) × utterance (5) pairs, we obtained the number of participants who gave each of the five weather ratings. After performing add-one Laplace smoothing on the counts, we computed a smoothed distribution over weather states given each context and utterance (Figure XX?). We call this distribution the *interpreted meaning* of an utterance in context. We found that participants often rated the speaker (e.g. Ann) as judging the actual weather state to be different from what she described in her utterance, suggesting that participants interpreted these utterances non-literally. To further examine the relationship between literal meaning, interpreted meaning, and judgements of irony, we computed the Kullback-Leibler divergence between the literal interpretation of the utterance and the distribution over weather states given the context and utterance. We found that adding the KL divergence measure to the linear regression model captured significantly more variance in the irony ratings, with an R-squared of 0.93 (more STATS). This suggests that even beyond inconsistencies in polarity between utterance and context, people

what is th
literal dis-
tribution?
it's delta
the word
respondin
state, then
isn't KL j
the interp
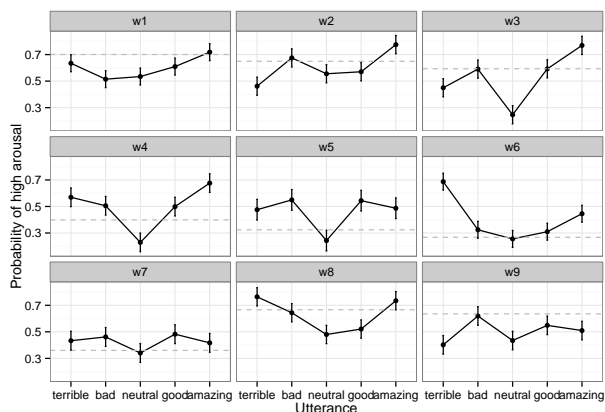
are sensitive to the fine-grained differences between literal and interpreted meanings in their judgements of irony.

Finally, we examined the emotions expressed by utterances in context. We used the loadings from the PCA on emotion ratings from Experiment 1 to predict the projected values of emotion ratings in Experiment 2 and performed the same transformations on the scores to convert them into values between 0 and 1. Figure XX shows the average probability of positive valence given a weather context and an utterance. The dotted gray lines are the average probabilities of positive valence associated with each weather context without any linguistic input, taken from Experiment 1. We see that for the *positive* weather contexts (w_1, w_2, w_3), the speaker is interpreted as more likely expressing positive valence given the extremely negative utterance "terrible" than given the less negative utterance "bad." On the other hand, for the *negative* weather contexts (w_7, w_8, w_9), the speaker is interpreted as less likely expressing positive valence given the extremely positive utterance "amazing than the less positive utterance "good" (STATS). This suggests that extreme utterances that are inconsistent with the valence of the context are more likely to express an opposite affect than its literal meaning (MAKE THIS CLEARER).

Figure XX shows the average probability of high arousal given a weather context and an utterance. The dotted gray lines are the average probabilities of high arousal associated with each weather context without any linguistic input, taken from Experiment 1. We see that regardless of the weather context, more extreme utterances are more likely to express high arousal (MAKE THIS CLEARER). (Generally, just say that the utterance shifts interpreted valence and arousal away from the baseline).

Model Evaluation

In this section, we evaluate the model's performance using results from our behavioral experiments. To produce an interpretation of an utterance in context, the model requires the following input values: (1) the prior probability of a weather state s for a given weather context, $P(s)$ (2) the joint probability of positive/negative valence given a weather state, $P(v, a|s)$
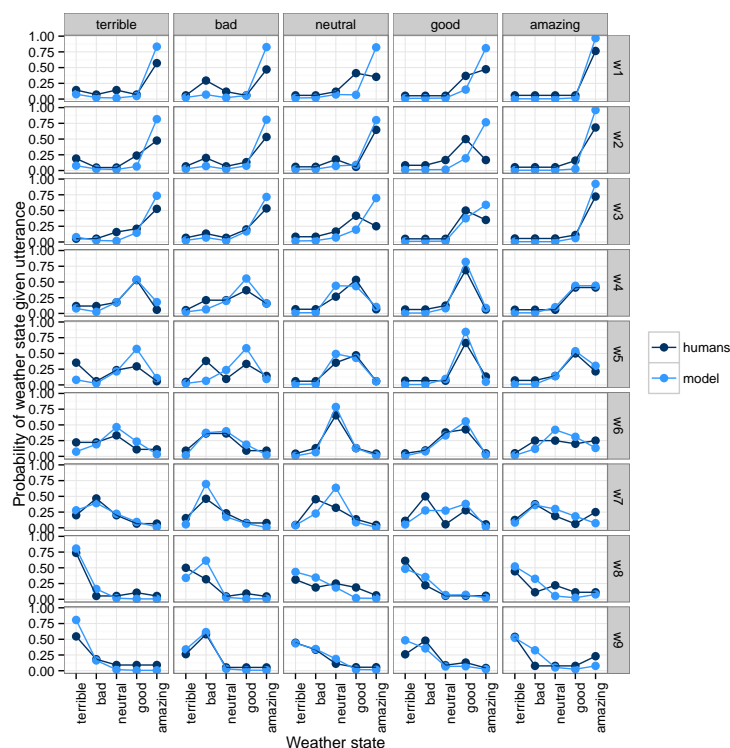
(3) the prior probability of a particular goal $P(g)$ (4) the speaker optimality parameter λ. We derived the values for (1) and (2) from Experiment 1 and fit (3) and (4) to the data from Experiment 2 (TODO: add actual fit values in a clear manner: $\lambda = 1$, p(state goal) = 0.2, p(valence goal) = 0.3, p(arousal goal)=0.4).

The model produced an interpretation for each of the 45 utterance × weather context pairs. Each interpretation is a joint posterior distribution $P(s, v, a|u)$. We first examine the model's performance on recovering the actual weather state $P(s|u)$ by marginalizing over v and a. Figure XX shows participants' and the model's interpretation of the actual weather state given an utterance and a weather context. We see that the model predictions closely match humans' interpretations, with a correlation of 0.86 (show scatter plot?). Next, we examine the model's performance on recovering the speaker's valence by marginalizing over s and a. The model's prediction for emotional valence match humans' extremely closely, with a correlation of 0.96. Finally, we examine the model's performance on recovering the speaker's emotional arousal by marginalizing over s and v. The model's prediction for emotional arousal match humans' with a correlation of 0.66 (TODO: add split half; figures???). This suggests that the model is able to incorporate background knowledge and reasoning about multiple affective goals to produce the appropriate ironic interpretations as well as the associated affects.

Discussion

With our behavioral experiments, we presented results showing the fine-grained effects of background knowledge on people's interpretations of ironic utterances, and identified the main affective dimensions that are conveyed by irony. In addition, we presented a computational model that predicts peoples' interpretations of ironic utterances using general communicative principles. In effect, the model is able to tell when an utterance is meant to be literal or ironic, and when meant to be ironic, what types of affect the speaker intends to convey. By reasoning about the speaker's communicative goals, the model goes beyond the literal meaning of an utterance to infer the actual state of the world. In addition, it recovers important aspects of the speaker's affect about the situation and captures the social and affective uses of irony. Together, these results suggest that basic principles of communication—background knowledge and reasoning about informativeness with respect to the speaker's communicative goals—may be an important driver of irony understanding.

Since Grice's original proposal, there have been two competing theories of irony understanding. The echoic mention theory (?, ?, ?) proposes that ironic utterances are intended to remind the listener of a previous utterance that turned out to be false or irrelevant. The affective subtext of irony then arises from the contrast between what was said earlier and what is actually the case. On the other hand, pretense theory (?, ?) argues that when a speaker produces an ironic utterance, she is not genuinely making the utterance, but only pretending to be someone who would make such an utterance. Understanding an ironic statement then involves recognizing the pretense as well as the preposterousness of the person being enacted. (could make this whole part shorter)

Given the prevalence of irony in natural language, many researchers in natural language processing aim to automatically detect sarcasm in large bodies of texts in order to recover the correct sentiment from an ostensibly positive or negative utterance (e.g. "I was overjoyed to pay $30 for an overcooked steak") (?, ?, ?). A critical insight that emerges from these approaches is that irony detection requires information far beyond surface linguistic cues, often calling upon a deep understanding of context and common knowledge between speaker and listener, which computers currently lack. Indeed, without sufficient contextual information, even human beings exhibit poor judgment of verbal irony (?, ?, ?).

It is important to note that the model we present here is only minimally different from the hyperbole model described earlier in the paper (?, ?). In fact, the only difference is that instead of considering a single affect dimension (negative valence), this extended model takes seriously the circumplex model of affect (cite), which identifies valence and arousal as two main dimensions underlying the slew of emotions people experience. The similarities between these two models suggests that hyperbole and irony may be understood using similar principles of communication. A deeper point we wish to make with this work is that communication in general, and nonliteral language understanding in particular, relies on reasoning about the speaker's communicative goals during interpretation. Furthermore, these goals are often social or affective in nature, and speakers are adept at harnessing shared background knowledge with the listener to convey rich affective meanings without explicitly stating them.

Our experimental paradigm and modeling framework lends itself well to a more detailed and precise account of irony understanding. For example, what are the range of social and affective meanings involved in irony understanding? We identified emotional valence and arousal in these experiments, but there may be others ....(attitudes? opinions? affect? social closeness?) (TODO: finish) What are the social functions of irony? Irony is often used to signal social closeness with the listener, presumably because it expresses the speaker's assumption that she and the listener share a great deal of common ground (cite). Given an ambiguous utterance that could be interpreted either literally or ironically, if a speaker supplements the utterance with prosodic cues to signal that it is meant ironically, will listeners judge the speaker as being closer to them? Can our model more directly test the predictions of the echoic mention and pretense theories of irony and help distinguish between them? Given the prevalence of irony in everyday language, we believe it would be *amazing* (literally) to address these questions in future research and understand irony's mysterious ability to be interpreted as the opposite of what it is.

References

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

Colston, H. L. (1997a). " i've never seen anything like it": Overstatement, understatement, and irony. *Metaphor and Symbol*, *12*(1), 43–58.

Colston, H. L. (1997b). 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.

Colston, H. L., & O'Brien, J. (2000). Contrast of kind versus contrast of magnitude: The pragmatic accomplishments of irony and hyperbole. *Discourse Processes*, *30*(2), 179–199.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning* (pp. 107–116).

Filatova, E. (2012). Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Lrec* (pp. 392–398).

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

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

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies: short papers-volume 2* (pp. 581–586).

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. (1967). Logic and conversation. *The Semantics-Pragmatics Boundary in Philosophy*, 47.

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

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.

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

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

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

Wallace, B. C., Do Kook Choe, L. K., & Charniak, E. (2014). Humans require context to infer ironic intent (so computers probably do, too). *ACL.*

Wilson, D. (2006). The pragmatics of verbal irony: Echo or pretence? *Lingua*, *116*(10), 1722–1743.