Regularized Conventions:
Equilibrium Computation as a Model of Pragmatic Reasoning

Athul Paul Jacob
apjacob@mit.edu

Gabriele Farina
gfarina@mit.edu

Jacob Andreas
jda@mit.edu

Abstract

We present a model of pragmatic language understanding, where utterances are produced and understood by searching for *regularized equilibria* of signaling games. In this model (which we call ReCo, for Regularized Conventions), speakers and listeners search for contextually appropriate utterance–meaning mappings that are both close to game-theoretically optimal conventions and close to a shared, "default" semantics. By characterizing pragmatic communication as equilibrium search, we obtain principled sampling algorithms and formal guarantees about the trade-off between communicative success and naturalness. Across several datasets capturing real and idealized human judgments about pragmatic implicatures, ReCo matches or improves upon predictions made by best response and rational speech act models of language understanding.

1 Introduction

Meaning in language is fluid: speakers can use the word *blue* to pick out a color that in other contexts would be described as *purple*, or identify a friend as *the one with glasses* even in a room in which everyone is wearing glasses (Figure 1). Such context-dependent meanings can arise as **conventions** within groups of language-users communicating repeatedly to solve a shared task (Hawkins et al., 2017). But remarkably, they can also arise *without any interaction at all*, between pairs of language users who share only common knowledge of words' default meanings.

What makes this kind of context-dependent language use possible? Almost all existing computational models of pragmatics are implemented as **iterated response** procedures, in which listeners interpret utterances by reasoning about the possible intentions of less-sophisticated speakers (or vice-versa) (Golland et al., 2010; Degen, 2023). These models have been successful at explaining a number of aspects of pragmatic language use. But they can be challenging to fit to real data: because they specify behavior in terms of an algorithm that speakers and listeners implement, rather than an objective that they optimize, iterated response models can be highly sensitive to low-level details of initialization and runtime.

We present an alternative model of pragmatic understanding based on **equilibrium search** rather than iterated response. In this model (which we call Regularized Conventions, or ReCo), speakers and listeners solve communicative tasks like the ones in Figure 1 by searching for utterance–

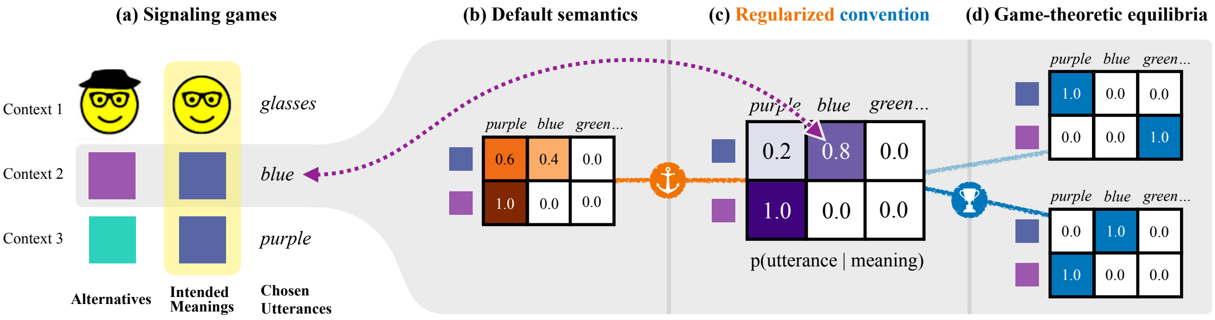

Figure 1: The ReCo model. To communicate (or resolve) an intended meaning from a set of possibilities **(a)**, language users search for a joint distributions over utterances and interpretations that is close to a distribution encoding "default semantics" **(b)** and close to some (game-theoretically) optimal signaling convention **(d)**. The resulting "regularized conventions" **(c)** predict human judgments on a variety of implicature tasks.

meaning mappings that are both close some to a (game-theoretically) optimal convention and close to a set of default semantics. In Figure 1, for example, this “regularized convention” assigns high probability to the use of *blue* to signal the intended color, and low (but nonzero) probability to the use of *purple* instead. This strategy is both close to one of many optimal strategies (in which every utterance arbitrarily, but uniquely, picks out one color), and close to color terms’ standard interpretation (in which the target color is improbably, but not impossibly, described as *blue*).

ReCo is by no means the first application of game-theoretic tools to model pragmatic language understanding (Parikh, 2000; Franke, 2013; Jäger, 2012)—many iterated response models (e.g. Franke, 2009a) also have a game-theoretic foundation. But by leveraging recently developed algorithmic tools for computing regularized equilibria of games, ReCo makes it possible to efficiently learn models of pragmatic communication from data, and to provide formal guarantees about their communicative success and deviation from default semantic. The algorithms that compute these equilibria turn out to have a very similar structure to probabilistic iterated response methods (Frank and Goodman, 2012), offering a possible bridge between algorithmic characterizations of pragmatic reasoning and ReCo’s optimality-based characterization.

Most importantly, ReCo gives a good fit to human data: on classic exemplars of pragmatic implicature, reference tasks eliciting graded human judgments, and tasks featuring perceptually complex meaning spaces, its predictions match or modestly outperform standard iterated response models. These results highlight the usefulness of modern game-theoretic tools in modeling language production and comprehension.

2 Background and Preliminaries

Consider again the example in Figure 1. We want to understand the process by which a SPEAKER might use *blue* to refer to the second color in the second row, and by which a LISTENER might resolve it correctly.

2.1 Signaling Games

This problem has often been formulated as a signalling game (Lewis, 1971), which features two players: the SPEAKER and the LISTENER. In this game, a **target meaning** (representing a communicative need) is first sampled from a space of possible meanings $m \in M$ with probability $p(m)$. To communicate this meaning, the SPEAKER produces an **utterance** according to a policy $\pi_{\mathsf{S}}(u \mid m)$. Finally, the LISTENER produces an **interpretation** according to a policy $\pi_{\mathsf{L}}(m' \mid u)$.

Informally, communication is successful if the LISTENER’s interpretation is the same as the SPEAKER’s intended meaning. More formally (and somewhat more generally), we may define communicative success in terms of **rewards**. Consider any (meaning, utterance, interpretation) combination (m, u, m'). The SPEAKER’s reward (or “payoff”) $r_{\mathsf{S}}(m, u, m')$ in this interaction is the sum of:

- a *cost* $-c(u)$ that the SPEAKER incurs for producing utterance u (all else equal, they may for example prefer short utterances); and
- a *success measure*, equal to 1 only when m' matches the target m, that is, $\mathbf{1}[m' = m]$.

Together,

$$r_{\mathsf{S}}(m, u, m') := -c(u) + \mathbf{1}[m' = m].$$

Most models assume that the LISTENER’s reward $r_{\mathsf{L}}(m, u, m')$ depends only on communicative success:

$$r_{\mathsf{L}}(m, u, m') = \mathbf{1}[m' = m].$$

Having specified rewards for all interactions, the *expected utility* of each player given policies $(\pi_{\mathsf{S}}, \pi_{\mathsf{L}})$ for the SPEAKER and LISTENER respectively is defined as the expected payoff when the meanings m are sampled from a prior distribution $p(m)$, and agents sample from their policies:

$$\bar{u}_i(\pi_{\mathsf{S}}, \pi_{\mathsf{L}}) := \mathop{\mathbb{E}}_{\substack{m \sim p \\ u \sim \pi_{\mathsf{S}}(\cdot|m) \\ m' \sim \pi_{\mathsf{L}}(\cdot|u)}} p_i(m, u, m') \qquad (1)$$

for $i \in \{\mathsf{S}, \mathsf{L}\}$.

2.2 Computing Policies for Signaling Games

How should a SPEAKER and LISTENER communicate to maximize the probability of success? We call a pair of policies, for the SPEAKER and for the LISTENER a *Nash equilibrium* if nobody wants to deviate. Notice that there may in general be multiple such policies: in Figure 1, for example, there is one equilibrium policy in which the intended meaning is called *blue* and the alternative is called *purple*, but another equilibrium policy in which the former is called *purple* and the latter called *green* (in clear violation of those words' standard use in English!). When a SPEAKER and LISTENER communicate for the first time, how can they ensure that their policies are compatible?

Iterated response methods A popular family of approaches attempt to ensure communicative success *algorithmically*. These approaches typically begin from an assumption that SPEAKERS' and LISTENERS' common knowledge of language consists of a **literal semantics** (which assigns context-independent meanings to utterances). Agents then derive policies by computing behaviors likely to be successful given an interlocutor communicating literally, or given an interlocutor themself attempting to respond to a literal communicator. Approaches in this family involve (Iterated) Best Response (I)BR (Jäger, 2007; Franke, 2009b,a) and Rational Speech Acts (RSA) (Frank and Goodman, 2012).

(I)BR is an iterative algorithm in which speakers (listeners) alternatingly compute the highest-utility action keeping the listener's (speaker's) policy fixed:

$$\pi_{\mathsf{L}}^{(t+1)}(m' \mid u) = \mathbf{1}\left[m' = \arg\max_{m} \pi_{\mathsf{S}}^{(t)}(u \mid m)\right]$$

$$\pi_{\mathsf{S}}^{(t+1)}(u \mid m) = \mathbf{1}\left[u = \arg\max_{u'} \pi_{\mathsf{L}}^{(t)}(m \mid u')\right]$$

RSA frames communication as one in which Bayesian listeners and speakers reason recursively about each other's beliefs in order to choose utterances and meanings:

$$\pi_{\mathsf{L}}^{(t)}(m \mid u) \propto \pi_{\mathsf{S}}^{(t)}(u \mid m) \cdot p(m)$$

$$\pi_{\mathsf{S}}^{(t)}(u \mid m) \propto \left(\pi_{\mathsf{L}}^{(t)}(m \mid u)/c(u)\right)^{\alpha}$$

In both approaches, "good" policies are obtained by assuming that speakers and listeners will run a specific algorithm from a specific starting point (rather than generically optimizing a known objective). As a result, a key feature of both algorithms is its sensitivity to the choice of initial ($t = 0$) policy; their convergence behavior remains poorly understood.

piKL-Hedge and regularized no-regret dynamics A set of principled techniques for solving games comes from the vast literature of online optimization and learning in games. Hedge (Littlestone and Warmuth, 1994; Freund and Schapire, 1997) is a popular iterative algorithm in this family that converges to a coarse correlated equilibrium (Hannan, 1957) and to a Nash equilibrium in the special case of two-player zero-sum games. However, in many cases, the equilibria that is of interest is one that is close to certain **anchor policies** – which Hedge does not guarantee.

In order to sidestep this issue while retaining the appealing properties of learning in games, Jacob et al. (2022) introduced **piKL-Hedge**. piKL-Hedge has been used in the context of board games, like Diplomacy (FAIR et al., 2022; Bakhtin et al., 2022) to find equilibria that are close to human imitation learned anchor policies. Recently, piKL-Hedge has been used in the context of language models, with the objective of increasing consensus between disriminative and generative approaches to language model generation (Jacob et al., 2023).

3 Our Approach: Pragmatic Inference as Regularized Equilibrium Search

Building on this past work, the key idea underlying the RECO model is to define an objective that makes it possible to directly optimize for both communicative success and adherence to shared background knowledge of language. As noted in Section 2.2, simply searching for high-utility equilibria of signaling games is unlikely to predict the behavior of human language users, or result in successful communication with new interlocutors: instead, we must guide inference toward policies that *look like natural language*. In RECO, we do so by optimizing utilities of the following form:

$$\tilde{u}_{\mathsf{S}}(\pi_{\mathsf{S}}, \pi_{\mathsf{L}}) := \bar{u}_{\mathsf{S}}(\pi_{\mathsf{S}}, \pi_{\mathsf{L}}) - \lambda_{\mathsf{S}} \cdot \mathrm{D}_{\mathrm{KL}}(\pi_{\mathsf{S}} \,\|\, \tau_{\mathsf{S}}),$$

$$\tilde{u}_{\mathsf{L}}(\pi_{\mathsf{S}}, \pi_{\mathsf{L}}) := \bar{u}_{\mathsf{L}}(\pi_{\mathsf{S}}, \pi_{\mathsf{L}}) - \lambda_{\mathsf{L}} \cdot \mathrm{D}_{\mathrm{KL}}(\pi_{\mathsf{L}} \,\|\, \tau_{\mathsf{L}}).$$

Here τ_{S} and τ_{L} represent the SPEAKER's and LISTENER's prior knowledge of language (independent of any specific communicative goal or context). We refer to these policies as the **default semantics** in the language used for communication. They play a similar role to the literal semantics used by RSA and other iterated response models. But here, we need not assume that they correspond specifically to literal semantics—instead, they model agents' prior expectations about how utterances are likely to be produced and interpreted in general by pragmatic language users.

The regularization parameters λ_{S} and λ_{L} control the amount of regularization towards the default semantics $\tau_{\mathsf{S}}, \tau_{\mathsf{L}}$. When the value of λ_i is large, Player $i \in \{\mathsf{S}, \mathsf{L}\}$ will be regularized towards only considering policies extremely close to τ_i; conversely, when λ_i is close to zero, the player will not be penalized for adopting semantics that differ significantly from τ_i.

3.1 Notation and representation of policies

In this subsection, we lay down the notation and representation details for the policies produced by our algorithm. Each agent's *policy* consists of a mapping from that player's observations to a distribution over actions. In order to provide a compact description of the algorithm, as well as an efficient vectorized implementation, we represent such a mapping as a row-stochastic matrix, with rows indexed by observations and columns indexed by actions. For the SPEAKER, the set of observations coincides with the set of meanings available in a given communicative context, and the set of actions coincides with the set of possible utterances. For the LISTENER, observations are utterances and actions are meanings. See Figure 2 for examples. We denote with $\mathbf{S}^{(t)} \in \mathbb{R}^{M \times U}$ the policy of the speaker at time t, and with $\mathbf{L}^{(t)} \in \mathbb{R}^{U \times M}$ that of the listener represented in this matrix form. Similar, we will also represent the anchor policies (*i.e.*, default semantics) $\tau_{\mathsf{S}}, \tau_{\mathsf{L}}$ in this representation as matrices $\boldsymbol{\tau}_{\mathsf{S}} \in \mathbb{R}^{M \times U}$ and $\boldsymbol{\tau}_{\mathsf{L}} \in \mathbb{R}^{U \times M}$. Instances of such matrix objects can be seen in Figure 2.

3.2 RECO: Computation of Approximate Convention-Regularized Equilibria

Given the regularized utilities $\tilde{u}_{\mathsf{S}}$ and u_{L} defined above, we use the piKL-Hedge algorithm (Jacob et al., 2022) to progressively refine the SPEAKER's and LISTENER's policy toward equilibrium (in the sense of Section 2.2). Intuitively, piKL-Hedge performs a variant of projected gradient ascent in the geometry of entropic regularization where projections are equivalent to softmax (normalized exponentiation). In order to apply piKL-Hedge, we start by computing the gradients of the unregularized utility functions $\bar{u}_{\mathsf{S}}, \bar{u}_{\mathsf{L}}$ defined in (1).

Let $\boldsymbol{p} \in \mathbb{R}^M$ be the vector whose entries correspond to $p(m)$, the prior distribution over meanings. Similarly, we let $\boldsymbol{c} \in \mathbb{R}^U$ denote the vector of utterance costs. Finally, let $\mathbf{P} \in \mathbb{R}^{M \times M}$ be the diagonal matrix whose diagonal equals $\boldsymbol{p}$. With this notation, it is straightforward to verify that the gradient of the unregularized utility function $\bar{u}_{\mathsf{S}}$ of the SPEAKER player, as a function of the matrix-form policies $\mathbf{S}, \mathbf{L}$, is given by

$$\nabla_{\mathbf{S}}(\mathbf{L}) := -\boldsymbol{p}\boldsymbol{c}^\top + \mathbf{P}\mathbf{L}^\top \in \mathbb{R}^{M \times U}. \qquad (2)$$

Similarly, for the LISTENER player we have

$$\nabla_{\mathbf{L}}(\mathbf{S}) := \mathbf{S}^\top \mathbf{P} \in \mathbb{R}^{U \times M}. \qquad (3)$$

With the above gradients, piKL-Hedge prescribes the following dynamics: first, at time 0, set $\bar{\mathbf{S}}^{(0)} = \bar{\mathbf{L}}^{(0)} := \mathbf{0}$; then, at each time $t \geq 0$, the next policy $\mathbf{S}^{(t+1)}, \mathbf{L}^{(t+1)}$ is chosen according to the update rules:

$$\begin{aligned}
\mathbf{S}^{(t+1)} &\overset{\text{row}}{\propto} \exp\left\{ \frac{\nabla_{\mathbf{S}}(\bar{\mathbf{L}}^{(t)}) + \lambda_{\mathsf{S}} \log \boldsymbol{\tau}_{\mathsf{S}}}{1/(\eta_{\mathsf{S}} t) + \lambda_{\mathsf{S}}} \right\}, \\
\mathbf{L}^{(t+1)} &\overset{\text{row}}{\propto} \exp\left\{ \frac{\nabla_{\mathbf{L}}(\bar{\mathbf{S}}^{(t)})^\top + \lambda_{\mathsf{L}} \log \boldsymbol{\tau}_{\mathsf{L}}}{1/(\eta_{\mathsf{L}} t) + \lambda_{\mathsf{L}}} \right\}, \\
\bar{\mathbf{S}}^{(t+1)} &= \frac{t}{t+1}\bar{\mathbf{S}}^{(t)} + \frac{1}{t+1}\mathbf{S}^{(t+1)}, \\
\bar{\mathbf{L}}^{(t+1)} &= \frac{t}{t+1}\bar{\mathbf{L}}^{(t)} + \frac{1}{t+1}\mathbf{L}^{(t+1)},
\end{aligned}$$

where $\overset{\text{row}}{\propto}$ denotes row-wise proportionality and exponentiation is intended elementwise.

piKL-Hedge dynamics have strong guarantees, including the following:

- the average correlated distribution of play of SPEAKER and LISTENER converges to the set of coarse-correlated equilibria of the game defined by the regularized utilities $\tilde{u}_{\mathsf{S}}, \tilde{u}_{\mathsf{L}}$;
- for any $i \in \{\mathsf{S}, \mathsf{L}\}$, the average policy of Player i lies within a distance of roughly $1/\lambda_i$ from the default semantics $\boldsymbol{\tau}_i$;
- the policies produced by piKL-Hedge guarantee that the player's regret will remain bounded by a functions whose growth is logarithmic in the number of training steps.

3.3 Special Case: Uniform Priors, No Costs

When the prior over the objects is uniform, and utterance costs are all set to zero, the gradients $\nabla_{\mathbf{S}}(\mathbf{L})$ and $\nabla_{\mathbf{L}}(\mathbf{S})$, defined in (2) and (3), simplify into

$$\nabla_{\mathbf{S}}(\mathbf{L}) = \frac{1}{|M|}\mathbf{L}, \quad \nabla_{\mathbf{L}}(\mathbf{S}) = \frac{1}{|M|}\mathbf{S}.$$

Hence, piKL-Hedge reduces to the simple algorithm that repeatedly updates and renormalizes policy matrices according to

$$\mathbf{S}^{(t+1)} \overset{\text{row}}{\propto} \exp\left\{ \frac{(\bar{\mathbf{L}}^{(t)})^\top + \hat{\lambda}_{\mathsf{S}} \log \boldsymbol{\tau}_{\mathsf{S}}}{1/(\hat{\eta}_{\mathsf{S}} t) + \hat{\lambda}_{\mathsf{S}}} \right\},$$

$$\mathbf{L}^{(t+1)} \overset{\text{row}}{\propto} \exp\left\{ \frac{(\bar{\mathbf{S}}^{(t)})^\top + \hat{\lambda}_{\mathsf{L}} \log \boldsymbol{\tau}_{\mathsf{L}}}{1/(\hat{\eta}_{\mathsf{L}} t) + \hat{\lambda}_{\mathsf{L}}} \right\},$$

where we let $\hat{\lambda}_i := |M|\lambda_i$ and $\hat{\eta}_i := \eta_i/|M|$ for all $i \in \{\mathsf{S}, \mathsf{L}\}$.

The above procedure is similar to the Rational Speech Acts model (Frank and Goodman, 2012), a widely used probabilistic iterated response model of pragmatics. In particular, using the same matrix notation from above, we may express RSA (with $\alpha = 1.0$) as:

$$\begin{aligned}
\bar{\mathbf{L}}^{(0)} &= \boldsymbol{\tau}_{\mathsf{L}} \\
\mathbf{S}^{(t+1)} &\overset{\text{row}}{\propto} (\bar{\mathbf{L}}^{(t)})^\top, \\
\bar{\mathbf{S}}^{(t+1)} &= \mathbf{S}^{(t+1)}, \\
\mathbf{L}^{(t+1)} &\overset{\text{row}}{\propto} (\bar{\mathbf{S}}^{(t)})^\top, \\
\bar{\mathbf{L}}^{(t+1)} &= \mathbf{L}^{(t+1)}.
\end{aligned}$$

Thus, it is also possible to interpret ReCo as an RSA variant in which (1) the final policy at level t is a weighted average of policies computed at lower levels, (2) both speakers and listeners downweight actions that are low-probability under the default semantics.

Having defined the ReCo objective and procedures for optimizing it, the remainder of this paper evaluates whether ReCo can successfully predict human judgments across standard test-beds for pragmatic implicature.

4 Two Model Problems: Q-implicature and M-implicature

We begin with two simple, widely studied "model problems" in pragmatics: Quantity implicature and Manner implicature. The experiments in this section aim to demonstrate that ReCo makes predictions that agree qualitatively with key motivating examples in theories of pragmatics.

4.1 Quantity Implicature

Quantity (or "scalar") implicatures are those in which a weak assertion is interpreted to mean that a stronger assertion does not hold. (For example, *Avery ate some of the cookies* ⇸ *Avery did not eat all of the cookies*, where ⇸ denotes pragmatic implication; (Huang, 1991)). The reference game we use as a model of scalar implicature is adopted from Jäger (2012); its associated default semantics is shown in Figure 2. Here, the utterances *none*, *some*, and *all* are used to communicate meanings `none`, `some (not all)`, and `all`. *Some* can (literally) denote *all* (as we may felicitously say *Avery ate some of the cookies; in fact, Avery ate all of them*), but is generally understood to *implicate* `not all`. The policy found by ReCo is shown in Figure 2, where it can be seen that this prediction is recovered by ReCo.

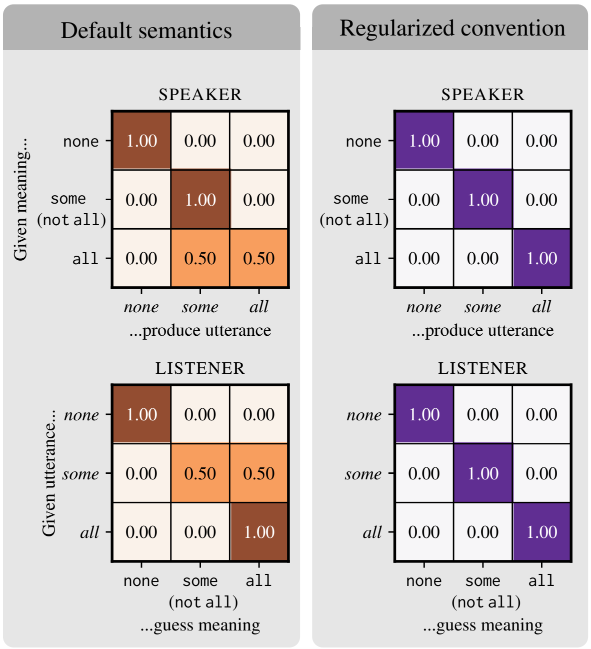

Figure 2: Quantity implicatures in ReCo. (Left) Matrices representing conditional probabilities that represent the default semantics τ_{S} and τ_{L}. (Right) Matrices representing conditional probabilities that represent the resulting regularized conventions π_{S} and π_{L}. In this setting, ReCo is able to predict the correct set of interpretations.

4.2 Manner Implicature

Another important class of implicatures are Manner implicatures, in (a subclass of) which an atypical utterance is used to denote that a situation occurred in an atypical way (*I started the car* ⇸ *The car started normally*; but *I got the car to start* ⇸ *The car started abnormally*; Levinson, 2000). The reference game we adopt as a model of such implicatures is due to Bergen et al. (2016). In this model, we assume we have two utterances (*short* and *long*) and two meanings (`freq` and `rare`) satisfying the following properties: (1) `freq` is more often the intended meaning than `rare`, (2) *long* is more costly to communicate than *short*, but (3) either *long* or

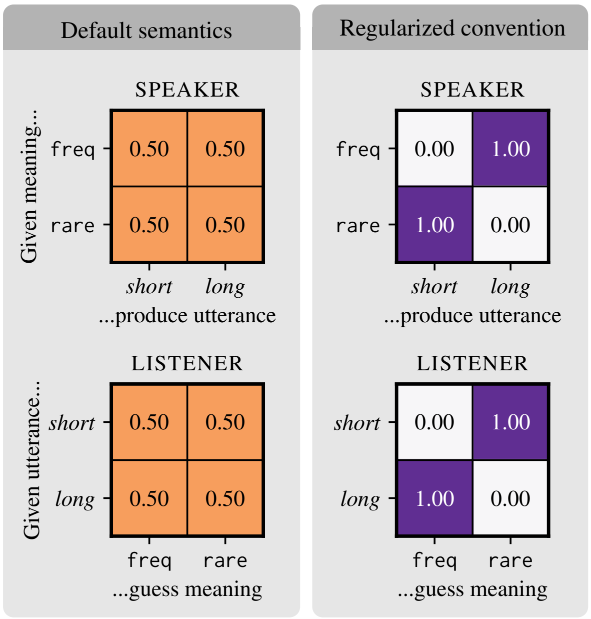

Figure 3: Manner implicatures in ReCo. (Left) Matrices representing conditional probabilities that represent the default semantics τ_S and τ_L. (Right) Matrices representing conditional probabilities that represent the resulting regularized conventions π_S and π_L. By incorporate prior probabilities of meanings and costs for utterances, ReCo is able to predict the correct set of interpretations.

	Literal LISTENER	BR SPEAKER	RSA	RD-RSA	ReCo
ALL	73.57%	90.04%	95.07%	94.98%	**95.96%**
SIMPLE	70.10%	88.16%	**96.02%**	**96.02%**	**96.02%**
COMPLEX	83.86%	97.83%	94.74%	94.35%	**98.18%**
TWINS	97.61%	93.43%	97.61%	**98.98%**	97.61%
ODDMAN	**94.97%**	**94.97%**	**94.97%**	**94.97%**	**94.97%**

Table 1: Correlation across different methods with graded human judgements in four reference games Frank (2016) (with the best hyperparameter settings). ReCo performs better than the alternatives in ALL.

short may, by default, denote `freq` or `rare`. In such situations, *short* is understood to implicate `freq` and *long* to implicate `rare`; as noted by Bergen et al. (2016), RSA and related theories require substantial modification to derive these predictions.

When using ReCo to perform equilibrium search with these costs and priors, it natively predicts the correct set of interpretations (Figure 3).

5 Graded Human Judgments

We next study a family of four reference tasks introduced by Frank (2016), which we refer to as SIMPLE, COMPLEX, TWINS and ODDMAN. We refer readers to the original work for the default meanings that define each of these tasks. Frank gathered graded human judgments about the likelihood that particular utterances might carry particular meanings. ReCo, like RSA-family models, captures probabilistic associations between utterances and meanings, we may evaluate the quality of its predictions in terms of correlations with human judgments.

.	Context			Utterance
1.	🟦	🟩	🟥	*purple*
2.	🟦	🟪	🟩	*blue*
3.	🟦	🟦	🟫	*blue*

Table 2: Example of the Colors in Context task (Monroe et al., 2017). The SPEAKER produces an utterance to refer to the reference color (the one within the black box) subject to the context to a LISTENER. As in Figure 1, notice how context affects the utterance.

	Literal LISTENER	BR SPEAKER	RSA	RD-RSA	ReCo
CIC (val.)	84.88%	75.90%	84.18%	84.18%	**85.17%**
CIC (test)	83.34%	74.28%	83.41%	83.41%	**83.62%**

Table 3: Performance of different models on Colors in Context (Monroe et al., 2017). All approaches aside from BR perform well on this task – as even literal models have access to all three referents. Note that, ReCo performs best.

Comparisons between ReCo, RSA, BR SPEAKER (i.e., best-response to a literal speaker) and RD-RSA (Zaslavsky et al., 2021) are shown in Table 1, with additional information about parameters in Figure 4. In these figures, ALL denotes correlations computed across all four tasks. It can be seen that ReCo modestly improves upon the best predictions of RSA across a range of speaker parameters.

6 Complex Referents and Utterances

Our final experiments focus on Colors in Context (CIC), a dataset of color reference tasks like the one in Figure 1 featuring a more complex space of meanings and a larger space of utterances. Another example from the dataset (introduced by Monroe et al., 2017) is given in Table 2. For this task, we use human-generated utterances collected by the authors across 948 games yielding a total of 46,994 utterances. We divide this data into an 80% / 10% / 20% train / validation / test splits. Here, we evaluate models by measuring the accuracy with which they can infer the intended meaning produced by a human SPEAKER.

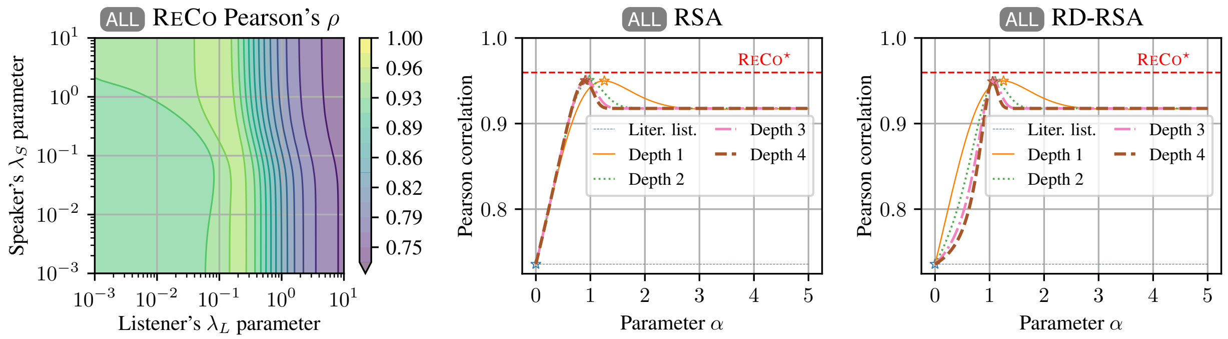

Figure 4: Pearson's correlation ρ on the full dataset of graded human judgments from (Frank, 2016). (Left) Correlation for RECo as a function of λ_L and λ_S represented as a contour plot. (Middle) Correlation between RSA at different levels of α and recursive depth (Right) Correlation between RD-RSA at different levels of α and recursive depth. (Middle, Right) RECo with the best setting of λ_L and λ_S is indicated with a red dashed line. Stars indicate the best α value at different depths.

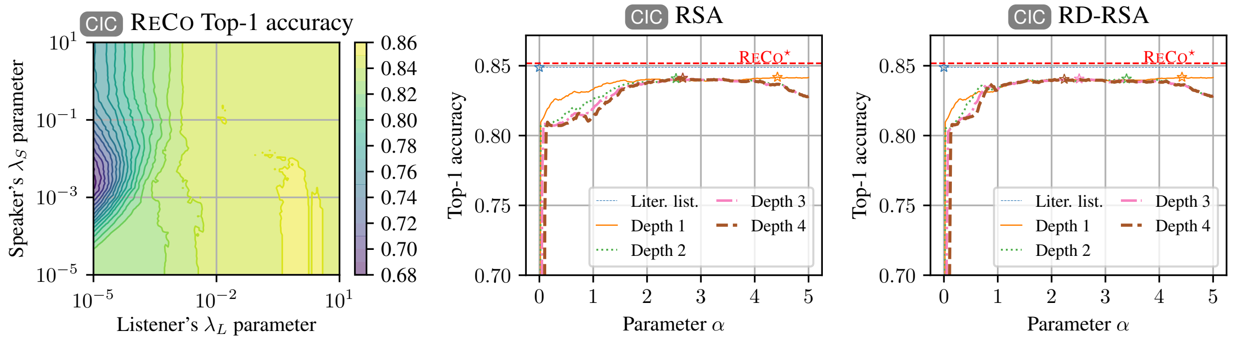

Figure 5: Top-1 accuracy of predicting meanings on the validation set of the Colors in Context task (Monroe et al., 2017). (Left) Accuracy for RECo as a function of λ_L and λ_S represented as a contour plot. (Middle) Accuracy of RSA at different levels of α and recursive depth (Right) Accuracy of RD-RSA at different levels of α and recursive depth. (Middle, Right) RECo with the best setting of λ_L and λ_S is indicated with a red dashed line. Stars indicate the best α value at different depths.

Base models Following past work (Monroe et al., 2017), we first train a transformer-based literal listener as a model that takes in the three colors and a natural language utterance, and uses these to predict the index of the referent. We also train a transformer-based speaker model, which takes in the context and target referent and generates a natural language utterance.

Candidate utterances The set of utterances are produced by first sampling 5 candidate utterances for each of the 3 possible targets from the speaker model along with the produced utterance, for a total of 16 candidates.

Results are shown in Figure 5 and Table 3. As with past work (McDowell and Goodman, 2019; Monroe et al., 2017), all models aside from BR perform well (even the literal listener); RECo matches (or perhaps slightly improves upon) these results.

7 Conclusion

We have presented a model of pragmatic understanding based on equilibrium search called RECo. In this model, speakers and listeners solve communicative tasks by searching for utterance-meaning mappings that that simultaneously optimize reward and similarity to a set of default meanings. RECo offers a link between "algorithmic" models of pragmatic reasoning and equilibrium-based models, and accurately predicts human judgments across several pragmatic reasoning tasks.

Acknowledgements

Thanks to Jennifer Hu for helpful suggestions about experiment design. This work was supported by the National Science Foundation under grant IIS-2212310.

References

Anton Bakhtin, David J Wu, Adam Lerer, Jonathan Gray, Athul Paul Jacob, Gabriele Farina, Alexander H Miller, and Noam Brown. 2022. Mastering the game of no-press diplomacy via human-regularized reinforcement learning and planning. In *The Eleventh International Conference on Learning Representations*.

Leon Bergen, Roger Levy, and Noah Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 9:ACCESS–ACCESS.

Judith Degen. 2023. The rational speech act framework. *Annual Review of Linguistics*, 9:519–540.

Meta Fundamental AI Research Diplomacy Team FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Michael C Frank. 2016. Rational speech act models of pragmatic reasoning in reference games.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Michael Franke. 2009a. Interpretation of optimal signals. *New perspectives on games and interaction*, pages 297–310.

Michael Franke. 2009b. *Signal to act: Game theory in pragmatics*. University of Amsterdam.

Michael Franke. 2013. Game theoretic pragmatics. *Philosophy Compass*, 8(3):269–284.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 410–419.

James Hannan. 1957. Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139.

Robert XD Hawkins, Mike Frank, and Noah D Goodman. 2017. Convention-formation in iterated reference games. In *CogSci*.

Yan Huang. 1991. A neo-gricean pragmatic theory of anaphora1. *Journal of linguistics*, 27(2):301–335.

Athul Paul Jacob, Yikang Shen, Gabriele Farina, and Jacob Andreas. 2023. The consensus game: Language model generation via equilibrium search. *arXiv preprint arXiv:2310.09139*.

Athul Paul Jacob, David J Wu, Gabriele Farina, Adam Lerer, Hengyuan Hu, Anton Bakhtin, Jacob Andreas, and Noam Brown. 2022. Modeling strong and human-like gameplay with kl-regularized search. In *International Conference on Machine Learning*, pages 9695–9728. PMLR.

Gerhard Jäger. 2007. Game dynamics connects semantics and pragmatics. In *Game theory and linguistic meaning*, pages 103–117. Brill.

Gerhard Jäger. 2012. Game theory in semantics and pragmatics. *Semantics: An international handbook of natural language meaning*, 3:2487–2516.

Stephen C Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

David K Lewis. 1971. Convention: A philosophical study. *Philosophy and Rhetoric*, 4(2).

Nick Littlestone and Manfred K Warmuth. 1994. The weighted majority algorithm. *Information and computation*, 108(2):212–261.

Bill McDowell and Noah Goodman. 2019. Learning from omission. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 619–628.

Will Monroe, Robert XD Hawkins, Noah D Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Prashant Parikh. 2000. Communication, meaning, and interpretation. *Linguistics and philosophy*, pages 185–212.

Noga Zaslavsky, Jennifer Hu, and Roger Levy. 2021. A rate–distortion view of human pragmatic reasoning? In *Proceedings of the Society for Computation in Linguistics 2021*, pages 347–348.