Verbal Irony, Pretense, and the Common Ground

Reuben Cohn-Gordon[1] and Leon Bergen[2]

[1] *Stanford University, Department of Linguistics*
[2] *University of California, San Diego, Department of Linguistics*

Abstract

We propose that verbal irony is a form of linguistic *countersignaling*, where agents engage in pretense about the state of the world or the perspective they hold in order to communicate about the common ground. We formalize this intuition using the Rational Speech Acts framework, by introducing a mechanism for pretense and a speaker whose goal is to be informative about the state of the common ground. In so doing, we resolve a number of the challenges facing Grice's original account for verbal irony. We show that our model extends to several types of non-declarative content in a modular way.

Keywords: verbal irony, countersignaling, pragmatics, Rational Speech Acts

Word Count: 14496

1. Introduction

Verbal irony[1] is characterized by a speaker saying something which on face value is misleading, but without the intent to mislead. In this respect, it differs from metaphor, where some aspect of the utterance's meaning is true, and hyperbole, where the degree of gradation is exaggerated but the direction is correct.

For instance, suppose Diogenes and Plato go to a disastrous play. The actors forget their lines repeatedly, the dialog is poorly written, and the plot is incoherent. Exiting the theater, Diogenes turns to Plato and remarks:

[1] Verbal (or discourse) irony is generally taken to be a superset of sarcasm, with the latter constituting cases of verbal irony which exhibit a disparaging attitude. We use *irony* to refer to verbal irony throughout, as opposed to *dramatic* or *situational* irony.

(1) I loved everything about that play.

In doing so, Diogenes does not communicate to Plato that he really did love everything about the play. But what *is* communicated, and how? Efforts to analyze irony, which stretch back to antiquity, should answer these two questions.

What does it do?. The Classical View (Cicero and Piderit, 1886; Quintilianus and Halm, 1869) is that an ironic utterance of P conveys the opposite of P. For instance, the utterance of (1) in the context described above would convey that Diogenes hated everything about the play.

Sperber and Wilson (1981) notes two problems with this account. First, irony appears in many settings where the notion of *opposite* is either ill-defined or not in keeping with what seems to be communicated, as in the following:

(2) Is she a professional singer?

(3) Tom Cruise was the actor in that wonderful movie we saw last night.

(4) That driver seems delighted/OK with what you just did.

One can utter (2) ironically about a clearly terrible singer, despite the absence of a well-defined notion of *opposite* for an interrogative. In (3), said of a terrible movie, it is not the at-issue meaning that is reversed, but rather the presupposition that the movie was wonderful. Finally, (4) can be uttered ironically, in reference to a clearly enraged driver. In one version (*delighted*) the true state of the world plausibly *is* the opposite of that conveyed by (4), but in the other (OK), it is not, since the speaker has merely understated, rather than reversed, the truth.

The second problem with the Classical view is that even in cases like (1) where "opposite(P)" is potentially being communicated, it is possible to just say what you mean directly, e.g. "I really hated that play," which is surely more direct. So there is no explanation of why irony is motivated in the first place.

How does it work?. As well as correctly describing what meaning is communicated by irony, an adequate account of the phenomenon should explain *how* it comes to communicate this meaning. That is, the account should describe the process by which a listener, on hearing an utterance u, concludes

that u is intended ironically, and further derives whatever meaning it really communicates.

While there are conventionalized ironic expressions (like "Fancy seeing you here!" and "Too bad"), as well as conventional markers of irony such as tone (Bryant and Fox Tree, 2005; Attardo et al., 2003), whether utterances like (1-4) are intended ironically can also depend on context in a way that requires reasoning about the beliefs of one's interlocutor. As such, a complete account should also allow for predictions to be made about what contexts do and do not permit irony, or more generally, what variables determine its felicitous production and interpretation.

Grice (1975) attempts to answer this question, by proposing that a listener B, on hearing an utterance from their interlocutor A with an improbable meaning reasons that "A must be trying to get across some other proposition than the one he purports to be putting forward. This must be some obviously related proposition; the most obviously related proposition is the contradictory of the one he purports to be putting forward."

First note that this inherits the problems of the Classical view, that irony always conveys the opposite of a declarative statement. Nevertheless, Grice's proposal has the merit of accounting for a key property of irony, that its use is heavily dependent on the listener's belief about the proposition that the ironic utterance literally means. For instance, the contextual factor that results in a ironic interpretation of (1) is Plato's prior belief that Diogenes did not love everything about the play. By contrast, in a different context, say one in which Plato judged the play to be good (or even mediocre), (1) could be taken at face value, so would less likely be interpreted ironically.

However, Grice's account only says how irony might be detected, and provides little insight into the eventual meaning that is inferred, or how it is obtained. Grice himself also observes that his original account overgenerates, by failing to rule out the use of irony in scenarios like the following:

> A and B are walking down the street, and they both see a car with a shattered window. B says, "Look, that car has all its windows intact". A is baffled. B says, "You didn't catch on; I was in an ironical way drawing your attention to the broken window." (Grice, 1991)

Similarly, in the spirit of example (1), if Diogenes said (5), it would be very difficult for Plato to interpret this as an ironic claim said to convey that Barack Obama was *not* in the play.

(5) Barack Obama was in that play.

We will refer to the question of why it is infelicitous to produce these ironic utterances (at least without further context) as the *car windows* problem.

Pragmatic Theories of Irony. Attempts to offer an improved pragmatic account of irony divide into two camps: *echoic* theories and *pretense* theories.

Echoic theories, originating with Sperber and Wilson (1981), propose that an ironic utterance like (1) is a mention, rather than a use of the utterance. In particular, the speaker is (either exactly or loosely) echoing an utterance to which they want to express an attitude, usually a negative one. For instance, (1) could echo a previous remark of Plato's that the play would be good, or an imagined person without taste who Diogenes intends to disparage.

A contrasting approach, from Clark and Gerrig (1984), sees verbal irony as a form of pretense, where one speaks *as if* some counterfactual obtained, relying on the common ground in order to make it clear to one's interlocutor that a pretense is being undertaken. For instance, in saying (1), Diogenes is pretending to inhabit a world in which he loved the play he just saw, with the intent that Plato will understand this to be a pretense (because, having seen the play, Plato would not be likely to believe Diogenes enjoyed it), and conclude something about the disparity of this pretense from the real world.

In the ensuing debate (Currie, 2006; Wilson, 2006), echoic theorists have claimed that the pretense theory fails to explain the derisory or allusory character of sarcasm and overgeneralizes to cases like (5), and the pretense theorists have argued that the echoic theory requires an extremely loose notion of an echo, of an utterance or stance which may never have been explicitly said.

Our proposal: Countersignaling Common Ground. We suggest that both accounts capture important insights. The ability of speakers to pretend is crucial to irony as Clark and Gerrig (1984) argues, but this alone is insufficient to explain cases where a speaker adopts the perspective (or refers to an action or utterance) of their interlocutor (or a salient third party) to express a derisory attitude towards that perspective (Sperber and Wilson, 1981).

However, we also believe that a key element is missing from both accounts, in understanding the function of pretending about the state of the world, or assuming someone else's perspective. This is the concept of *countersignaling*, coined in the game theory literature to describe situations where a signal is sent contrary to one's type:

> The nouveau riche flaunt their wealth, but the old rich scorn such gauche displays. Minor officials prove their status with petty displays of authority, while the truly powerful show their strength through gestures of magnanimity. People of average education show off the studied regularity of their script, but the well educated often scribble illegibly. Mediocre students answer a teachers easy questions, but the best students are embarrassed to prove their knowledge of trivial points. (Feltovich et al., 2002)

Taking the first example above, one explanation is that the "old money" are less likely to flaunt their wealth than the *nouveau riche*, precisely because flaunting wealth suggests one is "new money". So long as the agent's being poor is sufficiently unlikely, their failure to signal wealth (or the signal of lack of wealth) causes the receiver of the signal to infer that the agent is not only wealthy, but that this is sufficiently well known that the signal carries little risk of leading to the belief that the signaler is poor.

We propose that verbal irony is a form of linguistic countersignaling, sketched out as follows:

- Upon hearing an utterance u which entails, presupposes or implicates some unlikely state w, a listener can infer that the speaker said u as the result of pretending that w is actual.

- Thus, a speaker, with such a listener in mind, will be more inclined to utter some u which conveys a non-actual state of affairs w *when they believe that the listener already knows the true state of affairs* w'. Put another way, if the speaker believes that it is in the common ground[2] that w is not actual, then claiming, implicating or presupposing w is relatively more desirable than if this were not in the common ground. For instance, the more that Diogenes believes that Plato already knows Diogenes' views about the play they just saw, the less misleading it will be for Diogenes to say (1), since Plato will be capable of interpreting it as pretense.

- As a result, the listener, on hearing an utterance u which entails, presupposes or implicates w, can draw the following inference: the speaker

[2]We discuss our usage of the term *common ground* and how we operationalize it further in section 5.

believes that it was already (i.e. prior to their utterance) in the common ground that w was not actual.

For instance, supposing that Plato believes it unlikely that Diogenes really did love the play, Plato may infer, on hearing (1), *both* that Diogenes did not enjoy the play, *and* that Diogenes believed it was in Plato's prior knowledge that Diogenes did not enjoy it (since in that case, Diogenes' claiming he did carries less risk of miscommunication).

- Finally, a speaker who knows the listener is reasoning both about the world and the common ground *may choose to use irony in order to communicate what they believe to be the common ground*. In other words, Diogenes' goal in saying (1) is not just to convey that he did not love the play, but that the play was such that it was already in the common ground that he did not love it.

Our aim in this paper is to provide a formal model, in the spirit of the Rational Speech Acts framework (Goodman and Stuhlmüller, 2013), which captures the above process of reasoning. We propose that irony is a phenomenon perfectly suited to just such a Bayesian model, albeit one which incorporates two new mechanisms: the ability of the speaker to engage in pretense, and the ability of a listener to jointly reason about the state of the world and the common ground.

In doing so, we provide a detailed account of *how* irony works, as a process of inter-speaker reasoning, involving countersignaling. We are also making a claim about *what* irony does, namely that it communicates about the speaker's view of the listener's knowledge, which is closely related to the linguistic notion of the common ground.

This claim is key to addressing the question of motivation posed by (Sperber and Wilson, 1981) (i.e. of why irony should be used in the first place; for instance why Diogenes does not just say "I hated the play" instead of (1)?). Suppose that Diogenes' goal is to convey both that he hated the play, *and* that it was obvious, prior to any explicit statement of the fact, that he hated the play. Saying "I hated the play" would convey the former, but the opposite of the latter: in this respect, it would be counter to Diogenes' goal to say it, just in the same way that opulent displays of wealth would run counter to the old money's desire to show that their status is common ground.

By contrast, saying "I really loved the play" runs little risk of being taken at face value (since the listener can interpret it as a pretense), and further conveys that the speaker believes the listener already knows the speaker hated the play, i.e. that the speaker's opinion is already in the common ground.

For the most part, we focus our attention on cases of irony in line with the pretense theory, concluding with a discussion of how our model draws a natural distinction between these and more clearly derisory cases. We suggest that these are in principle compatible with our framework, and that in fact, our framework draws a natural distinction between the two.

Structure. The rest of the paper is structured as follows. We begin by outlining the type of Bayesian model we use to describe pragmatic inferences and language production, the *Rational Speech Acts* (RSA) framework. We then consider a previous model of verbal irony (Kao and Goodman, 2015) and show that while it can model certain cases, it fails to provide an appropriately general account, largely for the same reasons that the Classical account of irony falls short. We then introduce a model of pretense, which lays the foundation for our full countersignaling model formalizing the process of reasoning described above.

Finally, we show that the core mechanism of the countersignaling model can be extended to describe more complex instances of irony. This is achieved by combining the countersignaling model in a modular fashion with a range of other Bayesian models, designed to describe other phenomena, such as use-conditional meaning, communication under uncertainty, and questions. We view the potential for a unified account of irony as an advantage of our proposal over those which explain different subcategories of irony differently.

2. An introduction to Bayesian models of Gricean pragmatics

Under the Gricean view of pragmatics, linguistic agents enrich the meaning of an utterance by recourse to reasoning about their interlocutor, under the assumption of cooperativity (Grice, 1975).

A simple example is a scalar implicature. Assuming that the utterance in (6), which we will refer to as u_{some}, is compatible, under its semantics, with any world in which at least one chair is blue, then it is then compatible with a world (or equivalence class of worlds) w_{all} in which all the chairs are blue. However, on hearing (6), a listener may infer that not all the chairs are blue, since a listener aiming to be both truthful and informative would have

said (7), or u_{all}, if they had been able to do so truthfully; u_{all} is an utterance which is compatible with a strict subset of the worlds (6) is compatible with, and thus more informative. Therefore the listener concludes that they are in the world (or equivalence class of worlds) $w_{not-all}$ where some but not all chairs are blue.

(6) Some of the chairs in this room are blue.

(7) All of the chairs in this room are blue.

The idea to use the tools of game theory to formalize pragmatic reasoning originates with Lewis (1968) and was elaborated in (Benz et al., 2005; Franke, 2009; Jäger, 2012; Franke and Jäger, 2014). The type of model used in what follows was introduced by Frank and Goodman (2012), in the form of the Rational Speech Acts framework. RSA has the advantage of using an explicit semantics and probabilistic agents, which makes its models easy to simulate computationally. Recent work has focused on expanding the framework from simple implicatures to richer Gricean phenomena, including manner and embedded implicatures (Bergen et al., 2016; Potts et al., 2016), vagueness (Lassiter and Goodman, 2013, 2017), focus effects (Bergen and Goodman, 2015), inferences drawn from questions (Hawkins et al., 2015) and figurative uses of language (Kao et al., 2014b, a). By way of introduction to the RSA framework, we now describe a basic RSA model, which formalizes the Gricean reasoning involved in the interpretation of a scalar implicature.

RSA models are probabilistic, and define speakers and listeners as conditional probability distributions[3]. A speaker (about to produce an utterance at a particular turn of the conversation) is of the form $P(u|w)$, i.e. a distribution over which utterance $u \in U$ to say given the world they are in w. Conversely, a listener (about to interpret an utterance at a particular turn) is of the form $P(w|u)$, a distribution over the world $w \in W$ given a heard utterance u.

We first consider a particular model of a listener, one who only reasons about a semantics. We refer to this model as L_0:

[3]A distribution $p(A)$ over a set A is the pair (A, f), where f is a function $A \to \mathcal{R}$, assigning each element of A a real-valued weight between 0 and 1, such that $\sum_{a \in A} f(a) = 1$. A conditional distribution $P(A|B)$ is a function $B \to Dist(A)$, where $Dist(A)$ is the set of all possible distributions on A. In other words, a conditional distribution takes (i.e. is conditioned on) $b \in B$ and returns a distribution over A.

(8) $L_0(w|u) = \frac{[\![u]\!](w) \cdot P_L(w)}{\sum_{w' \in W} [\![u]\!](w') \cdot P_L(w')}$

Here, $P_L(W)$ is a distribution representing the listener's *prior beliefs* about the state of the world. The semantic interpretation function $[\![u]\!](w)$ is defined by:

$$[\![u]\!](w) = \begin{cases} 1, & \text{if } w \in [\![u]\!] \\ 0, & \text{otherwise} \end{cases}$$

L_0 can be understood as a listener which begins with a prior over states, receives an utterance, rules out any states semantically incompatible with that utterance, and renormalizes to obtain a posterior distribution over the states that remain.

For a concrete example, assume the following values for W, U, $P_L(W)$ and $[\![\cdot]\!]$ (for which the ordered pairs in the compatibility relation constituting the semantics are shown):

- W : $\{w_{all}, w_{not-all}\}$
- U : $\{u_{some},\ u_{all}\}$
- $P_L(W)$: $\{w_{all} : 0.5, w_{not-all} : 0.5\}$
- $[\![\cdot]\!]$: { $\langle u_{all},\ w_{all}\rangle$, $\langle u_{some},\ w_{not-all}\rangle$, $\langle u_{some},\ w_{all}\rangle$}

On these assumptions, L_0 assigns equal probability to $w_{not-all}$ and w_{all} on hearing u_{some}. To break this symmetry, we need two further layers of reasoning. First we define a speaker S_1 as follows:

(9) $T_{S_1}(u, w) = ln(L_0(w|u))$

(10) $S_1(u|w) = \frac{exp(T_{S_1}(u,w))}{\sum_{u' \in U} exp(T_{S_1}(u',w))}$

S_1 knows the state of the world w fully[4] and chooses a utterance u. S_1 consists of two parts, a utility function T_{S_1} and a decision function. Here, T_{S_1} is simply the listener's log probability of inferring w on hearing u, so that the goal of the speaker is to maximize this probability. The decision

[4]This assumption can be relaxed - see (Goodman and Stuhlmüller, 2013), which we discuss further in section 8.3.

function is softmax (Sutton and Barto, 2018). This decision function makes the speaker approximately rational: actions with higher utility will be chosen with greater probability than actions with lower utility. While T_{S_1} may vary from model to model (see section 8.3), the decision function remains the same.

Furthermore, S_1 will never say anything false (i.e. incompatible with w), since doing so would cause L_0 to assign w probability 0. As a result, S_1 can be understood as a cooperative speaker who respects the maxims of Quantity and Quality.

For instance, given the assumptions of the previous example, $S_1(u_{all}|w_{all}) > S_1(u_{some}|w_{all})$ ($Quantity$) and $S_1(u_{all}|w_{not-all}) = 0$ ($Quality$).

This puts us in a position to define a new model, L_1, capable of deriving the desired implicature above by reasoning about the world w that S_1 must have been in to have produced the heard utterance[5].

$$(11) \quad L_1(w|u) = \frac{S_1(u|w) \cdot P_L(w)}{\sum_{w' \in W} S_1(u|w') \cdot P_L(w')}$$

Note that in this model, we make the assumption that S_1 is fully knowledgeable, so that w represents the actual world. Note that this model represents higher-order knowledge among the speaker and listener (Fagin et al., 2004): L_1 is a model of a listener who believes that the speaker S_1 believes that L_0 will interpret utterances in a certain way.

Under the semantics and values for W and U provided above, L_1 prefers $w_{not-all}$ on hearing u_{some}, although w_{all} is still a possibility: $L_1(\mathrm{w}_{not-all}|u_{some}) > L_1(\mathrm{w}_{all}|u_{some})$. This corresponds to the calculation of a scalar implicature.

3. A previous model of ironic language

Kao and Goodman (2015) propose a model of irony within the RSA framework, which uses a mechanism also employed in (Kao et al., 2014b) and (Kao et al., 2014a) to circumvent S_1's strict adherence to Quality.

The core idea is to define a speaker S_1^Q which is parametrized not only by the world w, but by a partition function $q : W \to W$, inspired by the notion of a *question under discussion* (Roberts, 1996; Groenendijk and Stokhof,

[5]Note that in what follows, we often display RSA equations up to proportionality, without the normalizing term (denominator), since the numerator contains all the information necessary to derive the full equation. For instance, we would write (11) as $L_1(w|u) \propto S_1(u|w) \cdot P_L(w)$.

1984). S_1^Q cares only about communicating $q(w)$, the partition cell that w belongs to, rather than w itself.

$$T_{S_1^Q}(u, w, q) = ln(\textstyle\sum_{w'} \delta_{q(w)=q(w')} \cdot L_0(w'|u)) \quad (12)$$

$$S_1^Q(u|w, q) \propto exp(T_{S_1^Q}(u, w, q)) \quad (13)$$

Here δ is an indicator function:

$$\delta_{q(w)=q(w')} = \begin{cases} 1, & \text{if } q(w) = q(w') \\ 0, & \text{otherwise} \end{cases}$$

Importantly, this allows the speaker to avoid strict adherence to *Quality* (a property of the standard S_1). A listener L_1^Q can then jointly infer the world w and also the partition function q, which represents what aspect of the world the speaker cares about communicating (p_L^W and p_L^Q are prior over worlds and projections, respectively):

$$L_1^Q(w, q|u) \propto S_1^Q(u|q, w) \cdot p_L^W(w) \cdot p_L^Q(q) \quad (14)$$

The assumption made by Kao and Goodman (2015) is that the speaker cares about communicating either the state of the world, or one of two dimensions of their emotion regarding this state: either its valence or its intensity. For example, when communicating about the weather, the speaker may be trying to communicate that their attitude towards the weather is positively valenced (i.e., they are either mildly happy or very happy about the weather), or that their attitude has a high degree of intensity (i.e., they are either very happy or very unhappy about the weather).

Certain forms of irony can arise under this account. When the speaker says, "I really love this weather," they are literally communicating that their attitude towards the weather is positively valenced, and high intensity. A listener who already knows that the speaker dislikes the weather will try to infer why the speaker chose this utterance. The best explanation is that the speaker was only trying to communicate the intensity of their attitude, and that their true attitude is that they intensely dislike the weather.

Shortcomings of this account. L_1^Q corresponds closely the Grice's analysis and the Classical view of irony: the listener knows from their prior that the literal meaning is unlikely, and by identifying a suitable projection, they are able to infer that the true world gives rise to the opposite emotional valence

but the same intensity. However, precisely because the model conforms to the Classical view of irony, it falls prey to the problems discussed in section 1. In particular, it is capable of interpreting "That driver seems delighted with what you just did" as ironic when the driver is in fact furious (see (4)), on the assumption that *delighted* and *furious* give rise to the same degree of emotional intensity. However, it is unable to correctly interpret instances of ironic understatement, like "The driver seems OK with what you just did", as the driver being *OK* has a different degree of emotional intensity than the driver being furious.

4. Pretense

Returning to (1), repeated below in (15), the analysis offered by Clark and Gerrig (1984), inherited from Grice (1991) and in turn Aristotle (Rowe et al., 2002) is that Diogenes is *pretending* that the play is good. That is, Diogenes is talking *as if* the world were such that he had just seen a good play.

(15) I really loved that play.

This intuition of *pretense* is reinforced if we consider cases where the speaker does not directly assert a falsehood u (e.g. "You are a good singer!"), but merely acts as if the world is such that u is true, as in (2), repeated in (16):

(16) Are you a professional singer?

Under the pretense theory, the ironic speaker of (16) is pretending to be in a world where their interlocutor is a wonderful singer. In such a world, the question in (16) would be warranted, but in the actual world, it is not. Similarly in (3), the speaker pretends that the world is such that the existence presupposition of "that wonderful movie we saw last night" is met.

More generally, for every sort of thing which it is possible to pretend about, there seems to be a corresponding form of irony. This includes pretending an event took place which didn't, as in (17), pretending interest as in (18), pretending to be someone else or possess someone else's phonological characteristics, as in (19), and pretending to lack knowledge as in (20):

(17) Thanks for cleaning up after yourself!

(18) You're an accountant? Tell me more!

(19) OMG, this kale smoothie is to die for!

(20) MIT? Where's that?

It is also often possible to respond to a pretense implied by a use of irony in a way which continues or develops that pretense. For instance, in a context where a sports car is parked in the spot usually reserved for the school bus, Plato and Diogenes might have the following dialog, in which both (21) and (22) involve the same pretense. Indeed, as Clark and Gerrig (1984) point out, whole narratives can be written, like Jonathan Swift's satirical essay "A Modest Proposal" in which a pretense is continually maintained and developed.

(21) Plato: The school bus looks a little unusual today.

(22) Diogenes: Oh yeah, I'd never noticed it was convertible before.

The sheer variety of these examples lends appeal to a view in which irony relies on a quite general ability of the interlocutor to take on another *perspective*, which may be not only a pretend belief about the world but also a pretend manner of speaking, reasoning or acting. For these reasons, it seems natural that a theory of irony should be designed in conjunction with an appropriately general notion of pretense.

Note, however, that while irony requires a pretense, not all pretenses are ironic. A simple example of non-ironic pretense is child's play, in which children take on roles in imagined scenarios (see Clark (1996)).

In the spirit of producing a formal model from an intuitive, Gricean story, we begin by considering an informal *maxim of pretense*:

(23) *Make it clear when you're pretending*: if a pretend perspective could be taken by the listener to be the real one, don't pretend to have this perspective.

For instance, suppose that Plato and Diogenes go to see a play which turns out to be very experimental in nature. It is clearly the sort of thing that some people might enjoy but others not. In this scenario, Diogenes ought to think carefully about engaging in a pretense; if he says "I loved that play," he runs the risk of Plato taking him at face value. The more certain he is that Plato knows he hated the play, the less risk he runs.

4.1. L_1^P: a model of pretense

We now define a listener model L_1^P capable of reasoning about pretense, which will form a part of the full countersignaling model of irony detailed in section 6, but importantly, *does not itself constitute a model of ironic interpretation.* L_1^P is like the standard L_1, but considers the possibility that the speaker is speaking *as if* they are in some non-actual state w'. In order to define L_1^P, we introduce two new ingredients.

The first is a prior P_{prt} on the probability of pretense: this represents the listener's belief about the prior probability (i.e. before the speaker's utterance) of the speaker engaging in pretense.

The second is a *pretense channel* $P_{channel}(w'|w, P_L)$ that represents the listener's prior beliefs about how the speaker selects a pretend world. If the speaker is pretending to be in a different world, then this is drawn from a distribution which depends on the actual the actual world w and the prior over worlds P_L. The question of how $P_{channel}$ is defined (i.e. how a speaker selects a pretense state) is important, but useful to separate from the definition of L_1^P itself. For our purposes, we define $P_{channel}$ as shown in (27). Intuitively, it says to draw w' from the listener's prior distribution P, conditioning on w' being different from the actual world w. Note that when there are only two possible worlds, this channel is deterministic, but is stochastic in the general case.

(24) $L_0(w|u) \propto [\![u]\!](w) \cdot P_L(w)$

(25) $T_{S_1}(u, w) = ln(L_0(w|u))$

(26) $S_1(u|w) \propto exp(T_{S_1}(u, w))$

(27) $P_{channel}(w'|w) = \frac{P_L(w') \cdot \delta_{(w \neq w')}}{\sum_{w''} P_L(w'') \cdot \delta_{(w \neq w'')}}$

(28) $L_1^P(w|u) \propto (P_{prt}(\text{True}) \cdot P_L(w) \cdot \sum_{w_{pr}} P_{channel}(w_{pr}|w) \cdot S_1(u|w_{pr})) + P_{prt}(\text{False}) \cdot P_L(w) \cdot S_1(u|w)$

Here $P_{prt}(\text{True})$ is the probability of the speaker engaging in pretense (and therefore trying to communicate a pretend world), and $P_{prt}(\text{False})$ is the probability that the speaker is trying to communicate the actual world.

The purpose of L_1^P is not to model irony in full, but rather to incorporate a notion of pretense into a Bayesian pragmatic model, which will allow the construction of more complex models which can interpret irony. Consider an utterance such as example (30). If a non-pretending speaker could only have generated it by being in a world w that is improbable under the prior

distribution, then L_1^P should be able to infer that the speaker is pretending, and that the speaker is in some state other than w.

The first thing to note is that L_0 and S_1 remain precisely the same as in the vanilla RSA model in equations (8) and (10). Thus, the innovation in the model resides in the L_1^P, which can be understood in terms of the following generative process: having heard an utterance u, first flip a (possibly biased) coin to decide whether the speaker was pretending or not. If heads, condition, as in standard RSA, on the speaker having chosen u in order to be informative about w, the actual world. If tails, sample a state w' according to the pretense channel, and condition on the speaker having produced u in order to be informative about this pretend world.

To demonstrate how this model of pretense works, consider the following concrete example:

(29) Customer: Can I get a drink?

(30) Bartender: No.

Let us assume the following possible utterances, possible states, listener's prior over those states and pretense, and semantics. Let world T be the world in which the bartender is willing/able to serve the customer, and F the world in which they are not.

- $W : \{T, F\}$
- $U : \{Yes, No\}$
- $Prior_w : \{T : 0.8, F : 0.2\}$
- $P_{prt} : \{pretense = True : 0.1, \quad pretense = False : 0.9\}$
- $[\![\cdot]\!] = \{\langle \text{Yes}, T\rangle, \langle \text{No}, F\rangle\}$

Having fixed these details, we can inspect the distribution over worlds predicted by L_1^P. Figure 1 shows that on hearing *No*, L_1^P puts non-zero probability mass on the possibility that the true world is nevertheless T, which is possible since the speaker may merely be *pretending* that F is actual.

The model provides the listener with a way of rationalizing a heard utterance which stands in opposition to their prior beliefs. If the listener already believes that the speaker was likely to pretend, this explanation is all the more satisfactory.

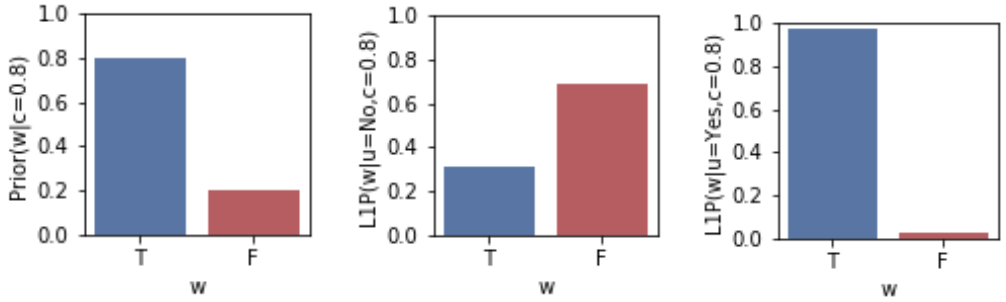

Figure 1: Barplots displaying the L_1^P prior distribution (left), posterior distribution after hearing *No* (middle) and after hearing *Yes* (right).

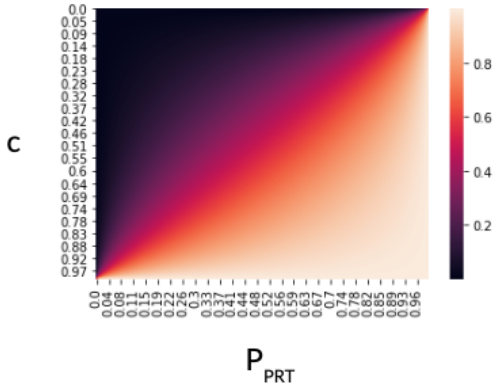

Figure 2: Each point on the heatmap displays $L_1^P(T|No, c)$, varying c from 0 to 1 on the y-axis and $P_{prt}(T)$ from 0 to 1 on the x-axis. As c increases, making $P(w = T|c)$ increasingly plausible, the probability $L_1^P(T|No, c)$ increases. It likewise increases with $P_{prt}(T)$.

What is important to note is that the degree to which L_1^P infers that the speaker is pretending when it hears *No* depends on L_1^P's prior belief that $w = T$, as well as the prior probability of pretense (see figure 2). The more certain L_1^P is that $w = T$, the more likely they are to assume that pretense is the explanation for hearing *No*.

To make this dependence of L_1^P on its prior over worlds explicit, we can rewrite (27-28) as follows:

(31) $P_{channel}(w'|w, c) = \frac{P(w'|c) \cdot \delta_{(w \neq w')}}{\sum_{w''} P(w''|c) \cdot \delta_{(w \neq w'')}}$

(32) $L_0(w|u, c) \propto [\![u]\!](w) \cdot P(w|c)$

(33) $T_{S_1}(u, w, c) = ln(L_0(w|u, c))$

(34) $S_1(u|w, c) \propto exp(T_{S_1}(u, w, c))$

(35) $L_1^P(w|u, c) \propto (P_{prt}(\text{True}) \cdot P(w|c) \cdot \sum_{w_{pr}} P_{channel}(w_{pr}|w) \cdot S_1(u|w_{pr}, c)) +$
$P_{prt}(\text{False}) \cdot P(w|c) \cdot S_1(u|w, c)$

Here, c is a term which parameterizes the listener's prior over W (this prior is shared between L_0 and L_1^P), which in this case is a Bernoulli distribution. It can therefore be written in this case as a single real number $0 \leq r \leq 1$, such that $P(w = T|c = r) = r$ and $P(w = F|c = r) = (1 - r)$. Having fixed a choice of c, L_1^P in (35) is precisely the same as defined in (28). Since c fully determines the listener's prior, we will often refer to c as the listener's prior.

Note that in L_1^P, pretense has no fundamental purpose - it is just something which people may do and which people know that people may do. As we shall see in section 5, however, given uncertainty over L_1^P's prior c, the mechanism of pretense allows a speaker to communicate information *about* this prior, by using irony.

4.2. S_2: A speaker who knows their interlocutor can detect pretenses

The speaker S_1 adheres strictly to *Quality*; there is no utility for it to choose an utterance which is false (with respect to the world that they are trying to communicate), since any such utterance is guaranteed to mislead L_0.[6]

However, a model of utterance production S_2 which reasons about the listener L_1^P who can detect pretenses, as in (37), is not constrained in this way, because L_1^P has some probability of not taking a given utterance at face value. The speaker S_2 therefore has the possibility of producing an utterance which is literally false.

(36) $T_{S_2}(u, w, c) = ln(L_1^P(w|u, c))$

(37) $S_2(u|w, c) \propto exp(T_{S_2}(u, w, c))$

For instance, $S_2(u = No|w = T, c = 0.9) > 0$. Note that S_2 still prefers to speak truthfully; that is, $S_2(u = Yes|w = T, c = 0.9) > S_2(u = No|w =$

[6]Note a subtlety here: the world that the speaker S_1 is trying to communicate about may be a pretend world or the actual one. We are assuming that *with respect to this world*, the speaker communicates honestly.

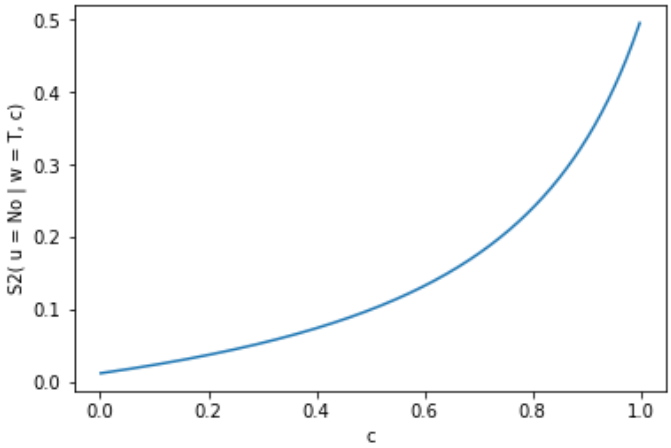

Figure 3: The probability of S_2 producing *No* falsely (i.e. when $w = T$) increases from 0 to 0.5 as c increases (and the listener puts more prior probability mass on T).

$T, c = 0.9$). So the S_2 does *not* yet have a motivation to deviate from the truth – the issue raised by Sperber and Wilson (1981) – which will require the concept of countersignaling, put forward in the section 6.

S_2 exhibits the following important property, shown in figure 3. Consider a pair u and w such that the semantics of u is *incompatible* with w. As we vary c so that $P(w|c)$ increases, $S_2(u|w, c)$ also increases. In other words, S_2's probability of deviating from the truth increases the more S_2 believes that the truth already has high probability under the L_1^P (and L_0) prior.

This behavior is precisely the "maxim of pretense" proposed in (23). S_2 obeys (23) simply out of the standard informativity utility.

5. Uncertainty and inference about interlocutor's prior knowledge

In conversation, a speaker may make assumptions about what their interlocutor already knows. This interlocutor, in turn, may have (higher order) beliefs about the speaker's assumptions, and draw inferences about them on hearing an utterance. We now discuss two particular inferences a listener hearing u with semantic meaning P can make, both of which could be construed as forms of presupposition accommodation:

Pattern 1. Saying u conveys P, but also *that P was not taken to be prior knowledge of the listener.* For instance, suppose Diogenes tells Plato: "My name is Diogenes". On hearing this, Plato learns Diogenes' name if he did

not already know it (assuming cooperativity), but also that Diogenes believed Plato not to know it, since this would best explain why he made this utterance.

Pattern 2. A second pattern of reasoning about higher order beliefs obtains when the utterance u has a literal meaning P which is considered very unlikely by the listener. Here, the listener may reason that the speaker is pretending. In order to account for the speaker following the maxim of pretense, the listener concludes that the speaker believes that the listener already believed that P was unlikely.

We propose that ironic interpretation is an example of this pattern of inference. For example, on hearing "I loved everything about that play", Plato may infer that Diogenes disliked the play, and feels licensed to pretend because he knows that Plato already believed that he disliked the play.

The common ground. A central notion in pragmatics, the *common ground* (Stalnaker, 1978, 2002), describes the set of propositions all conversational participants assume, assume everyone else assumes, and so on. Propositions that are in the common ground are pragmatically *presupposed*.

In our probabilistic model, the prior c shared between L_1^P and L_0 represents what the speaker S_2 believes that the listener believes (and believes that the speaker believes that the listener believes). For this reason, c will in general represent shared background between the speaker and listener, which the speaker can take advantage of during communication — much like the traditional concept of the common ground. Note however that the listener's prior is a distribution over worlds, rather than a set of propositions - so the notion are formally different (though see Clark and Marshall (1981); Lassiter (2012) for related discussion).

A listener can draw inferences about what the common ground must *previously* have been, given the utterance they just heard, for instance in the form of patterns 1 and 2. In terms of our model, this will amount to uncertainty over c itself (see section 5.1). In more traditional terms, this bears a resemblance to presupposition accommodation.

In view of this correspondence, we informally refer to the listener L_1^P's prior as the common ground, in particular the common ground *prior* to the speaker's utterance. Unlike the traditional common ground, however, we are not assuming that this prior distribution is *common knowledge* between the speaker and listener (Aumann, 1976). Rather, higher-order listeners (in

particular, the listener L_2) may have uncertainty over c (i.e. about what the speaker assumes they believe). The resolution of this uncertainty plays an important role in our account of irony as discussed informally above. We now describe how this type of inference may be modeled formally.

5.1. L_2: joint inference over prior and state

A listener model capable of carrying out both patterns of reasoning described in section 5 must jointly reason about w and c (see the model of Degen et al. (2015) for a comparable joint inference). In order to represent uncertainty over c, we introduce a prior over values of c, which we refer to as P_{hyp} (short for *hyperprior*). This will represent L_2's prior beliefs about what prior over W the speaker believes the listener to have. Since c itself represents a distribution, P_{hyp} is a distribution over distributions.

As an example, one possible value of P_{hyp} is shown in (38). This instantiation of P_{hyp} says that the prior is either $\{T : 0.9, F : 0.1\}$ or $\{T : 0.99, F : 0.01\}$.

(38) $\{c = 0.9 : 0.5, c = 0.99 : 0.5\}$

L_2 then has uncertainty over c *and* w, and draws inferences about both jointly, by reasoning about S_2:

(39) $L_2(w, c|u) \propto P_{hyp}(c) \cdot P(w|c) \cdot S_2(u|w, c)$

L_2 hears an utterance, and samples a possible prior c. It then samples a world w from c, and conditions on S_2 having produced the heard utterance given that S_2 was trying to convey w and assumed c to be the prior of the L_1^P and L_0. In other words, L_2 asks: given that S_2 produced u, how likely is it that the true state (according the S_2) is w and that S_2 assumed the common ground to be c.

To understand how it behaves, we can examine the L_2 posterior after hearing *Yes* or *No* (keeping to the example introduced in section 4.1 but noting that in general, it need not be that $|W| = |U| = 2$). To do this, we need to provide a concrete distribution for P_{hyp}. This distribution represents the L_2's prior uncertainty over the common ground, and is the core part of the context which determines whether an utterance will be interpreted ironically. In the following example, we use (38) as our distribution for P_{hyp}.

Results are shown in figure 4. On hearing *Yes*, L_2 reasons along the lines of *pattern 1* above. It puts the most weight on $w = T$, and $c = 0.9$, i.e. the version of the common ground in which saying *Yes* is most informative.

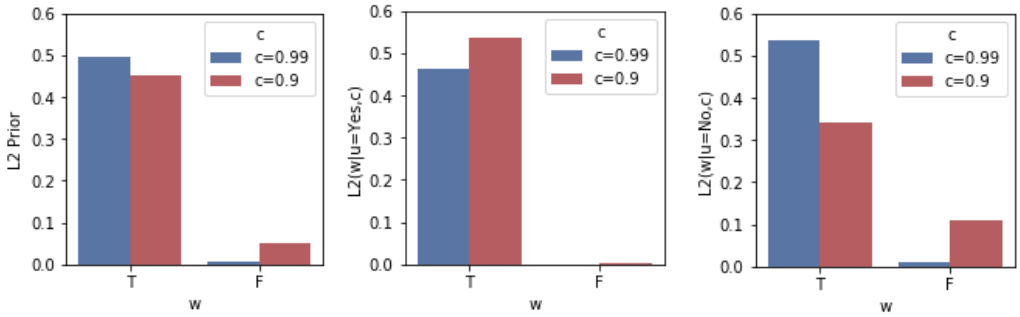

Figure 4: Barplots displaying the L_2 prior (left), the L_2 posterior distribution after hearing *No* (middle) and after hearing *Yes* (right).

On hearing *No*, L_2 reasons along the lines of *pattern 2*. It still puts the most weight on $w = T$ (because all possible priors under P_{hyp} heavily favor T), but now also on $c = 0.99$, the version of the common ground under which L_1^P is more certain that $w = T$ and thus for which S_2 is more licensed to pretend.

More generally, we can say that the L_2 posterior after hearing an utterance u is bimodal: in one mode, it favors a pair of state and common ground (w, c) where u is likely to have been produced by S_1 given w and where w has relatively low probability under c (*pattern 1*). In the other, it favors a pair (w, c) where u is *un*likely to have been produced by S_1 given w and where w has relatively high probability under c (*pattern 2*).

An illustration of this bimodality is shown in figure 5, where P_{hyp} is a Beta distribution with $\alpha = 1, \beta = 1$. In this setting, P_{hyp} is a continuous distribution, so ranges over an infinite set of possible values for c.

The role of the hyperprior in listener interpretations. The listener exhibits qualitatively interesting variation under different hyperpriors. In particular, different hyperprior values give rise to ironic interpretations to a greater or lesser degree.

While the space of all hyperpriors $D(H)$ is too large to investigate comprehensively here, we can restrict to a more tractable subset of $D(H)$. Figure 6 shows the degree of ironic interpretation over a range of values for P_{hyp}, namely those which distribute mass uniformly over two priors, respectively parametrized by c_1 and c_2, which we let range from 0.5 to 0.99. As the figure shows, the values of P_{hyp} which most allow for irony are those in which $\frac{c_2}{c_1}$ is far from 1, but where neither c_2 nor c_2 is too small. This is because, when

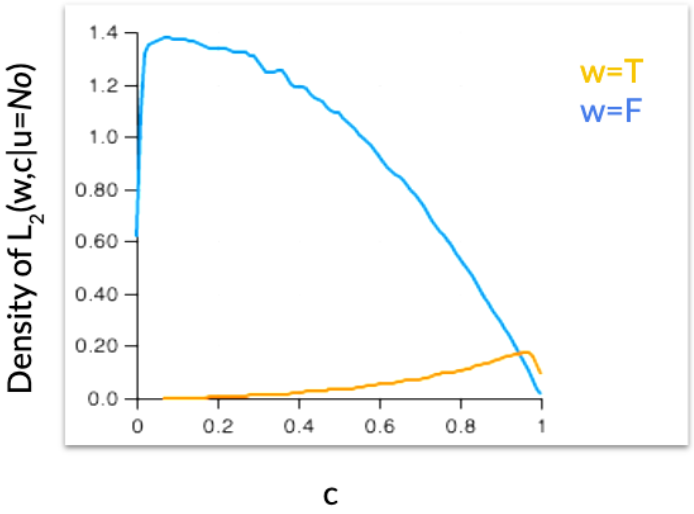

Figure 5: This figure shows the MCMC approximation of $L_2(\cdot|u = No)$ when P_{hyp} is a Beta distribution with $\alpha = 1, \beta = 1$. The mode corresponding to pattern 1 is the apex of the left hand curve, with the mode corresponding to pattern 2 is the apex of the right hand curve. Note that the ironic interpretation has low probability, on account of this particular prior assigning insufficiently high marginal probability to T.

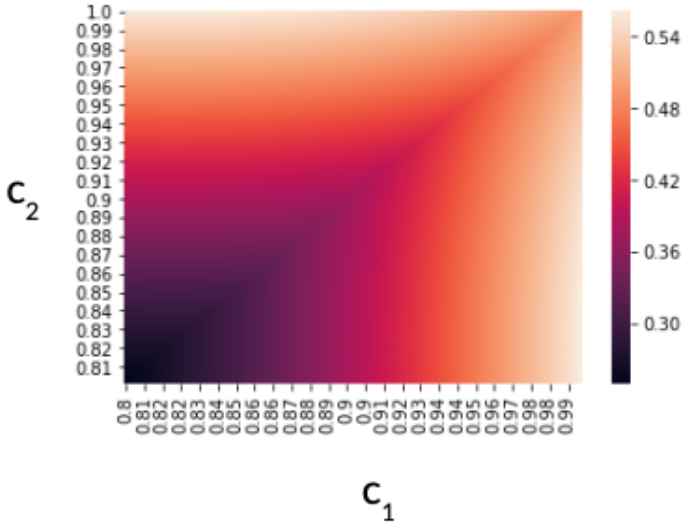

Figure 6: The heatmap shows the marginal probability $L_2(w = T|u = \text{No})$, as the values of the hyperprior c_1, c_2 are varied. Brighter regions correspond to values of the hyperprior where $L_2(w = T|u = \text{No})$ is higher. These are the values for which *No* receives an ironic interpretation.

either is too small, the marginal probability of the non-ironic interpretation of *No*, i.e. $w = T$, has sufficient probability to override the ironic interpretation. Intuitively, this corresponds to situations where the context makes it plausible that the face-value interpretation (e.g. for (30), that the bar doesn't serve drinks or for (1), that Diogenes loved the play) is the right one.

6. Countersignaling

So far, we have introduced a theory, formalized in L_2, of how irony is interpreted. However, three questions relating to the production of irony remain. The first is the question of motivation raised by Sperber and Wilson (1981). That is, why produce ironic language at all? The second is the *car windows* problem: why not produce irony in cases like that of the broken car windows described by Grice, or (5)? The third is the question of why irony is not produced in situations where there is too much uncertainty, for instance by saying (1) after a plausibly enjoyable play.

In answer to all of these questions, we propose that conveying information about the common ground is not only the effect of irony on a listener, but a goal in using irony on the part of a speaker. That is, the goal of the speaker is to communicate both the state of the world w and the state of the common ground c. Inspired by the corresponding term in game theory, we refer to this as *linguistic countersignaling*, in which a speaker undertakes to produce an utterance u which entails, implicates or presupposes a non-actual state w with the aim of having the listener employ the reasoning in *pattern 2*, and to thereby conclude that the speaker believed that the listener already believed (i.e. before u was said) that w was non-actual — i.e., that w being non-actual was already in common ground.

Consider (1) for instance. A speaker who forewent this ironic utterance and instead said "I hated that play" would successfully convey the state of the world, but fail to convey the speaker's view of the common ground, namely that the low quality of the play was in the common ground even before the utterance.

The *car windows* problem is closely related. Suppose we are in a situation where there is no uncertainty about whether the relevant proposition is in the common ground. For instance, two people, on seeing the car with its windows broken, know that the windows are not intact, know that the other knows this, and so on. Since there is no uncertainty about the common ground, at least with respect to the proposition that the windows are broken,

the speaker has nothing to communicate about c and by using irony, would only risk miscommunication (in which the listener searches for alternative pragmatic explanations of the speaker's strange utterance).

Finally, in a situation where the listener assigns relatively low probability to w in every possible prior c, the use of irony carries too much risk of being taken at face value. For instance, after seeing a play that he considered reasonably good, Plato may think it plausible that Diogenes really did love it, and uttered (1) without ironic intent.

6.1. S_3: a model of countersignaling

To model the speaker with the goal of communicating the common ground, we introduce S_3. S_3 has a common ground c which they wish to convey, as well as a world w, and with L_2 in mind chooses the utterance which will best convey both w and c. This is possible because L_2 is a model which makes inferences about the common ground - S_3's goal is to choose the utterance which gets L_2 to make the appropriate inference.

$$T_{S_3}(u, w, c) = ln(L_2(w, c|u)) \tag{40}$$

$$S_3(u|w, c) \propto exp(T_{S_3}(u, w, c)) \tag{41}$$

Contrast this to the superficially similar definition for S_2 in (37): the difference is that S_2 reasons about a model L_1^P which does not itself reason about the common ground c. Thus, S_2 is not choosing an utterance in order to communicate c, but rather choosing an utterance on the assumption that c represents the L_1^P's prior.

The key consequence of the definition of the S_3 is that under certain circumstances, the model prefers to use irony, as shown in figure 7. In particular, it prefers to use irony when there is sufficient uncertainty about the common ground (ruling out the *car windows* case), but not excessive uncertainty, ruling out the use of irony in contexts where it would be likely to be misunderstood as non-ironic.

Example. For the value of P_{hyp} defined in (38), $S_3(No|T, c = 0.99) = 0.54 > S_3(Yes|T, c = 0.99) = 0.46$: irony is preferred. The reason is that saying *Yes* (i.e. choosing the non-ironic utterance in this context) would successfully communicate $w = T$ but be unsuccessful in communicating $c = 0.99$. This corresponds to our earlier observation that saying u with the meaning of P conveys both P and that P was not previously in the common ground. However, if the speaker wants to communicate the less certain common ground

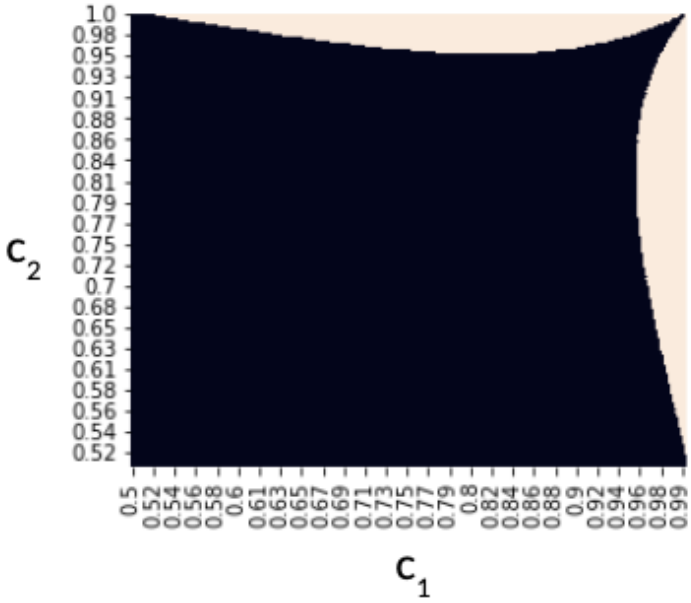

Figure 7: As in figure 6, each point corresponds to a pair of real numbers between 0 and 1, (c_1, c_2), which represent the two equally likely priors possible under P_{hyp}. Thus, each point determines a value for P_{hyp}. The area in white corresponds to values of P_{hyp} under which $S_3(No|T, c = max(C)) > S_3(Yes|T, c = max(C))$, where C is the support of P_{hyp}. This is the region where using irony is preferable to not using irony. When c_1 and c_2 are too similar (as in the top right of the heatmap), using irony is infelicitous. This corresponds to the *car windows* observation made by Grice. When either prior puts too much weight on *No* (moving left and down in the heatmap), irony is also infelicitous, corresponding to situations in which the ironic utterance is plausibly (literally) true, and may be misinterpreted.

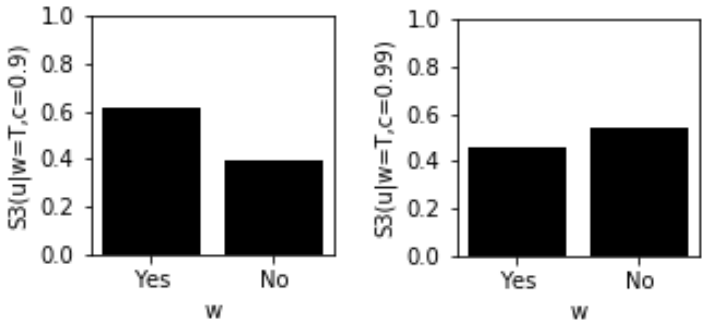

Figure 8: Barplots displaying $S_3(u|w = T, c)$, for $c = 0.9$ (left) and $c = 0.99$ (right).

0.9, $S_3(Yes|T, c = 0.9) = 0.61 > S_3(No|T, c = 0.9) = 0.39$: irony is not preferred. See figure 8 for an illustration of these results.

On this account, ironically saying "Yes" is not equivalent in meaning to un-ironically saying "No". It has a different effect on the common ground.

The behavior of S_3 also addresses the *car windows* problem. Supposing that there is no uncertainty under P_{hyp}, i.e. that all probability under P_{hyp} is on a single prior c_1, it will never be the case that $S_3(No|T, c_1) > S_3(Yes|T, c_1)$. This is because the only factor that would have justified a preference for saying *No* would have been to communicate about the common ground, but there is no uncertainty on this front.

In intuitive terms, the idea is that, when Diogenes and Plato both see a car with shattered windows, they each know that the windows are not intact, but also know that the other knows, and so on. That is, they know that the fact that the windows are intact is in the common ground. As such, there is no need for Diogenes to communicate about the nature of the common ground.

7. Interim Summary

We have proposed a theory of irony as countersignaling, building on the pretense theory proposed by Clark and Gerrig (1984). We formalized this theory within the RSA framework, culminating in the L_2 interpretation model of equation (39) and the S_3 production model of equation (41). From a modeling perspective, the key innovations are the ability of a listener model (L_1^P) to reason about whether the speaker is pretending, and the ability of a speaker model (S_3) to choose an utterance which conveys a particular com-

mon ground to their model of a listener (L_2), which jointly reasoning about the common ground and state of the world.

We observed that our account provides concrete answers to the question of what irony communicates (both the state of the world and the common ground) and how (through a process of Gricean reasoning we refer to as countersignaling). Furthermore, it is able to make precise predictions about the conditions under which irony is felicitous, in line with intuitions previously laid out in the literature.

8. Four extensions of the countersignaling model

So far, we have only shown the behavior of the model in the simplest possible domain, consisting of two states and two utterances. This does not account for any of the cases discussed as objections to the Classical view of irony, which are also the cases inaccessible to the previous model of irony proposed by Kao and Goodman (2015). However, we now show that simple extensions of this core model yield the desired generalizations to the other kinds of irony discussed in (1). We take the straightforwardness of combining RSA models of a variety of phenomena with our model of irony as one of the key virtues of our account.

8.1. Ironic Understatement

Example (42) is a simple case of irony involving a gradable adjective, when said about a driver who is clearly furious. Following Kennedy and McNally (2005), we say that the denotation of a gradable adjective is a real number, corresponding to a degree on some scale. For instance, in the context of modeling (42), we could assume a scale measuring the driver's degree of anger, and say that *furious* denotes a high positive number. The utterance *OK* denotes a number lower than *furious*, and the utterance *delighted* denotes an even lower number.

(42) That driver seems delighted/OK with what you just did.

We refer to the version of (42) with *OK* as ironic understatement. As discussed in section 1, this constitutes a simple case where the Classical theory of irony fails.

Modeling these cases in our framework requires no change to the model itself (i.e. to equations (39) and (41)) but instead a change to their interpretation; for this case, we want the state space W to range over degrees on a scale, and priors over this scale to be discretized Gaussian.

Figure 9: The 5 possible alternative utterances, with their semantics, and an example of a possible listener prior, from a Gaussian with $\mu = 2$ and $\sigma = 1.5$.

We make the arbitrary assumptions that these 5 adjectives are the possible alternatives, and that they correspond to the degrees shown in figure 9, but note that the behavior of our model is not contingent on these assumptions.

We assume that the semantics of these adjectives is bounded, so that, for example, *OK* is compatible with degrees ≤ 0, as indicated in figure 9. However, we note that similar results obtain for an exact semantics, where, for example, *OK* is compatible only with degree 0.

The prior over W represents an agent's belief that a given degree $w \in W$ represents the driver's attitude, along the scale measuring their anger.

In the previous case of (30), where $|W|$=2, any possible prior over W was a Bernoulli distribution. In this case, where $|W|$=5, there are more priors possible. We constrain the space by only considering (discretized) Gaussian distributions over W with mean at 5 (the point in the scale corresponding to *furious*). The parameter c over which L_2 has uncertainty is then the variance of the Gaussian. This corresponds to the degree of certainty the listener has that the driver's true degree of furiousness is 2 on the relevant scale.

We define a discretized Gaussian as follows. Let c be any non-negative real number. Then the probability of w given c is its probability under a Gaussian, normalized over this 5 point scale. We use $\mathcal{N}(\cdot|\mu, \sigma)$ for the probability density function of a Gaussian distribution with mean equal to μ and standard deviation equal to σ. See figure 9 for an example with $\mu = 2, \sigma = 1.5$.

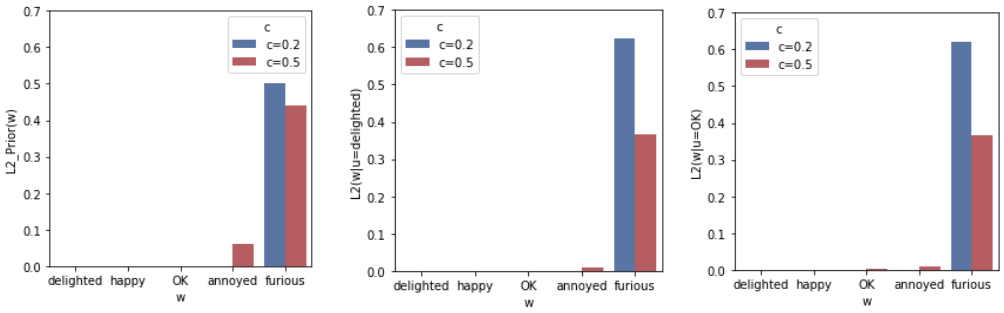

Figure 10: Barplots displaying the L_2 prior, and the posterior distribution after hearing *delighted* and *OK*, from left to right.

$$(43) \quad P(w|c) = \frac{\mathcal{N}(w|\mu{=}5,\sigma{=}c)}{\sum_{w' \in W} \mathcal{N}(w'|\mu{=}2,\sigma{=}c)}$$

For simplicity, we assume that L_2 only considers two possible values of c, 0.5 (relatively high variance, and thus low certainty that the listener knows the driver is furious) and 0.2 (relatively low variance, and thus high certainty of the same). This gives the following hyperprior:

$$(44) \quad \{c = 0.5 : 0.5, c = 0.5 : 0.2\}$$

Figure 10 shows the L_2 prior, and the L_2 posteriors on hearing *delighted* and *OK*. In both cases, we see that L_2 infers, based on prior knowledge, that *enraged* is the most probable state, but also that this must have been assumed to be prior knowledge: the probability of $(w = 2, c = 0.2)$ increases from the L_2 prior to the posterior, while the probability of $(w = 2, c = 0.5)$ decreases.

8.2. Irony in Use-conditional Meaning

A range of uses of irony share the feature that they do not involve utterances with clear truth-conditional content, or do not employ irony which relates to truth-conditional content.

Consider the following example. Diogenes and Plato are cleaning out the kitchen. On seeing a cockroach crawl out from under the fridge, Diogenes exclaims, in the voice with which one would conventionally address a cute animal:

(45) Hey little guy! (child-like tone)

The irony is (45) does not come from making a claim that is false. In this case, it is not even clear that (45) has truth-conditional meaning, but even for utterances with truth-conditional meaning like "You've got a lot of legs" said in the same voice as (45), the negation of the truth-conditional meaning is not being conveyed. Rather, (45) is an exclamation which would typically be uttered in worlds in which the object being addressed was a cute animal, rather than a cockroach. Thus, we take the speaker of (45) to be pretending that the cockroach is cute (or alternatively, that they are seeing a different animal), and to be communicating that it is in the common ground that the cockroach is not cute (or could not be mistaken for a cute animal).

Kaplan (1999) introduces the notion of use conditions (as opposed to truth conditions) to capture meaning where the notion of truth is not clearly applicable, such as the meaning of exclamations like *ouch*. Kaplan's suggestion is that such expressions should be described not by the conditions under which they are true, but rather the conditions under which they are used (see also (Gutzmann, 2015)).

An RSA model of use-conditional meaning. Qing and Cohn-Gordon (2018) propose a model of use-conditional meaning in which the conventional probability of a speaker producing u when in state w is represented by a conditional probability distribution $S_0(u|w)$. We show that by combining this model in a modular fashion with our countersignaling model of verbal irony, cases of use-conditional irony like (45) can be handled.

Consider, for example, two states of the world, one in which an animal being addressed is cute (w_1), and one in which it is not (w_2). In addition consider two utterances, a cute form of address u_1 (e.g. *Hey little guy!*) and a normal one u_2 (e.g. *Hello*). We encode in S_0 that the probability of the cute form of address u_1 (here expressed by lexical choice and tone) is higher for $w = w_1$ than $w = w_2$:

- $U = \{u_1 \text{ (cute form)}, u_2 \text{ (normal form)}\}$
- $W = \{w_1 \text{ (cute animal)}, w_2 \text{ (unpleasant animal)}\}$
- $S_0(u|w) : w = w_1 \mapsto \{u_1 : 0.9, u_2 : 0.1\}, w = w_2 \mapsto \{u_1 : 0.1, u_2 : 0.9\}$

Combining the model of use conditions with the model of countersignaling. With respect to the case in (45), where the convention that cute forms of address are directed towards cute animals is being used for ironic effect, we

can incorporate S_0 into our countersignaling model simply by modifying L_1^P to reason about S_0 instead of S_1:

(46) $P_{hyp} = \{c = 0.01 : 0.5, c = 0.1 : 0.5\}$

(47) $L_1^P(w|u,c) \propto (P_{prt}(\text{True}) \cdot P(w|c) \cdot \sum_{w_{pr}} P_{channel}(w_{pr}|w) \cdot S_0(u|w_{pr},c)) +$
$(P_{prt}(\text{False}) \cdot P(w|c) \cdot S_0(u|w,c))$

Beyond L_1^P, the model is identical to before (see equations (37,39,41)). $P(w_1|c) = c$ and $P(w_2|c) = 1 - c$. Since P_{hyp} encodes that in either possible prior, it is likely that the animal in question is not cute (but with some uncertainty about how certainly this is known), on hearing *Hey little guy!*, L_2 infers not only that the animal is not cute, but that it is assumed to be prior knowledge of the listener that the animal is not cute. Quantitatively, this case resembles the case shown in figure 4 very closely.

8.3. Pretenses about belief

A further set of cases of irony arise from pretenses about beliefs, either of the speaker or listener, rather than pretenses about the world.

In one type of case, a speaker pretends to know more than they actually do. For instance, suppose Diogenes and Plato are examining a machine in their office building of enormous complexity. Hundreds of dials and switches are set to seemingly arbitrary positions, in a way that clearly does not allow for any immediate understanding of their function. Diogenes looks at it for a moment pensively, and says:

(48) Switch 47 should be set to 122.4 not 122.5

In this case, Diogenes pretends to have more knowledge about the correct configuration of the machine than he really does. In doing so, he communicates (if his irony is successful) that it's obvious that he lacks this knowledge, on account of the machine's complexity.

To model this case, we first need a model of a speaker who communicates not the state of the world *per se*, but their belief state, i.e. a distribution over states of the world representing an agent's belief. Such a model is proposed by Goodman and Stuhlmüller (2013), where a speaker chooses the utterance which minimizes the Kullback-Leibler (KL) divergence between their belief state and their listener's posterior after hearing u (Cover and Thomas, 2012).

(49) $L_0(w|u,c) \propto [\![u]\!](w) \sum_k P(w|k) \cdot P(k|c)$

(50) $T_{S_1}(u, k, c) = -KL(k||L_0(\cdot|u, c))$

(51) $S_1(u|k, c) \propto exp(T_{S_1}(u, k, c))$

Here k is a belief state, which is represented as a probability distribution over worlds, and $P(w|k)$ is the probability of world w under this distribution. The key difference to the standard S_1 is that now, S_1's utility function is defined in terms of KL-divergence: the speaker wants to minimize the KL-divergence between their belief state and the listener's posterior distribution after hearing their utterance. Importantly, due to the properties of KL divergence, S_1 will only utter u if it knows that u is true. The model thus enforces the knowledge norm of assertion (Williamson, 2000).

We can then introduce a version of L_1^P which infers the speaker's belief state, allowing for pretenses over the belief state. Formally, reasoning about pretense is inserted into the model in the same way as was done in previous models. Note that in this case, c parametrizes a distribution over belief states, rather than states of the world:

(52) $L_1^P(k, w|u, c) \propto (P_{prt}(\text{True}) \cdot P(k|c) \cdot P(w|k) \cdot \sum_{k_{pr}} P_{channel}(k_{pr}|k) \cdot S_1(u|k_{pr}, c)) + (P_{prt}(\text{False}) \cdot P(k|c) \cdot P(w|k) \cdot S_1(u|k, c))$

(53) $L_1^P(w|u, c) = \sum_k L_1^P(k, w|u, c)$

As in equation (51), the speaker S_2 tries to choose an utterance that will minimize the KL-divergence between their belief state and that of the listener.

(54) $T_{S_2}(u, k, c) = -KL(k||L_1(\cdot|u, c))$

(55) $S_2(u|k, c) \propto exp(T_{S_2}(u, k, c))$

(56) $L_2(k, c|u) \propto P_{hyp}(c) \cdot P(k|c) \cdot S_2(u|k, c)$

The new L_2, in (56), then has uncertainty both over the speaker's belief state and the prior on belief states, and as before, performs joint inference.

To model the case in (48), we first assume a state space W ranging over possible true configurations of the machine in question. For simplicity, suppose the machine has only two states, *On* and *Off*. For the set of utterances U, we have *The machine is on*, *The machine is off*, and *silence*, with the obvious semantics.

We assume only three possible belief states, for simplicity: k_1 (which is fully certain that the state is *On*), k_2 (which is fully certain that the state

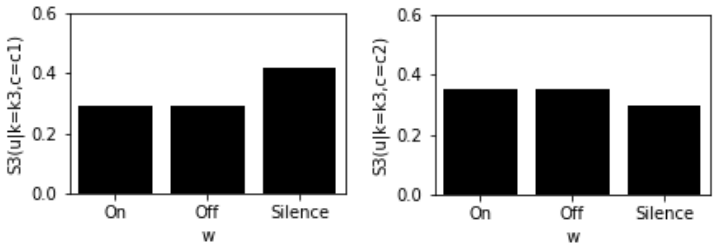

Figure 11: Barplots displaying the S_3 knowledge model predictions when trying to convey $k = k_3$, $c = c_1$ (left), where irony is not preferred, and the predictions when trying to convey $k = k_3$, $c = c_2$ (right), where irony *is* preferred.

is *Off*), and k_3 (which is maximally uncertain between *On* and *Off*). We include both k_1 and k_2 for reasons of symmetry, so that an agent who assigns equal possibility to all three knowledge states has an equal expectation of the machine being *On* and being *Off*.

(57) $k_1 : \{w = \text{On} : 1.0, w = \text{Off} : 0.0\}$

(58) $k_2 : \{w = \text{On} : 0.0, w = \text{Off} : 1.0\}$

(59) $k_3 : \{w = \text{On} : 0.5, w = \text{Off} : 0.5\}$

We then have a hyperprior with two priors over belief states in its support, both of which put the majority of weight on k_3 (the uncertain state), but with c_1 putting relatively less weight on it, and c_2 relatively more.

(60) $c_1 : \{k = k_1 : 0.2, k = k_2 : 0.2, k = k_3 : 0.6\}$

(61) $c_2 : \{k = k_1 : 0.005, k = k_2 : 0.005, k = k_3 : 0.99\}$

(62) $P_{hyp} : \{c = c_1 : 0.5, c = c_2 : 0.5\}$

In this scenario, the speaker S_2 knows that if they say *On* or *Off*, the listener L_1^P will infer that it is implausible that they actually know enough to use either of these utterances. This listener will therefore infer that the speaker must be merely pretending to be in one of these worlds, and that the speaker's actual knowledge state is k_3, which has a high degree of uncertainty about the state of the world. The speaker S_2 can therefore use *On* or *Off* to communicate this high-uncertainty knowledge state, and the listener L_2 will infer that this knowledge state was intended when they hear *On* or *Off*.

For this reason, on hearing *On* or *Off*, L_2 puts more weight on ($k = k_3, c = c_2$) that it did in its prior. Accordingly, a speaker who wishes to convey not

only k_3, but also c_2 (i.e. that k_3 was already likely under the listener's prior) prefers using irony (either saying *On* or *Off*) over silence (see the righthand plot in figure 11).

8.4. Ironic Questions

Ironic interrogatives are a clear case where what is communicated is not the opposite of the literal meaning of the utterance, since the notion of "opposite" is not even well-defined for questions. An example is (63), said in reference to someone performing a tuneless and painful rendition of a song:

(63) Is she a professional singer?

Suppose that the questioner is fairly certain, because of the quality of the singing, that the singer is not a professional. It is possible that this questioner is uncertain, however, as to whether the singer has any experience at all.

By asking a question, namely (63), whose answer they are certain of, the questioner is sacrificing their opportunity to ask a question which they may have more uncertainty about (e.g. *Has the singer ever taken lessons?*), signaling low uncertainty regarding this alternative question too.

In other words, the questioner communicates that the singer is sufficiently terrible that it is obvious they have never taken lessons, let alone performed professionally.

In order to explain how irony can arise in questions, we first need to introduce a model of the pragmatics of questions. Our model is closely related to that of Hawkins et al. (2015), though it is simplified in several respects. We leave it to future work to fully investigate the differences between these models.

The structure of the model is a variation on the normal RSA equations:

(64) $T_{Q_0}(q,k) = \lambda \sum_w k(\llbracket q \rrbracket(w)) \ln \frac{1}{k(\llbracket q \rrbracket(w))}$

(65) $Q_0(q|k) \propto exp(T_{Q_0}(q,k))$

(66) $L_1(k|q,c) \propto P(k|c) \cdot Q_0(q|k)$

The model starts with a literal questioner, Q_0. This questioner has a knowledge state k, which is a distribution over worlds. The equation T_{Q_0} defines the utility of the question q. We assume that the semantic interpretation $\llbracket q \rrbracket$ defines a partition function on worlds (Groenendijk and Stokhof, 1984); $\llbracket q \rrbracket(w)$ maps the world w to its cell in the partition. The term $k(\llbracket q \rrbracket(w))$ is then the marginal probability of w's partition cell under the distribution w:

(67) $k([\![q]\!](w)) = \sum_{w'} \delta_{[\![q]\!](w)=[\![q]\!](w')} k(w')$

The utility of the question q for Q_0 is the entropy of the distribution over answers to q. The agent receives higher utility if they have more uncertainty about how q will be answered. The constant $\lambda > 0$ determines the degree of optimality of the questioner.[7] The listener L_1 tries to infer the knowledge state of the questioner Q_0 using Bayes' rule. Here $P(k|c)$ is the prior probability of knowledge state k given common ground c. This provides a simple model of pragmatic question interpretation. The listener knows that the questioner wants to ask informative questions. From the perceived question, they try to infer which knowledge state would have made this an informative question to ask. For example, if the questioner is an expert on basketball, it is unlikely that they would ask, *Who is Michael Jordan?* A listener who hears this will infer that the questioner does not have much knowledge about the subject.[8]

In order to model ironic questions, we can introduce pretense and common ground inference in the same way as we have done in previous cases:

(68) $L_1^P(k|q,c) \propto P_{prt}(True) \cdot P(k|c) \cdot \sum_{k_{pr}} (P_{channel}(k_{pr}|k) \cdot Q_0(q|k_{pr})) + P_{prt}(False) \cdot P(k|c) \cdot Q_0(q|k)$

(69) $T_{Q_1}(q,k,c) = \ln L_1(k|q,c)$

(70) $Q_1(q|k,c) \propto exp(T_{Q_1}(q,k,c))$

(71) $L_2(k,c|q) \propto P_{hyp}(c) \cdot P(k|c) \cdot Q_1(q|k,c)$

Here L_1^P is a listener who believes that the questioner may be engaging in pretense. In particular, the questioner may be asking a question from the perspective of a knowledge state k_{pr} which has been sampled from the pretense channel $P_{channel}$. The listener L_1^P needs to jointly infer whether the questioner was engaging in pretense, and what the questioner's knowledge state is.

The questioner Q_1 chooses a question in order to communicate their knowledge state to the listener L_1^P; intuitively, their goal is to communicate

[7] An alternative interpretation is that λ determines the base of the logarithm which is used to calculate entropy.

[8] Note that this model ignores an important aspect of question interpretation, namely inferences about the questioner's intended QUD. It is straightforward to incorporate QUD inferences into this model, and the discussion of irony here remains largely unchanged in this setting.

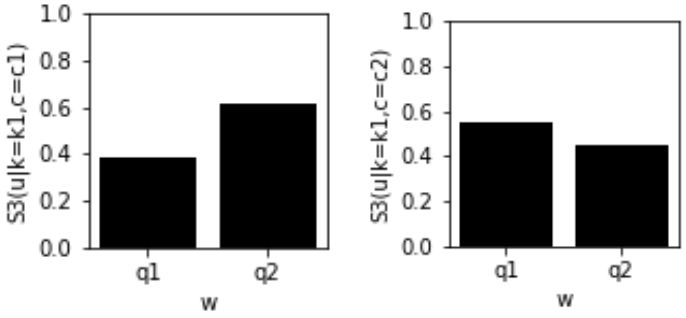

Figure 12: Barplots displaying the S_3 question model predictions when trying to convey $k = k_1$, $c = c_1$ (left), where irony is not preferred, and the predictions when trying to convey $k = k_1$, $c = c_2$ (right), where irony *is* preferred. q_1 is the question *Is the singer a professional?*, and q_2 is the question, *Has the singer ever received lessons*

this knowledge state so that the listener will know how to provide useful information in subsequent communication turns. The listener L_2 jointly infers the common ground and the questioner's knowledge state from the question.

As a concrete example, we model (63). We allow three possible states: that the singer is a professional (w_1), has received lessons but isn't a professional (w_2) and has never even received lessons (w_3). Further, we allow two questions: *Is the singer a professional?* (q_1) and *Has the singer ever got lessons?* (q_2). Finally, we define two possible knowledge states: k_1, in which the questioner is certain that the singer is not a professional (but uncertain as to whether they have had any training) and k_2, in which the speaker is certain that the singer took lessons (but uncertain as to whether they are a professional).

(72) $k_1 : \{w = \mathrm{w}_1 : 0.0, w = \mathrm{w}_2 : 0.5, w = \mathrm{w}_3 : 0.5\}$

(73) $k_2 : \{w = \mathrm{w}_1 : 0.5, w = \mathrm{w}_2 : 0.5, w = \mathrm{w}_3 : 0.0\}$

We then have a hyperprior with two priors over belief states in its support, both of which put the majority of weight on k_3 (the uncertain state), but with c_1 putting relatively less weight on it, and c_2 relatively more.

(74) $c_1 : \{k = k_1 : 0.9, k = k_2 : 0.1\}$

(75) $c_2 : \{k = k_1 : 0.999, k = k_2 : 0.001\}$

(76) $P_{hyp} : \{c = c_1 : 0.5, c = c_2 : 0.5\}$

This hyperprior represents a context in which all possible common grounds heavily favor the belief that the singer is not a professional. When trying to communicate not only k_1 but also c_2 (i.e. that k_1 was very probable under the listener's prior) S_3 prefers the ironic question q_1 over q_2, whereas when trying to communicate k_1 and c_1, q_2 is preferred (see figure 12).

9. Concluding remarks

By proposing a core mechanism of countersignaling behind ironic language, we are able to provide a unified account of irony across many utterance types, which addresses the following key questions:

- What does irony communicate?
- Why use irony at all?
- What situations do and do not permit irony, and why?

All three questions are answered in a unified way: irony communicates about the common ground, and this is the reason to use it, in certain situations, to speaking at face value. In particular, it should be used in situations where there is some uncertainty about the common ground, but not enough marginal uncertainty about the world to make it unclear that a pretense is being perpetrated.

The uses of verbal irony are very broad, and we have said nothing of ironic imperatives, performatives or apologies, to name a few. While we have discussed semantic presuppositions intended ironically, like (3)), we have not offered a corresponding model. However, our belief is that the approach we employ, in conjunction with an appropriate account of each of these phenomena, will yield models of their ironic usage straightforwardly.

There are, however, a number of cases that we do not expect to be able to account for so easily. One class of challenging cases involve *criticisms of reasoning*. One of the original attested cases of irony provides an example: in response to Plato claiming that humans are the unique featherless biped, Diogenes walks in the forum holding a plucked chicken and utters (77). Diogenes' utterance communicates that Plato's belief is obviously wrong.

(77) Behold, a man!

The challenge in modeling this case is that any model of Plato's beliefs in which he performs deductive closure, and in which he already knew that a plucked chicken is a featherless biped must assign no probability to the belief that humans are the only such creature. Building a model of belief without deductive closure is well beyond the scope of this project (see Garrabrant et al. (2017)).

9.1. What about disparagement?

Sperber and Wilson (1981) connect irony (and sarcasm in particular) with disparagement and reference to a previous event, noting that "...what is missing from non-echoic versions of the pretense account is precisely what is emphasized by the echoic account: that the attitude expressed in irony is primarily to a thought or utterance that the speaker attributes to some identifiable person or type of person, or to people in general."

Indeed, and as Grice (1991) himself notes, saying "The car windows are intact" can be a felicitous use of irony if one's interlocutor previously commented repeatedly on the safety of the neighborhood. Similarly, (5) could be said to mock a friend who had previously and wrongly claimed to have spotted a former president. In these contexts, an allusion is being made to a previous event (Kumon-Nakamura et al., 1995).

On the other hand, we agree with Currie (2006) that irony is not always disparaging. For instance, in (78), the speaker is not disparaging their interlocutor, and we see no reason to think that there is another implicit agent who is being disparaged either. Instead, the speaker is engaging in a pretense that "Euler's method" is a surprising name for a method invented by Euler.

(78) This method was invented by Euler. You'll be shocked to hear that it's called Euler's method.

Cases of mockery, by contrast, seem to involve the speaker taking on the perspective of either their interlocutor or a salient third party. Such perspective-taking is not explicitly represented in our model. We posit, however, that the basic architecture of this model can be used to represent perspective-taking, and capture the pragmatic effects that echoic theorists like Sperber and Wilson (1981) are interested in.

There are two changes to the model that would be required for perspective-taking. First, speakers are currently parameterized only by the world or beliefs that they are trying to communicate. In a model of perspective-taking, this would need to be enriched, to include the other factors that

determine a speaker's perspective, such as attitudes and mannerisms. Second, the pretense channel would need to be modified, allowing speakers to take on non-actual perspectives which are salient in the conversation. While likely feasible, these changes are sufficiently non-trivial that we leave their full investigation for future work.

9.2. What about conventionalized markers of irony?

Verbal irony, and sarcasm in particular, is often associated with a tone of voice which makes clear that the utterance is intended ironically, at least in English. While we do not offer a theory of how conventional markers interact with the process of Gricean reasoning we have presented, we speculate that our model can be extended in a way which would allow this sort of conventional knowledge about irony to be incorporated.

The solution would involve encoding in S_1 the conventional association between being in a pretense world and using ironic tone (or whatever other features are conventional of irony). As a result, a listener who heard language marked in this way would be inclined to assume a pretense was being undertaken. This would encourage the use of conventional markers in situations where there was insufficient context to allow for the successful use of unmarked sarcasm.

9.3. Future work

Our proposal for verbal irony involves two core components: countersignaling and pretense. Of these, we suspect that the latter, which amounts to a mechanism for speaking counterfactually, can be extended in many ways, to more complex varieties of irony, like (79) and (21-22) as well as other figurative language out of the scope of previous models (Kao et al., 2014b, a), like (80).

(79) Plato: Did you say you were planning to buy a Tesla?
Diogenes: Yes, and then I'm going to move to the Bay and found a start-up.

(80) What were the dinosaurs like?

Example (79) communicates a probabilistic relationship between two variables, namely that a person who buys a Tesla is also likely to work at a start-up in the Bay Area. The ability to communicate this sort of complex

world knowledge is a powerful function of pretense and irony which merits significant further work.

In (21-22), two agents jointly construct a counterfactual world over successive turns of a conversation. Note that our mechanism of pretense is well suited to the assumption that there is a fixed pretend world, about which agents can be informative, and which can be maintained across conversational turns.

In (80), the speaker talks as if they are in a pretend world in which it is presupposed that the speaker was born before the extinction of the dinosaurs, a hyperbolic utterance which communicates that the speaker is old. Cases of this sort are challenging for existing models of hyperbole, e.g. (Kao et al., 2014b) but seem amenable to models which incorporate a notion of pretense.

Bibliography

Salvatore Attardo, Jodi Eisterhold, Jennifer Hay, and Isabella Poggi. Multimodal markers of irony and sarcasm. *Humor*, 16(2):243–260, 2003.

Robert J Aumann. Agreeing to disagree. *The annals of statistics*, pages 1236–1239, 1976.

Anton Benz, Gerhard Jäger, Robert Van Rooij, and Robert Van Rooij. *Game theory and pragmatics*. Springer, 2005.

Leon Bergen and Noah D Goodman. The strategic use of noise in pragmatic reasoning. *Topics in cognitive science*, 7(2):336–350, 2015.

Leon Bergen, Roger Levy, and Noah Goodman. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 9, 2016.

Gregory A Bryant and Jean E Fox Tree. Is there an ironic tone of voice? *Language and speech*, 48(3):257–277, 2005.

Marcus Tullius Cicero and Karl Wilhelm Piderit. *De oratore*. BG Teubner, 1886.

Herbert H Clark. *Using language*. Cambridge university press, 1996.

Herbert H Clark and Richard J Gerrig. On the pretense theory of irony. 1984.

Herbert H Clark and Catherine R Marshall. Definite knowledge and mutual knowledge. 1981.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Gregory Currie. Why irony is pretence. *The architecture of the imagination*, pages 111–133, 2006.

Judith Degen, Michael Henry Tessler, and Noah D Goodman. Wonky worlds: Listeners revise world knowledge when utterances are odd. In *CogSci*, 2015.

Ronald Fagin, Joseph Y Halpern, Yoram Moses, and Moshe Vardi. *Reasoning about knowledge*. MIT press, 2004.

Nick Feltovich, Richmond Harbaugh, and Ted To. Too cool for school? signalling and countersignalling. *RAND Journal of Economics*, pages 630–649, 2002.

Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

Michael Franke. *Signal to Act: Game Theory in Pragmatics*. ILLC Dissertation Series. Institute for Logic, Language and Computation, University of Amsterdam, 2009.

Michael Franke and Gerhard Jäger. Pragmatic back-and-forth reasoning. In Salvatore Pistoia Reda, editor, *Pragmatics, Semantics and the Case of Scalar Implicatures*, pages 170–200. Palgrave Macmillan UK, London, 2014.

Scott Garrabrant, Tsvi Benson-Tilsen, Andrew Critch, Nate Soares, and Jessica Taylor. A formal approach to the problem of logical non-omniscience. *arXiv preprint arXiv:1707.08747*, 2017.

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

H Paul Grice. Logic and conversation. *1975*, pages 41–58, 1975.

H Paul Grice. *Studies in the Way of Words*. Harvard University Press, 1991.

Jeroen Antonius Gerardus Groenendijk and Martin Johan Bastiaan Stokhof. *Studies on the Semantics of Questions and the Pragmatics of Answers*. PhD thesis, Univ. Amsterdam, 1984.

Daniel Gutzmann. *Use-conditional meaning: Studies in multidimensional semantics*, volume 6. OUP Oxford, 2015.

Robert XD Hawkins, Andreas Stuhlmüller, Judith Degen, and Noah D Goodman. Why do you ask? good questions provoke informative answers. In *CogSci*. Citeseer, 2015.

Gerhard Jäger. Game theory in semantics and pragmatics. *Semantics: An international handbook of natural language meaning*, 3:2487–2425, 2012.

Justine T Kao and Noah D Goodman. Let's talk (ironically) about the weather: Modeling verbal irony. In *CogSci*, 2015.

Justine T. Kao, Leon Bergen, and Noah D. Goodman. Formalizing the pragmatics of metaphor understanding. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, pages 719–724, Wheat Ridge, CO, July 2014a. Cognitive Science Society.

Justine T. Kao, Jean Y. Wu, Leon Bergen, and Noah D. Goodman. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007, August 2014b.

David Kaplan. The meaning of ouch and oops: Explorations in the theory of meaning as use. *Manuscript, UCLA*, 1999.

Christopher Kennedy and Louise McNally. Scale structure, degree modification, and the semantics of gradable predicates. *Language*, pages 345–381, 2005.

Sachi Kumon-Nakamura, Sam Glucksberg, and Mary Brown. How about another piece of pie: The allusional pretense theory of discourse irony. *Journal of Experimental Psychology: General*, 124(1):3, 1995.

Daniel Lassiter. Presuppositions, provisos, and probability. *Semantics and Pragmatics*, 5:2–1, 2012.

Daniel Lassiter and Noah D Goodman. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and linguistic theory*, volume 23, pages 587–610, 2013.

Daniel Lassiter and Noah D Goodman. Adjectival vagueness in a bayesian model of interpretation. *Synthese*, 194(10):3801–3836, 2017.

David Lewis. *Convention: A philosophical study*. John Wiley & Sons, 1968.

Christopher Potts, Daniel Lassiter, Roger Levy, and Michael C Frank. Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics*, 33(4):755–802, 2016.

Ciyang. Qing and Reuben. Cohn-Gordon. Use-conditional meaning in rational speech act models. *Sinn und Bedeutung*, 111, 2018.

Marcus Fabius Quintilianus and Karl Halm. *Institutio oratoria*, volume 2. Teubner, 1869.

Craige Roberts. Information structure in discourse: Towards an integrated formal theory of pragmatics. 1996.

Christopher J Rowe, Sarah Broadie, et al. *Nicomachean ethics*. Oxford University Press, USA, 2002.

Dan Sperber and Deirdre Wilson. Irony and the use-mention distinction. *Philosophy*, 3:143–184, 1981.

Robert Stalnaker. Assertion. 1978.

Robert Stalnaker. Common ground. *Linguistics and philosophy*, 25(5):701–

721, 2002.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. 2018.

Timothy Williamson. *Knowledge and its Limits*. Oxford University Press on Demand, 2000.

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