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 in the next section, 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). 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 all state, 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 Experiment 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 Xs Yed". 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. In Experiment 3 we explore participants' intuitions about whether the world is normal in the scenarios

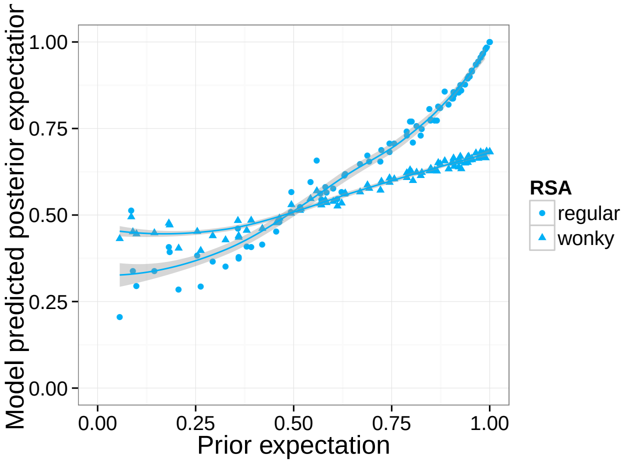

Figure 1: For each item, rRSA and wRSA model predicted mean empirical proportion of affected objects after observing *Some of the X Yed*, as a function of prior mean proportion of affected objects.

of Experiment 2. We then 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 resolves the puzzle of the prior's muted effects, and predicts people's normalcy judgements.

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:

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

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

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

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 it's usual literal meaning: $F_{u_{\text{none}}}(s){=}\delta_{s=0}$, $F_{u_{\text{some}}}(s){=}\delta_{s>0}$, $F_{u_{\text{all}}}(s){=}\delta_{s=15}$. Let us imagine that the prior is binomial with single event probability θ: $P(s) = \text{Binomial}(s|\theta)$.

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

Include also plot of all-state probability: posterior empirical and predicted probability (rRSA, wRSA) as a function of prior all-state probability, for the SI people

Experiment 1

Exp. 1 measured participants' prior beliefs about how many times different objects would exhibit a certain effect (e.g., how many marbles sink), $P(s)$.

Method

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

Procedure and materials On each trial,[1] participants read a 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. Judgments were obtained for 90 items, of which each participant saw a random selection of 30 items. should be more specific about the materials... each item had a similar form, with action, category, and outcome differing? say that we constructed them to cover the range of probabilities as much as possible.

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 nonparametric density estimation for ordinal categorical variables (?, ?) using the `np` package in R (?, ?) see hayfield 2013 np package specification for li and racine ref — or laplace if that's what we use..

say that we succeed in getting items that cover the probability range – also maybe indicate how close the distributions are to binomial?

Experiment 2a

if we end up tight on space, the expt 2a and 2b sections can be combined.

i think the current 2b should go first, since it uses the same DM as expt 1....

Exp. 2a measured participants' posterior beliefs in different objects exhibiting a certain effect (e.g., marbles sinking), after observing an utterance, $p(s|u)$.

Participants We recruited 120 participants over Amazons crowd-sourcing platform Mechanical Turk.

Procedure and materials [2]

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"*, and were asked to rate on sliding scales with endpoints labeled "very unlikely"

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

[2] This experiment can be viewed at $https : //web.stanford.edu/ \quad jdegen/sinking - marbles - nullutterance/sinking - marbles - nullutterance.html$

and “very likely”, how likely they thought 0%, 1-50%, 51-99%, or 100% of the marbles sank.

Each participant saw 10 “some” trials and 20 fillers, 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, e.g. *What a stupid thing to do.* The utterances were randomly paired with 30 random items for each participant.

Results XXX question

$p(s_\forall|u_{\text{some}})$ increased with increasing talk about both $p(s_\forall$ and the prior expectation of the distribution? $p(s_\forall)$ (β=.1, *SE*=.01, *t*=6.9, *p*<.0001); however, mean $p(s_\forall|u_{\text{some}})$ was never higher than .26, suggesting that a) participants drew strong implicatures in this paradigm and b) the effect of $P(s)$ is much smaller than predicted by rRSA.

Experiment 2b

Exp. 2b replicates the effect of the prior on participants’ posterior estimates of different objects exhibiting a certain effect (e.g., marbles sinking) using a different dependent measure.

Participants We recruited 120 participants over Amazons crowd-sourcing platform Mechanical Turk.

Procedure and materials [3]

The procedure and materials were identical to those of Exp. 2a with the exception of the dependent measure. Rather than providing point estimates of the probability of different numbers of objects sinking, participants performed the task from Exp. 1, i.e., they were 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.

Results and discussion XXX question

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, see also Figure 2), replicating the effect observed in Exp. 2a. Again, the effect of the prior was much smaller than predicted by rRSA and resulted in mean proportions of affected objects between 30% and 65%, where rRSA predicts a range from XXX to XXX for these items.

need to discuss fillers, and that this means the muted effect of prior is not because Ss are insensitive to it.

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). earlier be clear about the prediction that as prior goes to one allprob and expectation should go to one. reference that qualitative prediction of RSA here and in 2a results.

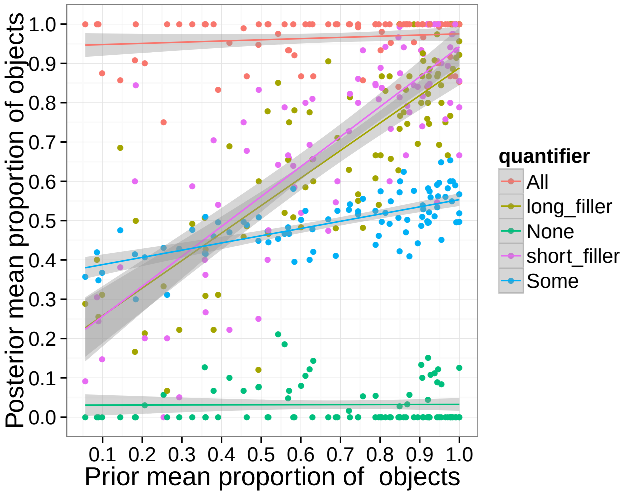

Figure 2: For each item, mean empirical proportion of affected objects after observing an utterance, as a function of prior mean proportion, for different quantifiers and filler conditions. two panels, one for 2a and one for 2b?

Discussion of why; wonky world intuition

Experiment 3

Exp. 3 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 [4]

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., 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 3. 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

[3] This experiment can be viewed at $https://www.stanford.edu/ \ jdegen/13_sinking-marbles-priordv-15/sinking-marbles.html$

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

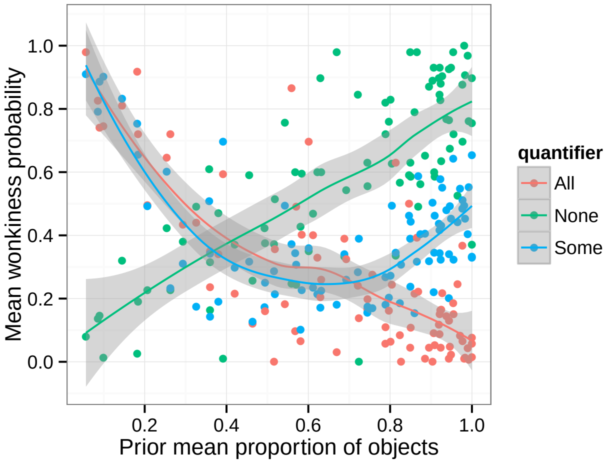

Figure 3: For each item, mean wonkiness probability after observing an utterance, as a function of expected prior proportion of affected objects, for different quantifiers.

Model

The wonky RSA model we propose for capturing the defeasibility of listeners' world knowledge is an extension of regular RSA: in wRSA, the listener infers the value of θ_{wonky} jointly with s. θ_{wonky} captures for each utterance and item, how likely the objects involved in the event (e.g., marbles) are in fact "wonky" (in which case the computation draws on a uniform prior, i.e. disregards prior beliefs) or not (in which case the model draws on the smoothed empirical prior distribution for that item, obtained in Exp. 1). The resulting $p(s|u)$ is a mixture of computations based on the uniform and empirical prior, with mixture parameter θ_{wonky}. The inferred value of θ_{wonky} itself depends on $p(u|s)$: the more surprising a particular utterance is given prior beliefs, the higher the probability of θ_{wonky}.

Model evaluation

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.

Discussion and conclusion

- what is wonky?
- other ways of asking about wonkiness

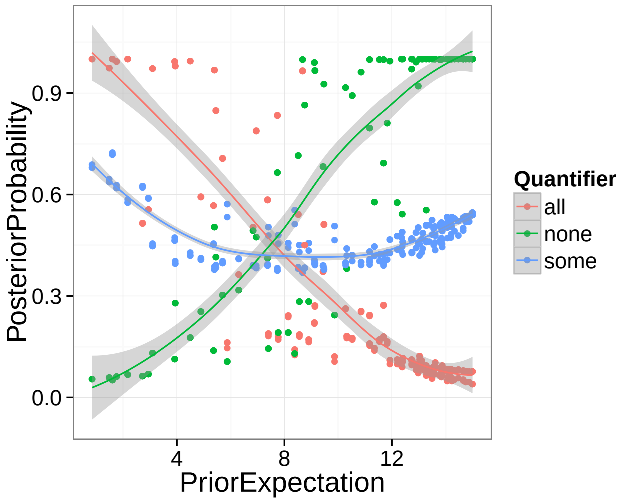

Figure 4: For each item, predicted proportion of 'wonky' ratings after observing an utterance, as a function of prior mean proportion, for different quantifiers.

- 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.