Wonky marbles, or: everyone's ongoing quest for the right prior

Judith Degen, Michael H. Tessler, and Noah D. Goodman (Stanford University)

World knowledge enters into the interpretation of utterances in complex ways. While effects of world knowledge on syntactic and semantic processing are well-established (McRae et al., 1998; Chambers et al., 2004; Hagoort et al., 2004), there is to date a surprising lack of systematic investigation into the effect of world knowledge in pragmatics. Here, we provide a quantitative model of, specifically, the situational *revision* of world knowledge in the face of unexpected utterances like *Some of the marbles sank*.

Recent Bayesian Rational Speech Act (RSA) models of scalar implicature (Goodman & Stuhlmüller, 2013; Degen & Goodman, 2014) make clear predictions about how world knowledge in the form of prior beliefs about states of the world *s* 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 the most likely state of the world given the speaker's utterance, or $p(s|u)$. By Bayes' rule: $p(s|u) \propto p(u|s)p(s)$. We refer to the state in which none, one, …, all marbles sink as $s_0, s_1, ..., s_\forall$ and utterances of the form *Some of the X Y-ed* as u_{some}. Without further modification, RSA makes two qualitative predictions (Fig. 1): for u_{some},

1. the expected value of the posterior distribution should roughly track that of the prior distribution for Xs that a priori are unlikely to not Y; that is, the mean number of marbles expected to have sunk should be very similar before and after observing *Some of the marbles sank.* This is because the main effect of u_{some} is to redistribute the probability mass that was a priori assigned to s_0 over the remaining states.

2. the posterior probability $p(s_\forall|u_{\text{some}})$ increases with increasing prior probability $p(s_\forall)$, such that for $p(s_\forall)$ close to 1, $p(s_\forall| u_{\text{some}})$ approaches 1 (i.e., implicatures are very weak).

These predictions are at odds with an observation in Geurts, 2010: for events with very high prior probability of occurrence (e.g., marbles sinking), the implicature that not all of the marbles sank upon observing u_{some} is intuitively very strong, that is, $p(s_\forall|u_{\text{some}})$ should be close to 0.

Our contribution is three-fold: first, we collect empirical estimates of $p(s)$ and $p(s|u)$ to investigate the empirical effect of participants' prior beliefs on implicature strength; in particular, we test whether the above two predictions are borne out. Second, we extend RSA to incorporate a free variable θ_w that captures the extent to which the listener believes the described event is abnormal and she should thus discount her prior beliefs (world knowledge presumed to be in common ground) when interpreting *u*. We refer to this model as *wonky RSA* (wRSA) in contrast to *regular RSA* (rRSA) and collect empirical judgments of world wonkiness. Third, we discuss the importance and challenge of collecting reliable estimates of participants' prior beliefs to the general enterprise of integrating facts about listeners' world knowledge into formal models of pragmatic reasoning.

Model. In wRSA, the listener infers the value of θ_w jointly with *s*. θ_w 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_w. The inferred value of θ_w itself depends on $p(u|s)$: the more surprising a particular utterance is given prior beliefs, the higher the probability of θ_w.

Exp. 1 (n=60) measured $p(s)$ for 90 items (of which each participant saw one third). 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 how many Xs Y-ed, e.g. *How many of the marbles do you think sank?*, on a sliding scale from 0 to 15.

Exp. 2a (n=120, Fig. 2, left) collected participants' posterior estimates of the expected mean number of Xs that Y-ed. Participants performed the same task as in Exp. 1, but 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"*. 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*. Filler trials were included to assess whether participants are in principle tracking information about the prior. There was a main effect of the expected value of the prior on participants' interpretation (β =.18, SE=.02, t=7.4, p<.0001), but the effect was much smaller than predicted by rRSA.

Exp. 2b (n=120, Fig. 2, right) collected participants' posterior estimates of p(s|u). The procedure and materials were the same as in Exp. 2a, but participants' task was different, in order to directly evaluate the effect of the prior on scalar implicature strength: they 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. p($s_{\forall}$|u_{some}) increased with increasing p($s_{\forall}$) (β =.1, SE=.01, t=6.9, p<.0001); however, mean p($s_{\forall}$|u_{some}) was never higher than .26, suggesting that a) participants drew strong implicatures in this paradigm and b) the effect of the prior p(s) is again much smaller than predicted by rRSA.

Comparing the fit of rRSA and wRSA model predictions to the posterior state estimates from Exp. 2 yields a much better fit for wRSA (see Figures 1 and 2), suggesting that listeners use speakers' utterances as a cue 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 speaker meaning. Further support for wRSA comes from an additional experiment (n=60), in which the same items from Exps. 2a and 2b were presented, but participants were asked to rate how likely it is that the Xs involved in the event were "normal" (or, by inversion, "wonky"). Participants' wonkiness judgments for *some*, *all*, and *none* qualitatively reflected the model's predictions.

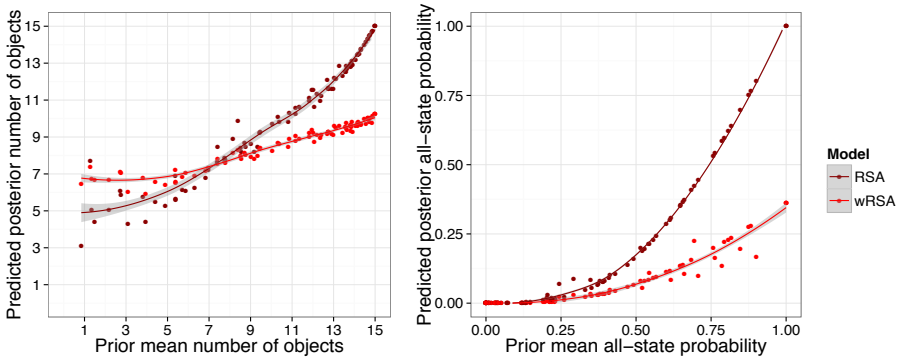

Figure 1. For each item in Exps. 1-2, model predictions for expected value of distribution (left) and posterior probability of all-state (right) after observing *Some of the X Y-ed*.

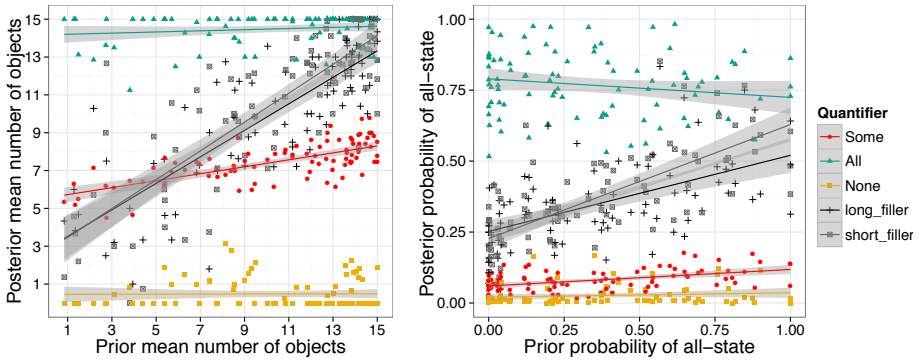

Figure 2. For each item in Exps. 1-2, empirical mean number of objects (Exp. 2a, left) and posterior probability of all-state (Exp. 2b, right).

Naturally, obtaining good estimates of participants' world knowledge is crucial to this enterprise, and the inability to know a priori how participants' underlying beliefs will surface in different dependent measures presents a huge methodological challenge. We discuss a variety of dependent measures and analysis methods that we have employed in estimating priors, and the differences between them.