Wonky worlds: Listeners reconsider common ground when utterances are odd

Judith Degen, Michael H. Tessler, Noah D. Goodman
{jdegen,mtessler,ngoodman}@stanford.edu
Department of Psychology, 450 Serra Mall
Stanford, CA 94305 USA

Abstract

World knowledge enters pragmatic utterance interpretation in complex ways. Sometimes, a speaker's utterance suggests that listeners should disregard their world knowledge, yet current models of pragmatic interpretation either disregard the role of world knowledge or overestimate its role in interpretation. Here we provide an extension to the Rational Speech Act model of scalar implicature that captures whether listeners believe they are in an abnormal–or 'wonky'–world after observing a speaker's utterance, in which case they downweight their prior beliefs in the computation of speaker meaning. We show in four experiments that a) listeners have varying prior beliefs about the probability of various objects exhibiting an effect (e.g., marbles sinking), b) these beliefs influence listeners' expectations about how many objects will show the effect after observing an utterance (like *Some of the marbles sank*), c) these beliefs influence scalar implicature strength, and d) listeners' world wonkiness judgments are affected by the surprisal of the observed utterance under their prior beliefs. The extended model is the first quantitative model that accounts for how rational listeners should integrate world knowledge in pragmatic utterance interpretation, and provides a close match to the empirically obtained data.

Keywords: scalar implicature; world knowledge; prior beliefs; experimental pragmatics; computational pragmatics

How often do you think marbles would sink in water? Probably extremely often, if not always. Now imagine reading *Max threw fifteen of his favorite marbles in the water. Some of them sank.* Have you begun to reconsider your assumptions? Perhaps you now suspect that these marbles are in fact made of plastic or the water is covered with thick algae? That is, that they are not just normal marbles in normal water. Here we explore how prior world knowledge enters into pragmatic utterance interpretation, and how this world knowledge is defeasible: some utterances lead listeners to conclude that the world under discussion is abnormal and has appropriately different prior probabilities. We refer to such an abnormal world as a *wonky* world.

The Rational Speech Acts framework (RSA) (Frank & Goodman, 2012; ?, ?), and related game-theoretic models (?, ?), treat communication as a signaling game (?, ?) between a speaker and a listener. The listener reasons by Bayesian inference about what the world is like given that a speaker who produced the utterance is trying to be informative (with respect to a naïve listener). Variants of these models have successfully captured listeners' quantitative behavior on a number of pragmatic inference tasks, including ad hoc Quantity implicature ref, M-implicature ref, scalar implicature ref, embedded scalar implicatures ref, and non-literal language (?, ?). A defining feature of Bayesian reasoning is that prior beliefs affect inferences that will be drawn. Bayesian models of language interpretation, accordingly, predict that prior beliefs about the world should affect the listener's interpretation of an utterance. While this impact of prior knowledge has been noted, and included in models, it hasn't been systematically studied.

Generalizing our opening example, take the case of *Some of the X-s Y-ed* (for category X and event Y). When the prior probability, $\theta_{X,Y}$, of an X Y-ing is not extreme, RSA leads to the standard scalar implicature: the posterior probability that all 15 of the Xs Yed, after hearing the utterance, is much lower than its prior probability. This is because a rational speaker would have been expected to say the more informative *all of the X-s Y-ed* if it had been true. As we will show below, RSA makes two strong predictions about the effect of the prior: (1) As $\theta_{X,Y}$ approaches 1, the interpretation probability that all X-s Y-ed approaches 1, that is, the scalar implicature disappears. This prediction follows because the extreme prior overwhelms the effect of the utterance. (2) The posterior expectation of the number of X-s that Y-ed should be approximately the same as the prior expectation, when the number of tries is large (e.g. 15)–UPDATE. . This prediction follows from the weak semantics of *some* and the isolated effect of the alternative *all*: because *Some of the X-s Y-ed* only restricts the interpretation to be greater than zero and the scalar implicature resulting from alternative *All of the X-s Y-ed* can at the most rule out the state in which all X-s Y-ed, the prior will dominate the inference of exactly how many X-s Y-ed.

However, intuition is at odds with these predictions: for example, Geurts (2010) has observed that for events with very high prior probability of occurrence (e.g. marbles sinking), observing an utterance of *Some of the marbles sank* leads to very strong implicatures, that is, the subjective probability that all of the marbles sank is intuitively close to 0. In Exp. 1 we collect prior probabilities for a variety of events and categories. In experiment 2a and 2b we collect posterior interpretations after hearing utterances such as *Some of the X-s Y-ed*. These experimental results confirm the intuition of relatively strong implicature—hence prediction (1) of RSA is incorrect—and show that the prior has a muted effect on posterior expectation—hence prediction (2) of RSA is incorrect. Given the previous success of RSA models, this constitutes a striking puzzle. To address this puzzle we pursue the intuition raised at the very beginning of this paper: that sometimes, the speaker's utterance will lead the listener to infer that the world under discussion is wonky and she should therefore down-weight her prior beliefs in the computation of speaker meaning. We introduce a variant of RSA, wRSA, in which the listener can revise her beliefs about the domain that the speaker intended. We show that this extension re-

solves the puzzle of the prior's muted effects. In Experiment 3 we explore participants' intuitions about whether the world is normal in the scenarios of Experiment 2 and find that wRSA predicts listeners' wonkiness judgements. make sure this follows the new organization

Experiment 1

Exp. 1 measured listeners' prior beliefs about how many objects exhibit a certain effect (e.g., how many marbles sink).

Method

Participants We recruited 60 participants over Amazon's crowd-sourcing platform Mechanical Turk.

Procedure and materials On each trial,[1] participants read a one-sentence description of an event like *John threw 15 marbles into a pool.* They were then asked to provide a judgment of an effect, e.g. *How many of the marbles do you think sank?*, on a sliding scale from 0 to 15. Each item had a similar form: the first sentence introduced the objects at issue (e.g., marbles). The question always had the form *How many of the X Yed?*, where *X* was the object noun phrase introduced in the first sentence (e.g., *marbles*, *cups*, *balloons*) and *Yed* was a verb phrase denoting an effect that the objects underwent (e.g., *sank*, *broke*, *stuck to the wall*). Each verb phrase occurred with three different objects, e.g., *sank* occurred with *marbles*, *cups*, and *balloons*. Items were constructed to intuitively cover the range of probabilities as much as possible, while also somewhat oversampling the upper range of probabilities to have more fine-grained coverage of this region that is of most interest for testing the RSA model. Judgments were obtained for 90 items, of which each participant saw a random selection of 30 items.

Results

Data from one participant, who gave only one response throughout the experiment, were excluded. Each item received between 12 and 29 ratings. Distributions of ratings for each item were smoothed using Laplace smoothing. As intended, items covered a wide range of probabilities. include measure or plot of this once you decide which smoothing to use ultimately.

In the next section, we use these empirically obtained prior beliefs to derive RSA predictions for the interpretation of utterances like *Some of the marbles sank*, before empirically measuring participants' interpretations.

Effect of the world prior in RSA

The basic Rational Speech Acts model defines a pragmatic listener $P_{\text{listener}}(s|u)$, who reasons about a speaker $P_{\text{speaker}}(u|s)$, who in turn reasons about a literal listener $P_{\text{literal}}(s|u)$. Each listener does Bayesian inference about the world state, given either the literal truth of utterance u or the speaker's choice of u; the speaker is a softmax-optimal decision maker, with the goal of being informative about the state s. RSA is defined by:

$$P_{\text{literal}}(s|u) \propto F_u(s) \cdot P(s) \tag{1}$$

$$P_{\text{speaker}}(u|s) \propto \exp(\lambda \ln P_{\text{literal}}(s|u)) \tag{2}$$

$$P_{\text{listener}}(s|u) \propto P_{\text{speaker}}(u|s) \cdot P(s) \tag{3}$$

Here $F_u : s \mapsto \{0,1\}$ is a truth-function specifying the meaning of each utterance.

For concreteness, assume that the set of states of the world is $S = \{s_0, s_1, s_2, \ldots, s_{15}\}$, where the subscript indicates the number of objects (e.g., marbles) that exhibit a certain effect (e.g., sinking). Assume also that the set of utterances *All/None/Some of the marbles sank* is denoted $U = \{u_{\text{all}}, u_{\text{none}}, u_{\text{some}}\}$ and each has its usual literal meaning: $F_{u_{\text{none}}}(s)=\delta_{s=0}$, $F_{u_{\text{some}}}(s)=\delta_{s>0}$, $F_{u_{\text{all}}}(s)=\delta_{s=15}$.

In Figure 1 we show the predictions of RSA for the items from Exp. 1 in two different ways: the left panel shows the the expected number of affected objects of the posterior distribution as a function of the expected value of the prior distribution; the right panel shows the posterior probability of the state in which all objects are affected, as a function of the prior probability of that state. We see that the prior has a strong effect, which can be summarized by the two predictions described in the Introduction.[2] We next turn to an empirical test of these predictions, or rather, of the intuition that they may be incorrect.

Experiment 2a and 2b

Exps. 2a and 2b[3] measured participants' posterior beliefs $P(s|u)$ about how many objects exhibited a certain effect (e.g., marbles sinking), after observing an utterance. The only difference between the experiments was the dependent measure. The dependent measures differed in order to directly and independently estimate the two values that the predictions laid out in the introduction are concerned with: $\mathbb{E}[P(s|u_{\text{some}})]$ and $P(s_{15}|u_{\text{some}})$.

Method

Participants For each experiment we recruited 120 participants over Amazons crowd-sourcing platform Mechanical Turk.

Procedure and materials Participants read the same descriptions as in Exp. 1. They additionally saw an utterance produced by a knowledgeable speaker about the event, e.g. *John, who observed what happened, said: "Some of the marbles sank"*. In Exp. 2a (just as in Exp. 1), they then provided a judgment of an effect, e.g. *How many of the marbles do you think sank?*, on a sliding scale from 0 to 15. In Exp. 2b

[1] This experiment can be viewed at https://www.stanford.edu/~jdegen/12_sinking-marbles-prior15/sinking-marbles-prior.html

[2] comment on the noise?

[3] These experiments can be viewed at https://www.stanford.edu/~jdegen/13_sinking-marbles-priordv-15/sinking-marbles.html and https://web.stanford.edu/~jdegen/16_sinking-marbles-sliders-certain/sinking-marbles-nullutterance.html

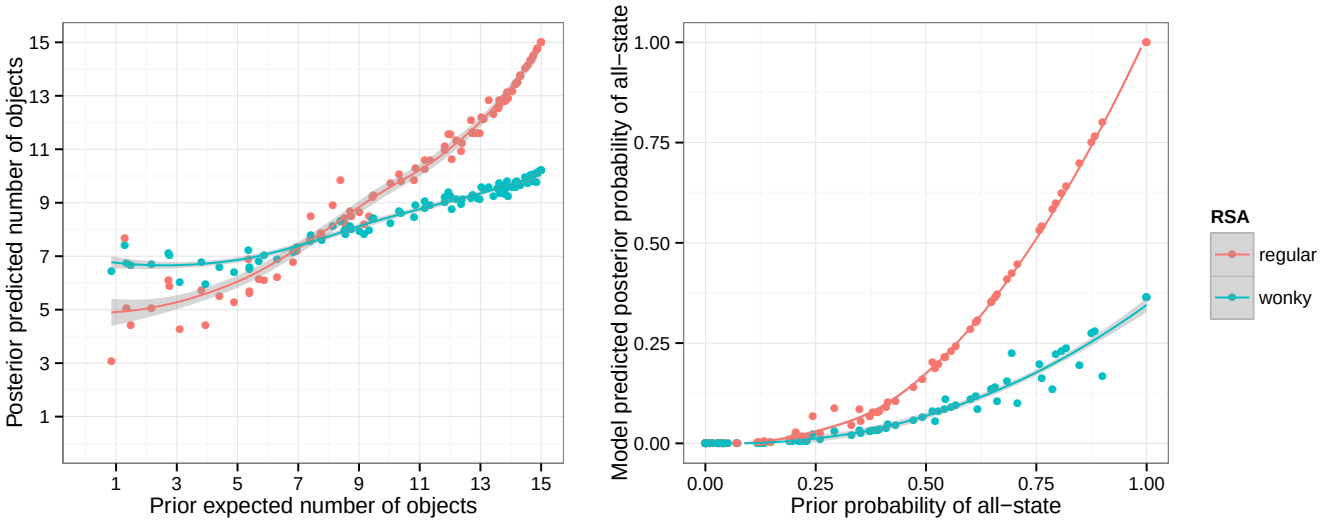

Figure 1: For each item, rRSA and wRSA model predicted $\mathbb{E}[P(s|u_{\text{some}})]$ as a function of $\mathbb{E}[P(s)]$ (left) and $P(s_{15}|u_{\text{some}})$ as a function of $P(s_{15})$.

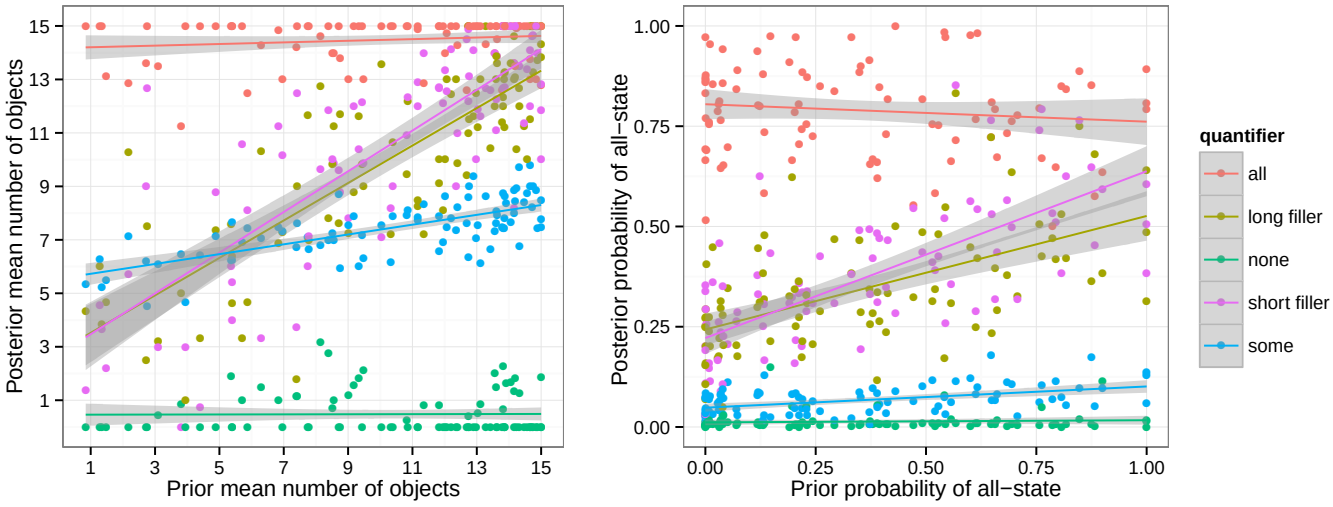

Figure 2: For each item and quantifier, empirical $\mathbb{E}[P(s|u_{\text{some}})]$ as a function of $\mathbb{E}[P(s)]$ (left, Exp. 2a) and $P(s_{15}|u_{\text{some}})$ as a function of $P(s_{15})$ (right, Exp. 2b).

they instead rated on sliding scales with endpoints labeled "definitely not" and "definitely", how likely they thought 0%, 1-50%, 51-99%, or 100% of the objects exhibited the effect.

Each participant saw 10 *some* trials and 20 filler trials, of which 10 contained the quantifiers *all* or *none*, and the rest were utterances that did not address the number of objects that displayed the effect. Of these, half were generic short fillers that were intended to communicate the prior, e.g., *Typical*. The rest addressed a different aspect of the described scenario, e.g. *What a stupid thing to do*. The utterances were randomly paired with 30 random items for each participant.

Results and discussion

Data from eight participants in Exp. 2b were excluded from the analysis because these participants assigned less than .8 probability to the interpretation corresponding to the correct literal interpretation on literal *all* and *none* trials.

The main question of interest was whether participants' judgments of how many objects exhibited the effect after hearing an utterance with *some* followed the predictions of the basic RSA model laid out in the previous section. Mean $\mathbb{E}[P(s|u)]$ and $P(s_{15}|u)$ are shown in Figure 2. For utterances of *Some of the X-s Y-ed*, the mean number of objects judged to exhibit the effect increased with increasing expectation of the prior distribution (β=.18, *SE*=.02, t=7.4, $p<.0001$). Similarly, the probability of all of the X-s Y-ing increased with increasing prior probability of all of the X-s Y-ing (β=.06, *SE*=.01, t=5.0, $p<.0001$). However, the size of this effect is astronomically smaller than that predicted by rRSA (for comparison, see Figure 1)

One possible explanation for this highly attenuated effect of the prior is that participants simply do not bring this information to bear on the interpretation of utterances. However, this possibility is ruled out by examining participants' performance in the filler conditions: in both Exps. 2a and 2b, the filer conditions closely tracked the prior (see Figure 2).

Exps. 2a and 2b demonstrate that there is an effect of listeners' prior beliefs on the interpretation of utterances with *some*. However, this effect is quantitatively much smaller than predicted by rRSA, and qualitatively does not show the critical limit effect (converging to the upper-right corner as seen in Fig. 1).

In the next section, we present an extension to rRSA that formalizes the intuition that listeners may infer that what they believed was in common ground may in fact not be; that is, that the prior beliefs they bring to bear on the utterance situation may not be the same as the speaker's.

Effect of the world prior in 'wonky RSA'

To capture the idea that the pragmatic listener is unsure what background knowledge the speaker is bringing to the conversation, we extend the basic RSA model by using a "lifted variable" (?, ?, ?, ?, ?) corresponding to the choice of state prior. That is, we posit that the prior, now $P(s|w)$, depends on a "wonkiness" variable w, which determines if it is the "usual" prior for this domain or a more generic back-off prior that we take to be uniform. The same prior is used in the literal and pragmatic listener, indicating that it is taken to be common ground. However, the w variable is reasoned about only by the pragmatic listener, which captures the idea that it is an inference the pragmatic listener makes about which communication system is relevant. Using the notation of the earlier modeling section:

$$P(s|w) = \begin{cases} P_{\text{usual}}(s) & \text{if not } w \\ \text{Uniform}(0,1) & \text{if } w \end{cases} \quad (4)$$

$$P_{\text{literal}}(s|u,w) \propto F_u(s) \cdot P(s;w) \quad (5)$$

$$P_{\text{speaker}}(u|s,w) \propto \exp(\lambda \ln P_{\text{literal}}(s|u,w)) \quad (6)$$

$$P_{\text{listener}}(s,w|u) \propto P_{\text{speaker}}(u|s,w) \cdot P(s|w) \cdot P(w) \quad (7)$$

We refer to this model as wRSA. Notice that the choice of w will depend on $p(u|s,w)$: if a given utterance can't be explained by the usual prior, because it is unlikely under any plausible world state s, then the pragmatic listener will back off to the uniform prior—inferring that the world is wonky.

To make predictions for Exp. 2 from wRSA we use the soothed empirical priors from Exp. 1 as $P_{\text{usual}}(s)$ for each item. The wonkiness prior $P(w)$ and the speaker optimality λ are fit to optimize correlation? with Exp. 2 data. point to figure, describe fit to exp 2 data.

These results are encouraging: wRSA is able to account for the qualitative and quantitative departures of participants' behavior from RSA, with respect to the effect of the prior. One could imagine various approaches to engineer this departure[4], is this actually because listeners are inferring from an utterance like "some of the marbles sank" that the world is unusual? The wRSA model makes predictions about the probability that a given world is wonky, after hearing an utterance ($P_{\text{listener}}(w|u)$); see Figure 3 for predicted wonkiness probabilities for *all*, *none*, and *some*. We can test these predictions directly by simply asking subjects whether the situation is normal.

Experiment 3

Exp. 3[5] measured participants' beliefs in world wonkiness after observing the scenarios and utterances from Exps. 2a and 2b.

Participants We recruited 60 participants over Amazon's crowd-sourcing platform Mechanical Turk.

Procedure and materials The procedure and materials were identical to those of Exps. 2a and 2b, with the exception of the dependent measure. Rather than providing estimates of what they believed the world was like, participants were asked to indicate how likely it was that the objects (e.g.,

[4]Though the authors tried several, which all failed by maintaining one or the other of the qualitative predictions of RSA identified in the introduction.

[5]This experiment can be viewed at `https://web.stanford.edu/~jdegen/17_sinking-marbles-normal-sliders/sinking-marbles-normal.html`

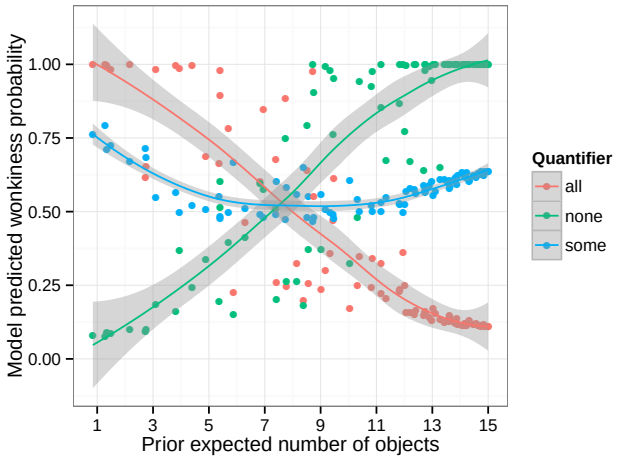

Figure 3: For each item, predicted wonkiness probability after observing an utterance (*all, none, some*), as a function of the prior expected number of affected objects.

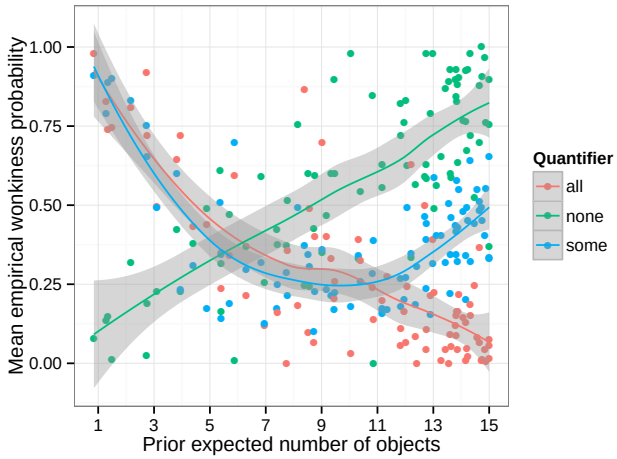

Figure 4: For each item, mean empirical wonkiness probability after observing an utterance (*all, none, some*), as a function of expected prior number of affected objects.

the marbles) involved in the scenario were normal objects, by adjusting a slider that ranged from *definitely not normal* to *definitely normal*.

Results The extreme ends of the sliders were coded as 1 (*definitely not normal*, i.e., wonky) and 0 (*definitely normal*, i.e., not wonky). We interpret the slider values as probability of world wonkiness. Mean wonkiness probability ratings are shown in Figure 4. For *all* and *none*, increasing prior expectation of objects exhibiting the effect resulted in a fairly linear decrease and increase in the probability of wonkiness, respectively. For *some*, the pattern is somewhat more intricate: probability of wonkiness initially decreases sharply, but rises again in the upper range of the prior expected value.

yay!! XXX

From the abstract: Comparing the fit of rRSA and wRSA model predictions to the posterior state estimates from Exp. 2 yields a much better fit for wRSA. The better fit of wRSA suggests that listeners use speakers' utterances as cues to how strongly to incorporate world knowledge. wRSA also provided a better fit than a model which used only a uniform prior, confirming that listeners do make use of world knowledge in a systematic way in the computation of scalar implicature.

it's possible we'd get less noise form some more stable estimator of prior. consider trying the plots with prior mode and median as x-axis.... or inferred binomial prob fit to each prior, if the fits are at all decent.

NDG: note to self – difference between revising own beliefs and revising common ground....

Discussion and conclusion

Interlocutors bring a wealth of world knowledge to bear on any language interpretation task. While effects of world knowledge in different areas of language processing are well-established (psycholinguistics refs), there has to date been a surprising lack of quantitative investigation into the role of world knowledge–and its defeasibility–in pragmatic inference. Here we have shown that listeners' world knowledge in the form of prior beliefs enters into the computation of speaker meaning in a systematic way, yet the effect of the prior on interpretation was much smaller than predicted by a regular Bayesian RSA model of quantifier interpretation, suggesting that in certain situations, listeners update their prior beliefs in the computation of speaker meaning. We have provided empirical evidence that these types of situations are cases of wonky worlds, that is, situations in which the speaker's utterance is too unlikely under the listener's prior beliefs. Extending rRSA with a lifted wonkiness variable that captures precisely whether listeners think the world is wonky and allowing them to back off to a uniform prior (i.e., ignore entirely their previously held beliefs about the world), provided a good fit to both the empirical wonkiness posteriors and dramatically improved the fit to participants' comprehension data, compared to rRSA. This model constitutes the first attempt to explicitly model the quantitative effect of world knowledge and its defeasibility on language interpretation and raises many interesting questions.

- what is wonky? – objects, event, speaker? – connection to adaptation?
- what's the right prior to back off to?
- revising private beliefs vs revising common ground.
- connection to presupposition (cf stalnaker), and other phenomena
- implication for experiments on language understanding

References

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

Geurts, B. (2010). *Quantity implicatures*. Cambridge: Cambridge Univ Press.