Reevaluating Pragmatic Reasoning in Web-based Language Games

Les Sikos, Noortje Venhuizen, Heiner Drenhaus, and Matthew W. Crocker (Saarland University)

Background

Tests of formalizations of Gricean maxims using web-based reference games have led to mixed results:

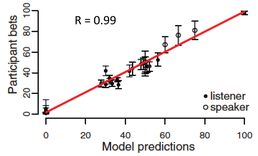
Frank & Goodman [2]

- One-shot paradigm
- 3 object displays in 7 visual context types
- Collected separate judgments from Speakers, Listeners, and for Salience

  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?

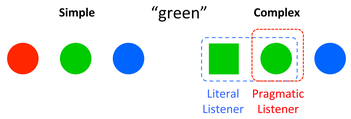
  Salience Task. Imaging someone is talking to you and uses a word you don't know to refer to one of the objects. Which object are they talking about?
- Rational Speech Act (RSA) model (Eq 1) closely predicted aggregate listener judgments



- This result was interpreted as indicating that participants reasoned pragmatically in this task

Alternative Explanation

- The reasoning required ranged from simple to more complex



- 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 and found that the basic RSA model was a poor predictor of their data
- [4] found that while listeners responded pragmatically in simpler contexts, they were at chance for more complex contexts
- To account for these results, [3] and [4] proposed various modifications to the RSA model (e.g., adding parameters for speaker/ listener degree of rationality)

→ **Participants rarely go beyond the literal meanings of words in such studies**

Goal and Methods

Goals

- Do Listeners in such tasks reason as pragmatically as presumed?
- Or can a simpler rather than more complex model better explain human behavior than RSA?

Participants

- 4642 participants recruited via Amazon Mechanical Turk
- 1137 excluded for identifying as non-native English (24%)
- 118 excluded for incorrect attention questions (3%)
- Remaining 3387 were randomly assigned as:

  Speaker N = 1143
  Listener N = 1111
  Salience N = 1133
- Demographics:

  Gender F 53%
  M 48%

  Age 18-25 17%
  26-35 40%
  36-45 22%
  46-55 13%
  56-65 6%
  over 65 2%

Procedure

Introduce Interlocutor

Attention Check

Task Response

Likelihood Judgment

Exit Survey

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"?

The target was always object B

Order of options were counterbalanced for shape/color first

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?

Given words were counterbalanced for shape/color

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?

Stimuli – 3-object displays in 17 visual context types

Number of objects that share target shape/color	Competitor objects have different/same shape	Competitor objects have different/same color	Example (target in middle)	
1s1c	ds	dc		
1s1c	ds	sc		
1s1c	ss	dc		
1s1c	ss	sc		
1s2c	ds	dc		
1s2c	ss	dc		Δc
1s3c	ds	sc		
1s3c	ss	sc		Δc
2s1c	ds	dc		
2s1c	ds	sc		Δs
2s2c.a	ds	dc		
2s2c.b	ds	dc		Δcs
2s3c	ds	sc		Δc
3s1c	ss	dc		
3s1c	ss	sc		Δs
3s2c	ss	dc		Δs
3s3c	ss	sc		

Δ - models make different predictions for Listener response, given a color (c) or shape (w) word

Analysis – We compared observed Listener responses to predictions from 4 models, as well as to observed Salience responses

- Basic RSA
- Basic RSA with uniform salience (RSA.us)
- Literal Listener + Salience (LL+S)
- LL with uniform salience (LL.us)
- Salience observations

Models

$$P(r_s|w,C) = \frac{P(w|r_s,C)P(r_s)}{\sum_{r' \in C} P(w|r',C)P(r')}$$

(Listener model; Speaker model; Salience)

Eq 1. RSA model for inferring the speaker's intended referent r_s in context C, given speaker's uttered word w. The Listener model combines a Speaker model (i.e. the likelihood that speakers use a particular word to refer to the target) with empirically measured salience.

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

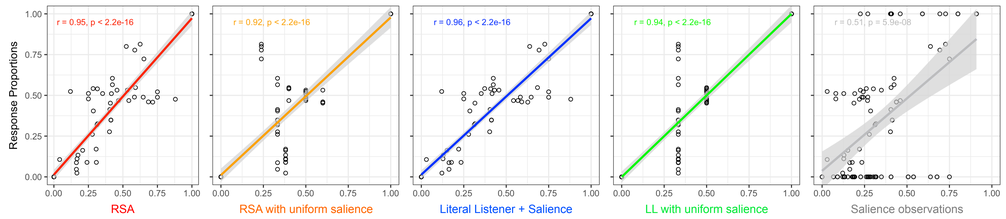
where: $R = \{r \in C \mid w \text{ can refer to } r\}$

(Listener model; Salience)

Eq 2. Literal Listener + Salience model. This model does not contain a speaker model. Instead, it provides a distribution over the set of referents in context C that can be referred to with word w, weighted based on empirically measured salience.

Results

Predictions vs Observed over All Visual Context Types



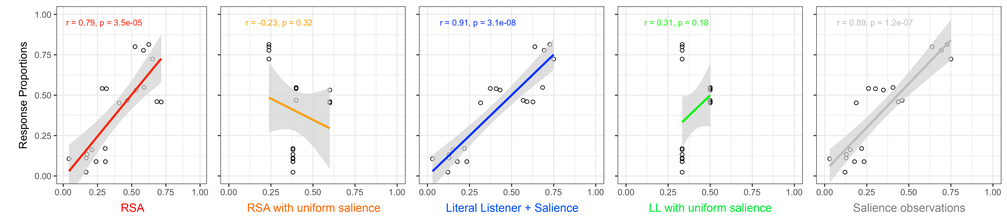
Correlations

Rank	Model	R	Adj R sq	*t*	*p*	
1	LL+S	0.964	0.929	36.429	< .0001	***
2	RSA	0.954	0.909	31.805	< .0001	***
3	LL.us	0.944	0.889	28.488	< .0001	***
4	RSA.us	0.918	0.84	23.085	< .0001	***
5	Salience	0.506	0.248	5.862	< .0001	***

Kullback–Leibler Divergence (base 2)

Rank	Model	summed KLD	mean KLD
1	LL+S	0.954	0.028
2	RSA	1.474	0.043
3	LL.us	4.166	0.123
4	RSA.us	5.466	0.161
5	Salience	6.418	0.189

Predictions vs Observed for Visual Contexts in which Model Predictions Differed (Δ)



Correlations

	Rank	Model	R	Adj R sq	*t*	*p*	
Contexts for which Models make **Same Predictions**	1	RSA.us	0.988	0.976	57.184	< .0001	***
	1	LL.us	0.988	0.976	57.184	< .0001	***
	2	RSA	0.970	0.940	35.727	< .0001	***
	2	LL+S	0.970	0.940	35.727	< .0001	***
	3	Salience	0.448	0.190	4.478	< .0001	***
Contexts for which Models make **Different Predictions**	1	LL+S	0.908	0.815	9.209	< .0001	***
	2	Salience	0.893	0.786	8.410	< .0001	***
	3	RSA	0.790	0.603	5.460	< .0001	***
	4	LL.us	0.313	0.048	1.400	0.178	n.s.
	5	RSA.us	-0.233	0.002	-1.018	0.322	n.s.

Kullback–Leibler Divergence (base 2)

Rank	Model	summed KLD	mean KLD
1	RSA	0.492	0.019
1	LL+S	0.492	0.019
2	RSA.us	1.055	0.041
2	LL.us	1.055	0.041
3	Salience	5.728	0.220

Rank	Model	summed KLD	mean KLD
1	LL+S	0.462	0.058
2	Salience	0.691	0.086
3	RSA	0.983	0.123
4	LL.us	3.111	0.389
5	RSA.us	4.411	0.551

Discussion and Conclusions

Results were consistent with the alternative hypothesis

- Although RSA had good fit to entire dataset (replicating [2]), LL+S performed better
- When considering only contexts for which predictions from RSA and LL models differed (i.e. the more challenging inferences):

  (a) LL+S performed best (b) Salience alone was a better predictor than RSA
- Comparing RSA and RSA.us models suggests salience essentially corrects for incorrect predictions in the basic RSA model
- Modified RSA models (*ala* [3, 4]) performed worse than the basic RSA model

Conclusions

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

Next Steps

- Increase engagement and cooperativity
- Allow multiple words and investigate ordering effects

References [1] Grice (1975). *Logic and conversation*. [2] Frank & Goodman (2012). Predicting pragmatic reasoning in language games. *Science*. [3] Qing & Franke (2015). Variations on a Bayesian theme: Comparing Bayesian models of referential reasoning. In *Bayesian natural language semantics and pragmatics*. [4] Frank, Emilsson, Peloquin, Goodman & Potts (2016). Rational speech act models of pragmatic reasoning in reference games. Preprint: osf.io/f9y6b.