Supplementary materials

Pragmatic interactions lead to efficient language structure and use

1 Rational Speech Act theory speaker and listener agents

RSA is a recursive Bayesian model of pragmatic language use, which can largely 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 use including hyperbole, metaphor, implicature and others (CITATIONS).
In the RSA framework, a ”speaker agent” defines a conditional distribution, mapping meanings $u \in U$ to utterances $m \in M$, 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, each of these functions is described in terms of the other. 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

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

2 Zipfian objective for linguistic system efficiency

2.1 Basic objective derivation

Zipf proposed that the particular distributional properties found in natural language emerge as a result of competing speaker and listener pressures. We operationalize this 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.

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

Let speaker effort be the negative log probability (surprisal) of a particular utterance. Intuitively, the number of bits needed to encode the utterance u. This particular formalization of speaker-cost is general enough to accommodate a range of instantiations that might in theory be related to production difficult via articulation effort, cognitive effort related to lexical access, etc [@BennettGoodman2015a].

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

Let listener effort be the negative log probability a listener disambiguates an intended meaning m given an utterance u. This operationalization of listener effort is intuitively related to existing work in sentence processing in which word difficulty during comprehension is proportional to surprisal [@Hale2001a; @Levy2008a].

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

Rewriting (1) we have

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

In general, these expectations are each taken over the set of possible communicative events $e \in E$ weighted by their probability, $P(E = e)$. Recall this is the set of all utterance, meaning pairs $< u, m >= e \in E$.

$$= \sum_{e\in E} p(e)[-log_2(p(u))] + \sum_{e\in E} p(e)[-log_2(L(m|u;\ell))] \tag{3}$$

We assume that the particular joint distribution over utterance-meaning $< u, m >= e$ pairs follows from a simple generative model. First, some meaning is sampled with probability $p(m)$. Our speaker attempts to convey this intended meaning to a listener via an utterance u by sampling from the speaker conditional distribution $S(u|m;\ell)$. Combining these terms leads to the *speaker's joint distribution over events* which we can write as $P(e) = S(u|m;\ell)p(m) = P_{speaker}(u, m;\ell)$.

$$\begin{aligned} &= \sum_{u,m} P_{speaker}(u, m;\ell)[-log_2(p(u))]+ \\ &\sum_{u,m} P_{speaker}(u, m;\ell)[-log_2(L(m|u;\ell))] \end{aligned} \tag{4}$$

Simplifying we arrive at (5):

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

Note that $L(m|u;\ell)p(u)$ is the listener-based joint distribution over all communicative events ($P_{listener}(u,m;\ell)$).

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

This is the simply the cross-entropy between the speaker an listener joint distributions.

$$\begin{aligned} &= \mathbb{E}_{P_{speaker}}[-log_2(P_{listener})] \\ &= H_{cross}(P_{speaker}, P_{listener}; \ell) \end{aligned} \tag{7}$$

From an information-theoretic perspective this objective is intuitive. Cross-entropy gives us 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, this is the difference between the joint distribution assumed by the speaker $P_{speaker}$ and listener $P_{listener}$. A good language ℓ used by a set of a pair of speaker-listeners will have properties which minimize this objective.

2.2 Baseline model objectives

For comparison, we also examine properties of optimal languages under two additional objectives. Zipf [-@Zipf1949a] proposed that the optimal speaker language $\ell^*_{speaker}$ should only optimize speaker effort. We operationalize this using the *first half* of equation (1).

$$\ell^*_{speaker} = argmin_{\ell \in L} \mathbb{E}_{P_{speaker}(u,m;\ell)}(-log_2(p(u))) \tag{1}$$

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

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

2.3 Update to cross-entropy objective to incorporate context

The argument outlined by Piantados et al. [-@Piantadosi2011a] is compatible with our current reference game setting with an addition – context. The authors argue that mapping two meanings m_1 and m_2 to a single utterance u_1 is useful when they can be disambiguated *in context*. In our case we consider a context c to specify a unique ordering of the need probabilities. That is, $p(m|c_1) \neq p(m|c_2)$, when there are two contexts $|C| = \{c_1, c_2\}$.
This in turn, leads to an update to our linguistic efficiency objective as we now would like to consider the average speaker-listener effort over all contextualized communicative events.

$$\begin{aligned} &\text{Efficiency}(S, L, \ell, C) \\ &= \sum_{c \in C} p(c) \sum_{u,m} S(u|m,c;\ell)p(m|c) - log_2[L(m|u,c;\ell)p(u)] \\ &= \mathbb{E}_{c \sim P(c)}[\mathbb{E}_{P_{speaker}(u,m,c;\ell)}[-log_2(P_{listener}(u,m,c;\ell))]] \end{aligned} \tag{1}$$

Note that in the case that $|C| = 1$, our objective simplifies to our original equation (7). With this update we can compare languages L while varying the degree to which context contains useful information, varying the number of contexts (the size of $|C|$).

3 Simulation 2

3.1 Updates to RSA speaker-listeners

We consider the same model of basic speaker and listener ($S_{vanilla}$, $L_{vanilla}$) models in Section 1 in conjunction with 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

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

and

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

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