Supplementary materials

The interactions of rational, pragmatic agents lead to efficient language structure and use

1 Rational Speech Act theory speaker and listener agents

Rational Speech Act theory (RSA) is a recursive Bayesian model of pragmatic language use, which can be seen as a mathematical formalization of essential Gricean principles. RSA has proven to be a productive framework for modeling a range of pragmatic phenomena in both language production and language understanding including hyperbole, metaphor, implicature and others (see Goodman & Frank, 2016 for review).

In the RSA framework, a "speaker agent" defines a conditional distribution, mapping meanings $m \in M$ to utterances $u \in U$, written as $S(u|m)$. We consider a prior over utterances $P(U)$ as well as a prior over meanings $P(M)$. A "listener agent" defines a conditional distribution mapping from utterances to meanings, written as $L(m|u)$. To capture recursive reasoning between interlocutors, these functions are mutually defined. That is,

$$S_i(u|m) \propto e^{-\alpha \times U(u;m)} \tag{1}$$

where

$$U(u;m) = -log(L_{i-1}(m|u)) - cost(u) \tag{2}$$

and

$$L_{i-1}(m|u) \propto S_{i-1}(u|m) \times p(m) \tag{3}$$

Defining nested speaker and listener agents could, in principle, lead to infinite regress. RSA defines a *literal listener*, denoted $L_0(m|u)$, as a base-case. The literal listener does not reason about a speaker model, rather this agent considers the literal semantics of the utterance.

$$L_0(m|u) \propto \delta_u(m) \times p(m) \tag{4}$$

with

$$\delta_u(m) = \begin{cases} 1, & \text{if } m \in [[u]] \\ 0, & \text{else} \end{cases} \tag{5}$$

where $[[u]]$ indicates a set of meanings, the denotation of u.

2 Zipfian objective for linguistic system efficiency

2.1 Basic objective derivation

Zipf (1949) proposed that the particular distributional properties found in natural language emerge from competing speaker and listener pressures. We operationalize this objective in equation (1) – the efficiency of a linguistic system ℓ being used by speaker and listener agents S and L is the sum of the expected speaker and listener effort to communicate over all possible communicative events $e \in E$. We assume a communicative event e is composed of an utterance-meaning-context triple ($e =< u, m, c >$)

$$\begin{aligned} \text{Efficiency}(S, L, \ell) = \mathbb{E}_{e \sim P(E)}[\text{speaker effort}] \\ + \mathbb{E}_{e \sim P(E)}[\text{listener effort}] \end{aligned} \tag{1}$$

We assume that speaker effort is related to the surprisal of an utterance in a particular context – intuitively, the number of bits needed to encode the utterance u.[1] This particular formalization of speaker-cost is general enough to accommodate a range of cost instantiations, such as production difficulty via articulation effort, cognitive effort related to lexical access, or others (Bennett & Goodman, 2015).

$$\text{speaker effort} = -log_2(p(u|c))$$

We assume listener effort is the surprisal of a meaning given an utterance. This operationalization of listener effort is intuitively related to existing work in sentence processing in which word comprehension difficulty is proportional to surprisal (Hale, 2001; Levy, 2008).

$$\text{listener effort} = -log_2(L(m|u, c; \ell))$$

Rewriting (1) we have

$$\text{Efficiency}(S, L, C, \ell) = \mathbb{E}_{e \sim P(E)}[-log_2(p(u|c))] + \mathbb{E}_{e \sim P(E)}[-log_2(L(m|u, c; \ell))] \tag{2}$$

We assume that the particular joint distribution over utterance-meaning-context triples $e =< u, m, c >$ follows from a simple generative model: First, some context is sampled with probability $p(c)$. Then some meaning is sampled with probability $p(m|c)$. Our speaker attempts to convey this intended meaning to a listener via an utterance u by sampling from the speaker conditional distribution $S(u|m, c; \ell)$. This allows us to re-write the objective as:

$$\begin{aligned} = -\sum_{u,m,c} P_{speaker}(u, m|c; \ell)p(c)[log_2(p(u|c))] - \\ \sum_{u,m,c} P_{speaker}(u, m|c; \ell)p(c)[log_2(L(m|u, c; \ell))] \end{aligned} \tag{3}$$

$$= -\sum_{c \in C} p(c)(\sum_{u,m} P_{speaker}(u, m|c; \ell)[log_2(p(u|c))] + \sum_{u,m} P_{speaker}(u, m|c; \ell)[log_2(L(m|u, c; \ell))]) \tag{4}$$

$$= -\sum_{c \in C} p(c) \sum_{u,m} P_{speaker}(u, m|c; \ell)[log_2(L(m|u, c; \ell)p(u|c)] \tag{5}$$

[1]In the current set of simulations we consider utterances costs as independent from context (i.e. $p(u|c) = p(u)$). Hence surprisal can be thought of more simply as cost.

Note that $P_{listener}(u, m|c; \ell) = L(m|u, c; \ell)p(u|c)$:

$$= -\sum_{c \in C} p(c) \sum_{u,m} P_{speaker}(u, m|c; \ell)[log_2(P_{listener}(u, m|c; \ell))] \tag{6}$$

The inner summation of (6) is the cross-entropy between speaker and listener conditional distributions over utterance-meaning pairs.

$$= -\sum_{c \in C} p(c) H_{cross}(P_{speaker}(u, m|c; \ell), P_{listener}(u, m|c; \ell)) \tag{7}$$

This final form is simply an expectation of speaker-listener cross-entropy over contextualized language use.

$$= \mathbb{E}_{c \sim P(C)}[H_{cross}(P_{speaker}, P_{listener})] \tag{8}$$

Note that in the case that $|C| = 1$, our objective simplifies to a simple Cross-Entropy between speaker-listener joint distributions over utterance-meaning pairs.

$$= H_{cross}(P_{speaker}, P_{listener}) \tag{9}$$

From an information-theoretic perspective this objective is intuitive: H_{cross} denotes the Cross-Entropy (CE), a measure of dissimilarity between two distributions – the average number of bits required to communicate under one distribution, given that the “true” distribution differs. In our case, we have an expectation over this term – the expected difference between the distributions assumed by the speaker $P_{speaker}$ and listener $P_{listener}$ given a set of contexts C. In other words, an ”efficient” language ℓ minimizes the distance between what speakers and listeners think.

2.2 Baseline model objectives

For comparison, we also examine properties of optimal languages under two additional objectives. Zipf (1949) proposed that the optimal speaker language $\ell^*_{speaker}$ should only optimize speaker effort. We operationalize this using the first half of equation (1) in Section 2.1.

$$\ell^*_{speaker} = argmin_{\ell \in L} \mathbb{E}_{c \sim P(C)}[\mathbb{E}_{P_{speaker}(u,m|c;\ell)}(-log_2(p(u|c)))] \tag{1}$$

The optimal listener language $\ell^*_{listener}$, by contrast, should only optimize listener effort. We operationalize this using the second half of equation (1) in Section 2.1.

$$\ell^*_{listener} = argmin_{\ell \in L} \mathbb{E}_{c \sim P(C)}[\mathbb{E}_{P_{speaker}(u,m|c;\ell)}(-log_2(L(m|u, c; \ell))] \tag{2}$$

3 Simulation 2

3.1 Updates to RSA speaker-listeners

We consider the same model of basic speakers and listeners (S_{vanilla}, L_{vanilla}) as in Section 1. We introduce discourse aware speaker-listeners ($S_{\text{discourse}}$, $L_{\text{discourse}}$) who can use the history of utterances (the discourse D) to infer the topic of conversation ($c \in C$):

$$S_{\text{discourse}}(u|m, c, D) \propto e^{\alpha U(u,c;,m,D)}$$
$$U(u, c; , m, D) = -log_2(L_{\text{discourse}}(m, c|u, D)p(c|D)) - cost(u)$$

where

$$p(c|D) \propto p(c) \prod_{i=0}^{|D|} S_{\text{vanilla}}(u_i|m_i)p(m_i|c)$$

and

$$L_{\text{discourse}}(m, c|u, D) \propto S_{\text{vanilla}}(u|m)p(m|c)p(c|D)$$

Note that $p(M|C = c)$ is simply the particular prior over meanings dictated by a topic c.

3.2 Language used in Simulation 2

We conduct $N = 600$ simulations, generating discourses of length $|D| = 30$ utterances with three different speaker models ($n = 200$ each). We consider a single language ℓ with $|U| = 6$ and $|M| = 4$ specified by the boolean matrix below. (Note that use of this particular language is not essential – the results are broadly generalizable languages that contain ambiguity.)

	m_1	m_2	m_3	m_4
u_1	1	0	0	0
u_2	0	1	0	0
u_3	0	0	1	0
u_4	0	0	0	1
u_5	1	1	0	0
u_6	0	0	1	1

We assume that $p(u_5) = p(u_6) > p(u_1) = \cdots = p(u_4)$. That is, the two ambiguous utterances (u_5 and u_6) are less costly than the non-ambiguous utterances.

References

[1] Bennett, E. & Goodman N. (2018). Extremely costly intensifiers are stronger than quite costly ones. *Cognition*.

[2] Goodman, N. & Frank M. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences 20(11)*. 818-829.

[3] Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceeds of the North American Chapter of the Association for Computational Linguistics*. 159-166.

[4] Levy, R. (2008). Expectation-based syntactic comprehension.*Cognition 106(3)*. 1126-1177.

[5] Zipf, G. (1949). *Human behavior and the principle of least effort*. New York, NY: Prentice-Hall.