Communicative Efficiency, Uniform Information Density, and the Rational Speech Act theory

Roger P. Levy
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
rplevy@mit.edu

Abstract

One major class of approaches to explaining the distribution of linguistic forms is rooted in communicative efficiency. For theories in which an utterance’s communicative efficiency is itself dependent on the distribution of linguistic forms in the language, however, it is less clear how to make distributional predictions that escape circularity. I propose an approach for these cases that involves iterating between speaker and listener in the Rational Speech Act theory. Characteristics of the fixed points of this iterative process constitute the distributional predictions of the theory. Through computer simulation I apply this approach to the well-studied case of predictability-sensitive optional function word omission for the theory of Uniform Information Density, and show that the approach strongly predicts the empirically observed negative correlation between phrase onset probability and rate of function word use.

Keywords: Communicative efficiency, Uniform Information Density, Rational Speech Act theory, syntactic optionality, pragmatics, computational modeling

Introduction

One central problem in the science of language is to account for the distribution of linguistic forms observed throughout the languages of the world. A handful of principles are leading candidates for contributing to a theory of this distribution, including innate constraint (Chomsky, 1968), social dynamics among speakers in a linguistic community (Labov, 2011), the acquisition bottleneck in intergenerational transmission (Christiansen & Kirby, 2003), and considerations of communicative efficiency (Zipf, 1949). The present paper focuses on this final principle. The past few decades have seen a resurgence of work on communicative efficiency-based approaches to linguistic distributions, taking advantage of new datasets, contemporary psycholinguistic theory and methods, and structured probabilistic models for language (Genzel & Charniak, 2002; Aylett & Turk, 2004; Pantadosi, Tily, & Gibson, 2011; Futrell, Mahowald, & Gibson, 2015). Here I focus attention on two prominent theories that have each advanced a general principle of what might constitute communicative efficiency: Uniform Information Density (UID; Levy & Jaeger, 2007, Jaeger, 2010) and the Rational Speech Act theory (RSA; Frank & Goodman, 2012, Goodman & Frank, 2016). These theories share the key feature of predicting speaker behavior on the basis of a utility calculation that involves a model of the listener as a probabilistic reasoning agent. This is in contrast to communicative-efficiency theories in which the utility calculation does not involve reasoning about a probabilistic agent (e.g., Hawkins, 1994). However, neither theory presents as straightforward a route to making distributional predictions as theories that do not involve reasoning about probabilistic agents, because the predictions of both UID and RSA are themselves contingent on distributional statistics already present in a language. This paper outlines an approach to circumventing this problem: UID provides a cost function, we allow RSA to perturb a language’s distributional statistics, and the properties of the fixed points of that perturbation process constitute the theory’s distributional predictions. I present a case study deriving the most-tested prediction of UID, that the rate of usage of an optional function word will be negatively correlated with the context-conditioned probability of the material immediately following the function word, more strongly than in previous work. This resolves a potential “levels” problem of UID, described below.

Uniform Information Density

Uniform Information Density proposes that communicative efficiency is maximized when the rate of information—as measured in bits—regarding a speaker’s intended meaning is distributed as uniformly as possible throughout an utterance. Multiple formalizations are possible. One might be called the “noisy-channel” approach. We assume that the listener might not necessarily successfully apprehend all parts of an utterance’s intended meaning, and that more elaborated material provides a stronger signal as to intended meaning. If meaning $m$ is coded within an utterance by material $w$, then the lower $P(m|\text{context})$ is, the greater $P(m|w)$ needs to be in order to ensure that $m$ is successfully apprehended, hence a rational speaker will modulate $w$ to ensure that $P(m|w)$ is high when $P(m|\text{context})$ is low. Instantiations of this idea can be seen in the Smooth Signal Redundancy Hypothesis (Aylett & Turk, 2004) and in the strategic-noise model of Bergen and Goodman (2015). This approach fits especially clearly to cases where “material” is taken to be the surface signal (e.g., acoustics of spoken language) and “meaning” is taken to be a non-surface property such as word identity. A range of results in probability-sensitive articulatory reduction are consistent with the proposal (e.g., Bell, Brenier, Gregory,
Girand, & Jurafsky, 2009), although there remains disagreement as to whether these results should be taken as reflecting audience design or speaker-internal pressures (e.g., Gahl, Yao, & Johnson, 2012). However, a “levels problem” lurks in this noisy-channel UID reasoning: the rational trade-off between meaning prior probability \( P(m|\text{context}) \) relied on conditional independence between the context and \( w \) given \( m \), but this trade-off would itself break this conditional independence. A rational language user who learns the statistics of her environment would learn the resulting dependency and a key premise of the UID derivation would break down.

A second and perhaps stronger formalization might be called the “clean-channel” approach: even setting aside the possibility of utterance misinterpretation, different utterances encoding the same meaning might vary in their communicative efficiency because they impose varying real-time demands on the comprehender. Incremental processing difficulty is probability-sensitive: when a word is less probable in its context, it induces processing difficulty that is probability in this context) RC onset, the fact that the utterance’s communicative efficiency. For example, if a word would be particularly improbable, then mechanisms available to the speaker to offload the fact of the RC’s onset predictability ignores the contextual probability of that itself, and it ignores the impact of using versus omitting that in the probabilities of subsequent words in the string. A more complete predictive theory would need to resolve this problem.

**The Rational Speech Act theory**

The Rational Speech Act theory, and the closely related class of game-theoretic approaches to pragmatics (Franke et al., 2009; Jäger, 2012; Degen, Franke, & Jäger, 2013), embed a notion of communicative efficiency in the choices of a speaker reasoning about a listener (and potentially vice versa). For each possible utterance \( u \) in a set of ALTERNATIVES considered by the speaker, its utility \( V(u;m) \) for communicating meaning \( m \) is defined as

\[
V(u;m) = \log P_L(m|u) - C(u)
\]

where \( P_L(\cdot|u) \) is the listener’s interpretation distribution for \( u \) and \( C(u) \) is the intrinsic cost of using \( u \). Given this utility, the speaker’s utterance preference distribution is defined as

\[
P_S(u|m) \propto \exp[\alpha V(u;m)]
\]

where \( \alpha \) is a softmax decision parameter (often assumed to be 1). RSA models recursively specify the listener probability \( P_L(m|u) \) as in Equation (III) below, and ground out in a “literal” listener \( L_0: \)

\[
P_L(m|u) \propto P_S(u|m)P(m)
\]

where \( \mathcal{L}(m,u) \) is typically an indicator function returning 1 if \( u \) is literally compatible with \( m \), and \( P(m) \) is the contextually-determined prior probability of the speaker intending meaning \( m \).

This basic form of RSA is already quite powerful for capturing a wide range of pragmatic interpretation phenomena, including quantity and informativeness implicatures (Levinson, 2000; Poppels & Levy, 2015), and enriched versions of the theory allow it to handle many more phenomena (see Goodman and Frank (2016) for a review). However, two challenges exist in applying it to making language-wide distributional predictions. First, to make predictions about an entire language, the complete network of alternative utterances for each utterance

\[
\begin{align*}
(1) \quad & \text{I bought some games hackers enjoy yesterday.} \\
(2) \quad & \text{I bought some games that hackers enjoy yesterday.}
\end{align*}
\]
must be specified. One might take the approach that
every utterance is an alternative to every other, but this
approach could find challenge in the body of evidence
for constraints on alternative sets, such as those provided
by lexical scales (Horn, 1989). Second, utterance cost
might plausibly be at least in part a function of its pro-
bability under some language model such as a probabi-
listic context-free grammar, so as to introduce speaker dis-
preferences for low-probability structures. However, ut-
erance probabilities will in general be perturbed by the
speaker function of Equation (II), so that it is unclear
whether the string distribution implied by her utterance preferences by
a pragmatic speaker reasoning about a literal listener should be taken
as the distinctive prediction of an RSA model.

Distributional predictions from RSA & UID

Here I consider a simple proposal for making distri-
butional predictions from RSA that can be incorpo-
rate utterance probabilities into the cost term. Asso-
icate with the base case—the “literal” \(L_0\) level—an initial
probability distribution over strings, \(P_0(u)\). For the
speaker reasoning about \(L_0\), who is conventionally no-
tated as \(S_1\), we can determine the probability distribu-
tion over strings implied by her utterance preferences by
marginalizing out intended utterance meaning: \(P_1(u) = \sum_m P_{S_1}(u|m)P(m)\). This distribution \(P_1\) will in general
be different from \(P_0\). In considering the second prag-
matic speaker, \(S_2\), who is reasoning about \(L_1\), we can take
\(P_1\) as the reference distribution for determining utterance
utility. We can thus obtain a distributions \(P_2, P_3, \ldots\) and so forth by further speaker–listener–speaker
RSA iterations. Typically this sequence will converge at
a FIXED POINT. The fixed point will in general depend
on \(P_0\), but consistent properties of the fixed points can be
taken to be distributional predictions of the model.\(^1\)

In the rest of this paper, we see how this approach can
be applied by using it to surmount the levels problem
of clean-channel UID. One simply incorporates a UID-
based component into the utterance cost function, and
examines the properties of the fixed points to which the
iterated RSA model converges. These properties consti-
tute UID’s distributional predictions.

**UID fixed points: a case study**

We examine the fixed points for clean-channel UID
found by RSA iteration for the case of an optional func-
tion word introducing a phrase, such as in (1)–(2) above.
The key questions are whether this type of function word
can in fact be optional at such fixed points (a priori, all
fixed points might well involve the function word disap-
ppearing or becoming obligatory), and if they can, what
correlation holds between phrasal onset probability and
optional function-word use at the fixed points.

**Language and initial string probabilities**

We consider a language (i.e., a set of strings) of the form:

\[
X((t)Y)Z
\]

where the variable \(X\) represents sentence-initial context,
\(Z\) represents sentence-final context, and parentheses ()
indicate optional material, so that \(Y\) represents an op-
tional modifying phrase that, when present, can be intro-
duced with a symbol \(t\) (see (1)–(2) for mappings of
these variables to a natural language example). Each of
\(X, Y, Z\) are assumed to be two symbols long, and to elim-
inate any possibility of incremental ambiguity (whether
material midway through a sentence is part of \(Y\) or \(Z\))
we draw their content from different symbol sets:

\[
X \in \{AA, Aa, aA, aa\} \\
Y \in \{BB, Bb, Bb, bb\} \\
Z \in \{CC, Cc, cC, cc\}
\]

We take the language to have no ambiguity, and all dif-
fences in string content to correspond to differences in
string meaning, with the exception that string meaning is invariant to the presence or absence of \(t\). The
language has 144 possible strings—\(M=16\) without the op-
tional phrase \(Y, N=64\) with \(Y\) unmarked by \(t\)—and
\(N=64\) with \(Y\) marked by \(t\)—expressing \(M+N=80\)
possible meanings.

We assign an initial probability distribution as follows.
Let the probability that the modifying phrase \(Y\) is present
be \(p_Y\), and the probability, given that a postmodifying
phrase is present, of it being introduced by the marker \(t\)
be \(p_t\). We then stochastically generate a multinomial dis-
tribution over possible strings in the language as follows:

\[
\begin{align*}
\text{Stringset} & \quad \text{Multinomial drawn from} \\
\{XZ\} & \sim (1 - p_Y) \text{Dir}(1, M) & (IV) \\
\{(XYZ, XtYZ)\} & \sim p_Y \text{Dir}(1, N) \otimes \{(1 - p_t), p_t\} & (V)
\end{align*}
\]

where Dir(1, \(K\)) is the uniform Dirichlet distribution over
a \(K\)-dimensional probability vector, and \(\otimes\) the out-
put product. This assignment distributes \((1 - p_Y)\) mass
among the \(M\) strings without the optional phrase \(Y\), and
\(p_Y\) mass among the \(N\) string pairs containing \(Y\). For each
string pair with the optional phrase \(Y, p_t\) of its proba-
bility mass is allocated to the variant with the optional
introductory marker \(t\), and the remainder is allocated to
the variant without the optional introductory marker. We
take \(p_Y\) and \(p_t\) each to be uniformly distributed on [0,1].

Since we have no ambiguity, the prior on meanings \(m\) is:

\[
P(m) = \begin{cases} 
P(XZ) & m \text{ expressible by } XZ \\
P(XYZ) + P(XtYZ) & m \text{ expressible by } X(t)YZ
\end{cases}
\]
Our initial assignment of string probabilities has the following key features:

- the probability that $Y$ will appear after $X$ is context-dependent, varying with the contents of $X$;
- when $Y$ appears, there is probabilistic dependence between its internal contents and the contents of its preceding context $X$ and following content $Y$;
- crucially, however, given that $Y$ appears, the presence of the optional marker $t$ is conditionally independent of all other properties of the string.

This small-scale setup could be generalized in the future to include larger and more variable languages, varying in size, string length, and optional phrase position.

**RSA speaker with UID cost function**

Given a probability distribution over the stringset, we model the behavior of an RSA speaker in the choice problem for every $(XYZ, XTYZ)$ string pair: given that the speaker wants to convey a meaning expressed by the strings in the pair, what will be her preferences over each of the two possible strings? In RSA, speaker $S$ determines her utterance preferences by balancing the effectiveness of each utterance $u$ in guiding listener $L$ to the intended meaning $m$ with utterance cost (Equation II). Since we assumed no ambiguity, $P_S(m|u)$ is simply 1 if $u$ is compatible with $m$ and 0 otherwise. This allows us to decompose the collection of speaker choices to a set of $N$ distributions among $(XYZ, XTYZ)$ pairs:

$$P_S(u|m) \propto \exp[\alpha C(u)]$$

Following UID principles, we include two components in the cost function: peaks and troughs in information density are costly, and long strings are costly. If an utterance $u$ is comprised of the symbol sequence $w_{1…n}$, then for parameters $k \geq 1$ and $c \geq 0$ we define:

$$C(w_{1…n}) = \sum_{i=1}^{n} \left[ -\log P(w_i|w_{1…i-1}) \right]^k + cn \quad (VI)$$

where larger values of $k$ more heavily penalize non-uniformity of information density throughout the string, and larger values of $c$ more heavily penalize longer strings. An objective function of this form was originally considered by Levy (2005, Equation 2.24) and the first term in the summand was analyzed in Levy and Jaeger (2007); as described on page 2, for any fixed joint probability $P(w_{1…n})$ to be distributed over a string, when $k > 1$ this objective is optimized when all conditional word probabilities are equal. We take $P(w_i|w_{1…i-1})$ to be the true (unsmoothed) distribution implied by the full joint distribution over strings.

**Iterating the RSA speaker**

Following frequent RSA practice, we can generalize our model beyond one level of pragmatic reasoning, indexing our original listener and speaker as $L_0$ and $S_1$ respectively, and recursively considering a “pragmatic listener” $L_i$ who reasons about a pragmatic speaker $S_i$ who reasons about a less-sophisticated listener $L_{i-1}$ and so forth, according to Equations II–III. Although $P_{S_i}(m|u)$ will always be deterministic, the distribution over strings $P(w_{1…n})$ implied by the pragmatic speaker’s preferences will in general be different than the distribution used in her cost function (Equation VI). Thus we can index the joint and conditional distributions over strings, with $P_0(w_{1…n})$ and $P_0(w_i|w_{1…i-1})$ the distributions under initialization, and $P_j(w_{1…n})$ and $P_j(w_i|w_{1…i-1})$ the distributions associated with the $j$-th level pragmatic speaker. We can then investigate the behavior of this distribution under repeated iteration (which never changes the prior over meanings, $P(m)$) to a fixed point.

**Fixed-point analysis**

A fixed point will have two key properties of interest to us. First is the marginal frequency of the marker $t$, which occurred at some non-zero rate in the initial distribution. In the fixed point, $t$ may have disappeared, it may have become obligatory, or it may have remained optional. We can refer to this third possibility as STABLE OPTIONALITY. In cases of stable optionality, we are interested in the relationship between the PHRASE ONSET PROBABILITY of $Y$ for any particular context $X$ and the conditional probability of the marker $t$ given that context’s phrase onset probability. The UID intuition is that the greater the phrase onset probability, the less useful $t$ is, and so phrase onset probability should be negatively correlated with $t$-rate. The two are by design uncorrelated in the initial distribution—are they correlated as predicted by UID in the fixed point?

Although the RSA equations are deterministic, the fixed point we reach will depend on the stochastic assignment of initial probabilities to the string set (Equations (IV)–(V)) and to the UID and RSA parameters $k$, $c$, and $\alpha$. (In future work, stochasticity might be introduced directly into the output of each RSA iteration, to assess fixed-point stability in the face of extrinsic noise.) Through repeated simulation, we can investigate the distribution of $t$ marginal frequencies, and the correlation between phrase onset probability and $t$-rate, as a function of these parameters. Here, we investigate through Monte Carlo simulation behavior for $k \in [1, 2]$ and $c \in [0, 2]$. We leave $\alpha = 1$, though note that high values of $\alpha$ will promote more categorical $t$-use.
Results

I take a fixed point to have been reached when both marginal frequency and the Pearson correlation between phrase onset probability and t-rate have intergenerational change less than a tolerance level of 0.001. Results reported here reflect 3.8 million Monte Carlo replicates.

In the special case where \( k = 1, c = 0 \), the RSA pragmatic speaker perfectly replicates the string probabilities, so the initial probabilities are never perturbed and we always have stable optionality. For other cases, Figure 1 shows the rate of stable optionality—operationalized here as a marginal frequency in the range \((0.001, 0.999)\)—as a function of \( k \) and \( c \). When \( k = 1 \), UID plays no role in the cost function and \( t \) always disappears. But stable optionality emerges for a range of \( k, c \) combinations, most frequently around the region of \( c \approx 2(k - 1) \). When stable optionality does not obtain, \( t \) is more likely to disappear when \( c \) is higher, and more likely to become obligatory when \( c \) is lower (not depicted).

Figure 2 shows the distribution of the marginal probability of \( t \) in cases when stable optionality is reached. This distribution is quite different from the uniform initializing distribution over \( t \) rate. In cases of stable optionality, marginal probabilities for \( t \) near 1 are less likely than marginal probabilities near 0, a new generalization that could be tested against known cases of optional function word omission.

Finally, and most crucially, Figure 3 shows the distribution of correlation coefficients between the probabilities of (i) onset of the optional phrase \( Y \) given a particular context \( X \), and (ii) the probability that \( t \) will be used given \( X \) and that the optional phrase \( Y \) appears. In the overwhelming majority of fixed points with stable optionality, onset probability and \( t \)-rate are negatively correlated—despite the fact that phrase onset probability and \( t \)-rate were by design conditionally independent upon initialization.

General Discussion

This paper has identified an intrinsic challenge in making typological predictions regarding language from theories of communicative efficiency that are probabilistic and whose efficiency metrics thus themselves depend on distributional properties. Although one might be concerned that this enterprise runs the risk of circularity, we have seen that circularity can be escaped if a Rational Speech Act speaker model is adopted and iterations of pragmatic inference determine a sequence of probability distributions over the strings in the language. In this setting, characteristics of fixed points to which these sequences converge constitute the distributional predictions of a theory of communicative efficiency. In this setting we can incorporate a Uniform Information Density-based utterance-cost structure. The resulting fixed points for the case of optional function word omission broadly share characteristics that have previously been argued to be predicted by UID and that are observed empirically.

The approach presented here is broadly applicable
and in future work could be applied to a wider variety of speaker choice problems and assumptions regarding utterance cost structures, including clean-channel UID, noisy-channel UID, and beyond. Finally: although the speaker—listener—speaker iterative process leading to the fixed points was presented here in terms of levels of pragmatic reasoning, it could equally be interpreted as representing intergenerational change in language transmission arising from differential communicative success or acquisition of different utterance variants, avoiding the need to appeal to many levels of theory of mind on the part of speakers and listeners.

Acknowledgments

I thank Naomi Feldman, Michael C. Frank, and four anonymous reviewers for helpful comments. NSF BCS awards 1456081 and 1829350 supported this work.

References


