Non-sinking marbles are wonky: defeasible world knowledge in language interpretation

Judith Degen (jdegen@stanford.edu)

Michael H. Tessler (mtessler@stanford.edu)

Noah D. Goodman (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 prominent Rational Speech Act model of scalar implicature that captures whether listeners believe they are in a '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, the first quantitative model that tries to account for how rational listeners should integrate world knowledge, provides a good fit to both the comprehension and the wonkiness data.

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

How often do you think marbles would sink in water? Probably quite often, if not always. Now imagine reading "Max threw 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 enriched with something that makes them float? 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 wonky.

Models of utterance interpretation

Recent Bayesian Rational Speech Act (RSA) (Frank & Goodman, 2012) and game-theoretic (?, ?) models treat communication as a signaling game (?, ?) between a speaker and a listener, in which a listener iteratively reasons about likely utterances a speaker will produce who is trying to be informative with respect to a naïve listener. These models make clear predictions about how prior beliefs about states s of the world should be integrated with listeners' expectations about utterances u a speaker is likely to produce to communicate s. The listener's task can be characterized as having to infer $p(s|u)$, i.e., the probability of the state of the world the speaker intended to communicate, given that the speaker produced u. By Bayes' rule: $p(s|u) \propto p(u|s) \cdot p(s)$. Thus, where the semantics of the utterance u is highly constraining with respect to s, prior beliefs will not affect the interpretation of u very strongly, if at all. However, where the semantics of u is weak – that is, if u does not constrain the state space the listener is considering very strongly – prior beliefs are predicted to have a large effect on the resulting listener belief distribution over s.

However, as evident in the introductory example, prior beliefs about the world are defeasible. The currently available models of utterance interpretation either do not take into account world knowledge in the form of prior beliefs at all some refs or do not allow for their defeasibility RSA/game-theoretic refs.

Here we measure the effect of world knowledge and its defeasibility in the interpretation of utterances that contain quantifiers like *all/none/some of the marbles sank*. While the semantics of *all* and *none* are highly constraining and limit the state space to just one state (that in which either zero or all of the marbles sank, respectively), the semantics of *some* is weak and is compatible with many different states of the world, as long as more than zero marbles sank. In addition, utterance with *some* give rise to scalar implicatures. These are inferences that arise in cases of utterances like *Some of the marbles sank*, which typically give rise to the inference that not all of the marbles sank, or else the speaker would have said so.

We refer to the state in which all marbles sink as $s_\forall$ and the utterance with "some" as u_{some}. Without further modification, RSA predicts that $p(s_\forall|u_{\text{some}})$ increases with increasing $p(s_\forall)$, such that for $p(s_\forall)$ close to 1, $p(s_\forall|u_{\text{some}})$ approaches 1, that is, implicatures are very weak (see Figure XXX create). Relatedly, with increasing expected value of the prior belief distribution, so should the expected value of the posterior belief distribution increase (see Figure 1). However, for events with very high prior probability of occurrence (e.g. marbles sinking), the implicature that not all of the marbles sank is intuitively very strong, that is, $p(s_\forall|u_{\text{some}})$ is intuitively close to 0 (Geurts, 2010).

Our contribution is two-fold: first, we collect empirical estimates of $p(s)$ and $p(s|u)$ to investigate the empirical effect of listeners' prior beliefs on implicature strength. Second, we

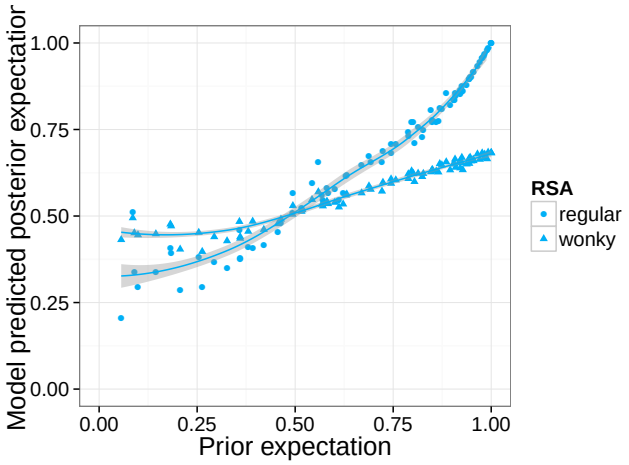

Figure 1: For each item, rRSA and wRSA model predicted mean empirical proportion of affected objects after observing *Some of the X Yed*, as a function of prior mean proportion of affected objects.

extend the RSA model to incorporate a free variable θ_{wonky} that captures the extent to which the listener believes the described event is abnormal or 'wonky' and she should thus discount her prior beliefs when interpreting u. We refer to this model as *wonky RSA* (wRSA) in contrast to *regular RSA* (rRSA). Wonkiness inferences in wRSA are triggered by the surprisal of a produced utterance u, given listeners' prior beliefs, capturing that listeners expect speakers' utterances to be both truthful and informative with respect to prior beliefs. To the extent that they are not, listeners will have to either infer that the speaker is being uncooperative, or else assume that they may need to revise their beliefs about the world. Here we pursue the latter possibility.

This paper is structured as follows. We first report the results of three experiments (1, 2a, 2b) that show that while there is an effect of the prior on listeners' interpretations of sentences like *Some of the marbles sank*, this effect is much smaller than predicted by rRSA. Exp. 3 provides evidence that listeners' beliefs about object or event wonkiness are indeed influenced by the surprisal of the utterance. Finally, we present an extension of rRSA that incorporates the idea of backing off to alternate prior beliefs if the observed utterance indicates a wonky world. This model provides a much better fit to the empirical data from Exps. 2a and 2b, and also provides a good fit to the wonkiness ratings obtained in Exp. 3.

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 $p(s)$, participants' prior beliefs about different objects exhibiting a certain effect (e.g., marbles sinking).

Method

Participants We recruited 60 participants over Amazons crowd-sourcing platform Mechanical Turk.

Procedure and materials On each trial, 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.

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.

Experiment 2a

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

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

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.

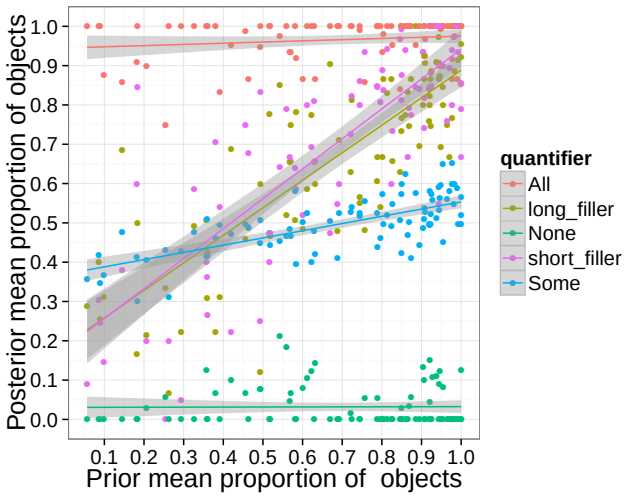

Figure 2: For each item, mean empirical proportion of affected objects after observing an utterance, as a function of prior mean proportion, for different quantifiers and filler conditions.

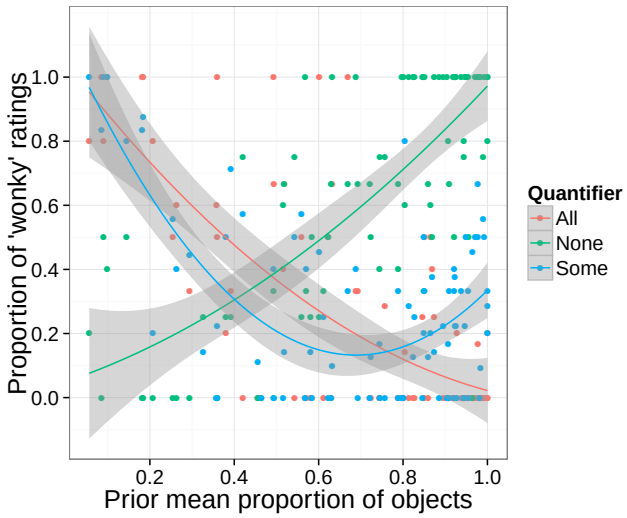

Figure 3: For each item, empirical proportion of 'wonky' ratings after observing an utterance, as a function of prior mean proportion, for different quantifiers.

Procedure and materials 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.

Exps. 2a and 2b demonstrate that there is a robust effect of listeners' prior beliefs on the interpretation of utterances with *some*. However, this effect is much smaller than predicted by rRSA.

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 Amazons 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 whether they believed the objects (e.g., the marbles) involved in the scenario were normal objects, by clicking a 'Yes' or 'No' radio button.

Results Proportion of 'No' ratings (where we take 'No' ratings to indicate participants belief in wonkiness of the world) for each item and quantifier are shown in Figure 3. For 'all' and 'none', increasing prior expectation of objects exhibiting the effect results in a monotonic 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 prior expectation range.

yay!! XXX

Model

continue exploring priors before reporting

From the CUNY abstarct: 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}.

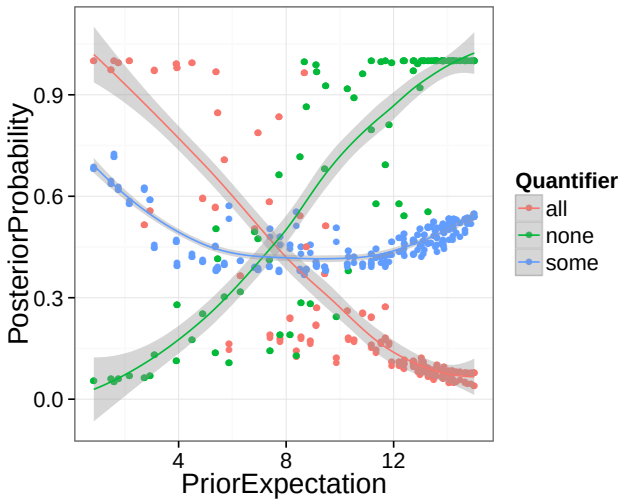

Figure 4: For each item, predicted proportion of 'wonky' ratings after observing an utterance, as a function of prior mean proportion, for different quantifiers.

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.

General discussion

- what is wonky?
- other ways of asking about wonkiness
- what's the right prior to back off to?

Conclusion

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.