Reevaluating Pragmatic Reasoning in Web-based Language Games

Les Sikos, Noortje Venhuizen, Heiner Drenhaus, and Matthew W. Crocker (Saarland University)
sikos@coli.uni-saarland.de

Recent work testing formalizations of Gricean maxims [1] using web-based reference games has led to mixed results. Some studies indicate that Bayesian (e.g., rational speech act (RSA)) models closely predict human (pragmatic) behavior [e.g., 2], while others suggest that participants rarely go beyond the literal meanings of words in such studies [e.g., 3-4]. For instance, [2] presented participants with three objects (Fig1) in 7 different context types. Using a one-shot paradigm (each participant sees a single trial), they collected separate judgments from speakers, listeners, and for salience. Results of the RSA model, which combines a speaker model (likelihood that speakers use a particular word to refer to the target) with empirically measured salience (Eq1), were highly correlated with aggregate listener judgments (Fig1d; R=0.99). This was interpreted as indicating that participants reasoned pragmatically in this task. However, the reasoning required in [2] ranged from simple (e.g., Fig1b) to more complex (e.g., Fig1c), such that the close fit of predicted to observed results might be driven by the simpler inferences. Consistent with this possibility, [3] attempted a close replication of [2], focusing on more challenging items like Fig1c, and found that the basic RSA model was a poor predictor of their data. Furthermore, [4] found that while listeners responded pragmatically in conditions similar to Fig1b, they were only at chance for conditions similar to Fig1c.

To account for these results, [3] and [4] proposed various modifications to RSA (e.g., adding parameters for speaker/listener degree of rationality). Here, we investigate another possibility: Listeners in such web-based tasks may not reason as pragmatically as presumed. Instead, they may simply interpret the utterance based on a combination of its literal meaning and the salience of particular referents. In other words, a simpler rather than more complex model may better explain human behavior than RSA. To test this hypothesis we employed the same general methods as [2] and systematically explored a wider variety of context types (34 in total). 3387 participants recruited via Amazon Mechanical Turk were randomly assigned to Speaker (N=1143), Listener (N=1111), and Salience (N=1133) tasks (Fig2). We then compared observed responses to predictions from the basic RSA model and a Literal Listener (LL) model that does not incorporate a model of the speaker. This basic LL model predicts that listeners should be equally likely to select any referent that a given word (e.g. “green”) can refer to. In order to provide a more direct comparison to RSA, which relies heavily on salience, we also tested a LL+Salience model that weights its probabilities based on salience (Eq2). For completeness, we also considered an RSA model that assumes uniform salience, and salience values alone.

Table 1 and Fig2d show that while RSA provided a good fit to the entire dataset (replicating [2]), both LL models performed better. Furthermore, when we analyzed only the contexts for which the predictions from RSA and LL+Salience models differed (i.e, the more challenging inferences), LL+Salience performed best (Table 2 bottom; Fig2e). In fact, salience alone was a better predictor than RSA. Moreover, comparing RSA and RSA-uniform-salience models suggests that salience essentially corrects for incorrect predictions in the basic RSA model. To the extent that one-shot web-based experiments accurately elicit the depth of pragmatic reasoning seen in typical human interactions, these findings indicate that a simpler model than RSA can better explain human behavior.

a. **Speaker Task.** Imagine you are talking to someone and you want him to refer to the middle object. Which word would you say, “green” or “circle”?

Listener Task. Imaging someone is talking to you and uses the word “green” to refer to one of these objects. Which object are they talking about?

Salience Task. Imaging someone is talking to you and uses a word you don’t know to refer to one of the objects. Which object are they talking about?

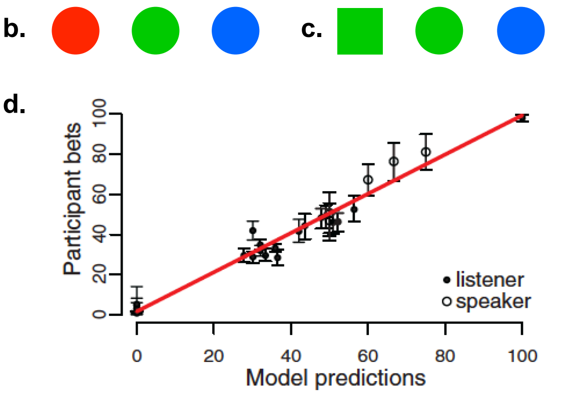

Fig 1. Overview of [2]. (a) Instructions. (b) Simple inference required. (c) Complex inference required. (d) RSA model predictions plotted against observed results.

$$\overbrace{P(r_s|w,C)}^{\text{Listener model}} = \frac{\overbrace{P(w|r_s,C)}^{\text{Speaker model}}\,\overbrace{P(r_s)}^{\text{Salience}}}{\sum\limits_{r'\in C} P(w|r',C)P(r')}$$

Eq 1. RSA model for inferring the speaker’s intended referent r_S in context *C*, given speaker’s uttered word *w*.

$$\overbrace{P(r_s|w,C)}^{\text{Listener model}} = \begin{cases} \overbrace{\frac{P(r_s)}{\sum_{r'\in R} P(r')}}^{\text{Salience}} & \text{if } r_s \in R \\ 0 & \text{otherwise} \end{cases}$$

$$\text{where: } R = \{r \in C \mid w \text{ can refer to } r\}$$

Eq 2. LL + salience model provides a distribution over the set of referents in context *C* that can be referred to with word *w*, weighted based on salience.

Table 1. Overall model fits (ranked in order of best fit).

	R	Adj R sq	t	p	
LL + Salience	0.89	0.79	16.36	0.00	***
LL uniform salience	0.88	0.77	15.52	0.00	***
RSA	0.87	0.75	14.44	0.00	***
RSA uniform salience	0.83	0.68	12.28	0.00	***
Salience alone	0.29	0.07	2.57	0.01	*

References

[1] Grice (1975). [2] Frank & Goodman (2012).

[3] Qing & Franke (2015). [4] Frank et al (2017).

a. **Speaker Task**. Imagine you are talking to Robert and you want him to pick out Item B. If you can only use one word, which word would you say, “green” or “fish”?

Listener Task. Robert wants you to pick one of the objects below but he can only say one word. He says, ”green". Which object do you think he is talking about: A, B, or C?

Salience Task. Robert wants you to pick one of the objects below, but due to background noise you cannot understand what he said. Which object do you think he is most likely talking about: A, B, or C?

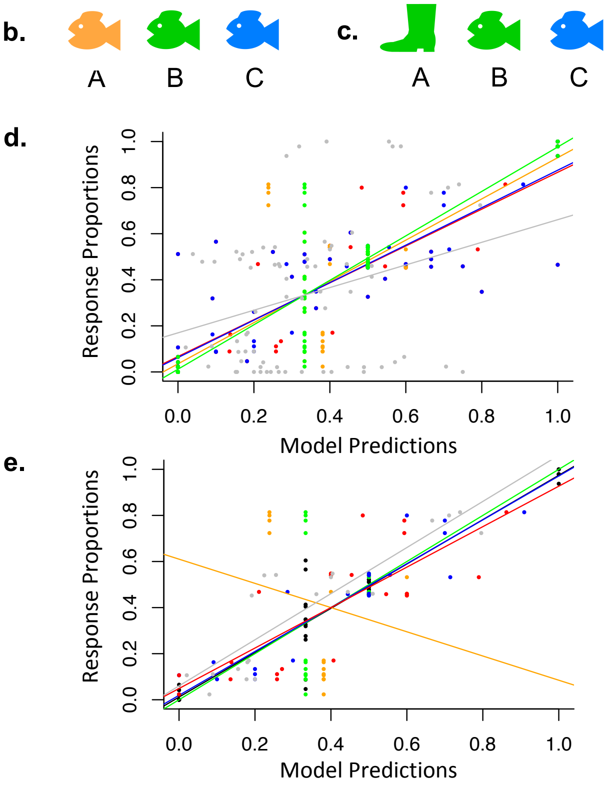

Fig 2. Overview of current study. (a) Instructions. (b) Simple inference required. (c) Complex inference required. (d) Predictions vs observed results over all visual contexts. (e) Predictions vs observed results for visual contexts in which models had identical predictions (black), and contexts in which model predictions differed. RSA, LL + salience, LL with uniform salience, RSA with uniform salience, Salience alone.

Table 2. Model fits for contexts in which models had identical predictions (top) and different predictions (bottom).

	R	Adj R sq	t	p	
Same prediction	0.97	0.94	28.56	0.00	***
Different predictions:					
LL + Salience	0.93	0.86	10.65	0.00	***
Salience alone	0.89	0.77	8.13	0.00	***
RSA	0.79	0.61	5.49	0.00	***
LL uniform salience	0.31	0.05	1.40	0.18	
RSA uniform salience	-0.23	0.00	-1.02	0.32	