Non-sinking marbles are wonky: world knowledge in scalar implicature computation

Judith Degen and Noah Goodman (Stanford University)
jdegen@stanford.edu

World knowledge enters into the interpretation of utterances in complex ways. While effects of world knowledge on syntactic processing are well-established, there is to date a surprising lack of systematic investigations into the effect of world knowledge in pragmatics. Here, we provide a quantitative model of the effect of world knowledge on scalar implicatures, which are inferences that arise in cases of utterances like *Some of the marbles sank*, which gives rise to the scalar inference that not all of the marbles sank.

Recent Bayesian Rational Speech Act (RSA) models of scalar implicature (Frank and Goodman, 2012) make clear predictions about how world knowledge in the form of 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 $p(s|u)$. By Bayes' rule: $p(s|u) \propto p(u|s)p(s)$. 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 $p(s_\forall|u_{some})$ increases with increasing $p(s_\forall)$, such that for $p(s_\forall)$ close to 1, $p(s_\forall| u_{some})$ approaches 1 (i.e., implicatures are very weak). However, for events with very high prior probability of occurrence (e.g. sinking marbles), the implicature that not all of the marbles sank is intuitively very strong, that is, $p(s_\forall|u_{some})$ 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 participants' prior beliefs on implicature strength. 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 when interpreting *u*. We refer to this model as *wonky RSA* (wRSA) in contrast to *regular RSA* (rRSA).

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 of an effect, e.g. *How many of the marbles do you think sank?*, on a sliding scale from 0 to 15.

Exp. 2 (n=120) collected participants' posterior estimates of $p(s|u)$. Participants read the same descriptions as in Exp. 1 and 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*. $p(s_\forall|u_{some})$ 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 $p(s)$ is 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, which 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 speaker meaning.