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 game-theoretic model of pragmatics that we call RECO (for Regularized Conventions). This model formulates pragmatic communication as a game in which players are rewarded for communicating successfully and penalized for deviating from a shared, "default" semantics. As a result, players assign utterances context-dependent meanings that jointly optimize communicative success and naturalness with respect to speakers' and listeners' background knowledge of language. By using established game-theoretic tools to compute equilibrium strategies for this game, we obtain principled pragmatic language generation procedures with formal guarantees of communicative success. Across several datasets capturing real and idealized human judgments about pragmatic implicature, RECO matches, or slightly improves upon, predictions made by Iterated Best Response and Rational Speech Acts models of language understanding.

1 Introduction

Meaning in language is fluid and context-sensitive: 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* in a room in which everyone is wearing glasses (Figure 1). Such context-dependent meanings can arise as **conventions** among language users communicating repeatedly to solve a shared task (Clark and Wilkes-Gibbs, 1986). But remarkably, they can also arise *without any interaction at all*, among language users who share only common knowledge of words' default meanings (Grice, 1975).

What makes this kind of context-dependent pragmatic language use possible? Almost all existing computational models of pragmatics are implemented as **recursive reasoning** procedures, in which listeners interpret utterances by reasoning about the intentions of less-sophisticated speakers (Golland et al., 2010; Degen, 2023). These models have been successful at explaining a number of aspects of pragmatics. But they can be challenging to fit to real data: because they specify speaker and listener behavior procedurally, rather than in terms of a shared objective, recursive reasoning models can be highly sensitive to implementation-level details (e.g. the number of "levels" of reasoning).

We present an alternative model of pragmatic understanding based on **equilibrium search** rather than recursive reasoning. In this model (which we call Regularized Conventions, or RECO), speakers and listeners solve communicative tasks like

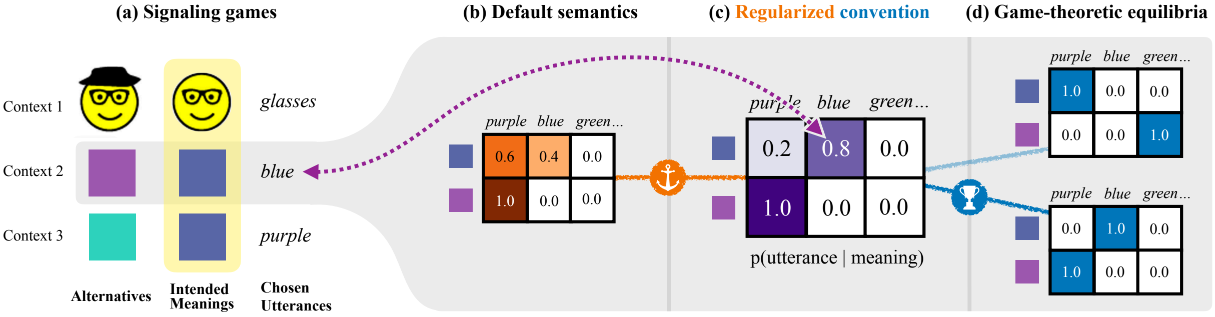

Figure 1: The RECO model. To communicate (or resolve) an intended meaning from a set of possibilities **(a)**, language users search for distributions over utterances and interpretations that are close to some "default semantics" **(b)** and close to a (game-theoretically) optimal signaling convention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

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2944–2955
June 16-21, 2024 ©2024 Association for Computational Linguistics

those in Figure 1 by searching for utterance–meaning mappings that are both close to a game-theoretically optimal communicative convention (a **signaling equilibrium**), and close to a shared initial semantics (which functions as a **regularizer**). In Figure 1, for example, 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 conventions (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)—in fact, many recursive reasoning models (e.g. Franke, 2009a) also have a game-theoretic interpretation. But by leveraging recently developed algorithmic tools for computing regularized equilibria of games, RECO can efficiently learn models of pragmatic communication from data, while providing formal guarantees about communicative success and deviation from default semantics. The algorithms that compute these equilibria turn out to have a very similar structure to some *probabilistic* recursive reasoning methods (e.g. Frank and Goodman, 2012), offering a bridge between procedural 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 (and sometimes modestly outperform) standard recursive reasoning models. These results show that game-theoretic approaches offer a viable foundation for expressive, learned models of pragmatic communication, and highlight the usefulness of the modern game-theoretic toolkit in more general systems for language production and comprehension.

2 Background and Preliminaries

Consider again the example in Figure 1. We wish 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

The problem depicted in Figure 1 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** $u \in U$ 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 $r_{\mathsf{S}}(m, u, m')$ in this interaction is the sum of:

- an *utterance 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, $\mathbf{1}[m' = m]$ (the SPEAKER wishes for the the LISTENER to identify their intended meaning).

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 reward when the meanings m are sampled from a prior distribution $p(m)$, and agents sample from their policies:

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

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 neither agent is incentivized to unilaterally modify their own policy given that the other agent's policy is fixed: formally,

$$\pi_i = \arg\max_{\pi} \bar{u}_i(\pi, \pi_{-i}) \ .$$

where π_{-i} denotes the policy used by the player other than i. In the bottom row of Figure 1(d), neither the SPEAKER nor LISTENER can improve their reward by unilaterally deciding that *blue* refers to a different color.

Notice that there may in general be multiple such policies: returning to Figure 1(d), the bottom row shows an equilibrium policy in which the intended meaning is called *blue* and the alternative is called *purple*, but the top row shows a different equilibrium policy in which the former is called *purple* and the latter called *green* (in clear violation of those words' standard use in English!).

This fact underlines a major limitation of signaling games (in their simplest form) as models of communication—while they can explain which utterance–meaning mappings correspond to stable conventions, they cannot explain why *particular* mappings are chosen in particular communicative contexts against the background of a shared language. In Figure 1(d), what prior knowledge of language allows us to identify the second row as more "natural" than the first one? When a SPEAKER and LISTENER communicate for the first time, how can they leverage this knowledge to ensure that they both identify the *same* mapping from utterances to meanings in context?

Recursive reasoning methods A popular family of approaches answers these questions *procedurally*. 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, 2009a,b) and the Rational Speech Acts model (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 a process 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 the same inference procedure from a specific starting point (rather than generically optimizing a fixed objective). As a result, a key feature of both algorithms is sensitivity to the choice of initial ($t = 0$) policy and number of iterations; their convergence behavior remains poorly understood in all but the simplest settings (though see Zaslavsky et al., 2021b for a discussion of the quantity optimized by single-step updates).

Hedge and game-solving algorithms While not widely used in the computational linguistics or natural language processing literature, techniques for directly optimizing for communicative success, as in Equation (1), may be found in the vast body of work on 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 general it provides no guarantees about *which* equilibrium will be found when multiple such equilibria exist. This presents a challenge not just in signaling, but in any game where strategies computed by equilibrium search will be used to interact with human players adhering to pre-established conventions.

Regularized search In order to sidestep this issue while retaining the appealing properties of learning in games, Jacob et al. (2022) introduced **piKL-Hedge**, a procedure for finding *regularized equilibria* that are close to chosen "anchor policies". piKL-Hedge (discussed in more detail below) has been applied to board games like Diplomacy (FAIR et al., 2022; Bakhtin et al., 2022) to find equilibria that are close to policies learned via imitation from human play. Recently, piKL-Hedge has also been applied to language model decoding, with the objective of increasing consensus between discriminative and generative approaches to language model generation (Jacob et al., 2023b).

Regularization toward an anchor policy has also been used in the context of recursive-reasoning models of pragmatics. Hawkins et al. (2020, 2023) model speakers as incurring a cost for deviating from a predictions made by a language model, corresponding to a form of "regularized RSA" in which the entropy or mutual information penalties present in RSA and RD-RSA respectively are replaced with a base model of language; below, we describe how to apply this same kind of regularization in the context of equilibrium search to derive an alternative model of pragmatics.

3 Our Approach: Pragmatic Inference as Regularized Equilibrium Search

The key idea underlying ReCo is to use regularized equilibrium concepts to describe pragmatic communication, by modeling LISTENERs and SPEAKERs as directly optimizing both communicative success and adherence to existing linguistic conventions. 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:

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

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 tradeoff between optimizing for communicative success and proximity to default semantics $\tau_{\mathsf{S}}, \tau_{\mathsf{L}}$. When the value of λ_i is large, an agent $i \in \{\mathsf{S}, \mathsf{L}\}$ will consider only policies extremely close to τ_i; conversely, when λ_i is close to zero, the agent will not be penalized for adopting semantics that differ significantly from τ_i.

3.1 Notation and Representation of Policies

Before describing how to optimize the utilities given above, we first establish some notation that will be useful for describing the optimization procedure and the policies it produces.

Each agent's policy consists of a mapping from that agent's observations to a distribution over 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.

In order to provide a compact description of the algorithm, as well as an efficient vectorized implementation, we represent this mapping as a row-stochastic matrix, with rows indexed by observations and columns indexed by actions. 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. We similarly 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 these 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 $\tilde{u}_{\mathsf{L}}$ defined above, we use the piKL-Hedge algorithm (Jacob et al., 2022) to progressively refine a pair of SPEAKER and LISTENER policies 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 Equation (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 $\boldsymbol{p}$. For notational convenience, define:

$$\begin{aligned}\nabla \bar{u}_{\mathsf{S}}(\mathbf{L}) &\coloneqq \nabla_{\mathbf{S}}(\bar{u}_{\mathsf{S}}(\mathbf{S}, \mathbf{L}))\\ \nabla \bar{u}_{\mathsf{L}}(\mathbf{S}) &\coloneqq \nabla_{\mathbf{L}}(\bar{u}_{\mathsf{L}}(\mathbf{S}, \mathbf{L}))\end{aligned}$$

With this notation, the gradient of the unregularized utility function $\bar{u}_{\mathsf{S}}$ of the SPEAKER, is a function of the matrix-form policy $\mathbf{L}$ only.

$$\nabla \bar{u}_{\mathsf{S}}(\mathbf{L}) = -\boldsymbol{p}\boldsymbol{c}^{\top} + \mathbf{P}\mathbf{L}^{\top} \in \mathbb{R}^{M \times U}. \quad (2)$$

Similarly, for the LISTENER we have:

$$\nabla \bar{u}_{\mathsf{L}}(\mathbf{S}) \coloneqq \mathbf{S}^{\top}\mathbf{P} \in \mathbb{R}^{U \times M}. \quad (3)$$

With the above gradients, piKL-Hedge (Jacob et al., 2022) prescribes the following algorithm for progressively refining policies: first, at time 0, we set

$$\bar{\mathbf{S}}^{(0)} = \bar{\mathbf{L}}^{(0)} \coloneqq \mathbf{0}; \quad (4)$$

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:

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

where $\overset{\text{row}}{\propto}$ denotes row-wise proportionality and exponentiation is performed elementwise. These dynamics strike a balance between playing proportional to the exponential of the utility gradient, and remaining in a neighborhood of the default semantics $\boldsymbol{\tau}$. Concretely, taking the SPEAKER player as an example, when $\lambda_{\mathsf{S}} = 0$, then the update rule for $\mathbf{S}^{(t+1)}$ reduces to $\mathbf{S}^{(t+1)} \overset{\text{row}}{\propto} \exp\{\eta_{\mathsf{S}} \cdot t \nabla \bar{u}_{\mathsf{S}}(\bar{\mathbf{L}}^{(t)})\}$, which corresponds to Hedge. Conversely, in the other extreme when $\lambda_{\mathsf{S}} \to \infty$, then the update rule for $\mathbf{S}^{(t+1)}$ reduces to $\mathbf{S}^{(t+1)} \overset{\text{row}}{\propto} \exp\{\log \boldsymbol{\tau}_{\mathsf{S}}\} = \boldsymbol{\tau}_{\mathsf{S}}$, that is, the dynamics do not move at all from the default semantics.

piKL-Hedge dynamics have strong guarantees, including the following (see Jacob et al., 2022):

- 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 K-L divergence between Player i's policy and the default semantics $\boldsymbol{\tau}_i$ scales as approximately $1/\lambda_i$.

3.3 Special Case: Uniform Priors, No Costs

When the prior over the meanings is uniform, and utterance costs are all set to zero, the gradients $\nabla \bar{u}_{\mathsf{S}}(\mathbf{L})$ and $\nabla \bar{u}_{\mathsf{L}}(\mathbf{S})$, defined in (2) and (3), simplify into:

$$\nabla \bar{u}_{\mathsf{S}}(\mathbf{L}) = \frac{1}{|M|}\mathbf{L}, \quad \nabla \bar{u}_{\mathsf{L}}(\mathbf{S}) = \frac{1}{|M|}\mathbf{S}.$$

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

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

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

The above procedure has a striking similarity 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 (in its simplest form) as:

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

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. In this interpretation, speakers *and* listeners incur an additional "communication cost" proportional to the log-probability of a given utterance or interpretation under the prior $\boldsymbol{\tau}$. As we will see, however, the more general formulation of RECO in Section 3.2 enables it to make predictions that are not achievable with RSA in its standard form.

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 it makes precisely this prediction.

4.2 Manner Implicature

Another important class of implicatures are Manner implicatures, in which (for example) 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 that our language contains two utterances (*short* and *long*) and two meanings (`freq` and `rare`) satisfying the following properties: (1) `freq` occurs as the intended meaning with probability $\frac{2}{3}$ and `rare` occurs with probability $\frac{1}{3}$; (2) *long* has production cost of 0.2 and *short* has a production cost of 0.1; finally (3) either *long* or *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 cannot make these predictions natively, and require substantial modification to derive them.

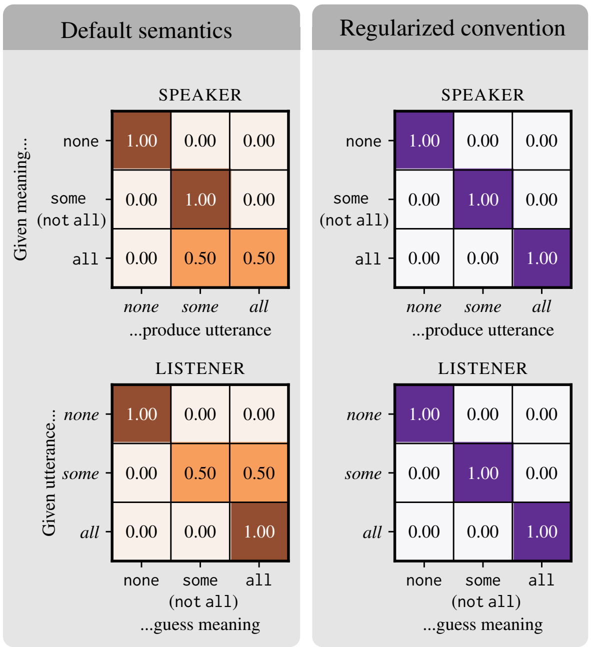

Figure 2: Quantity implicatures in ReCo. (Left) Matrices representing conditional probabilities that represent the default semantics τ_S and τ_L. (Right) Matrices representing conditional probabilities that represent the resulting regularized conventions π_S and π_L. In this setting, ReCo is able to predict the correct set of interpretations.

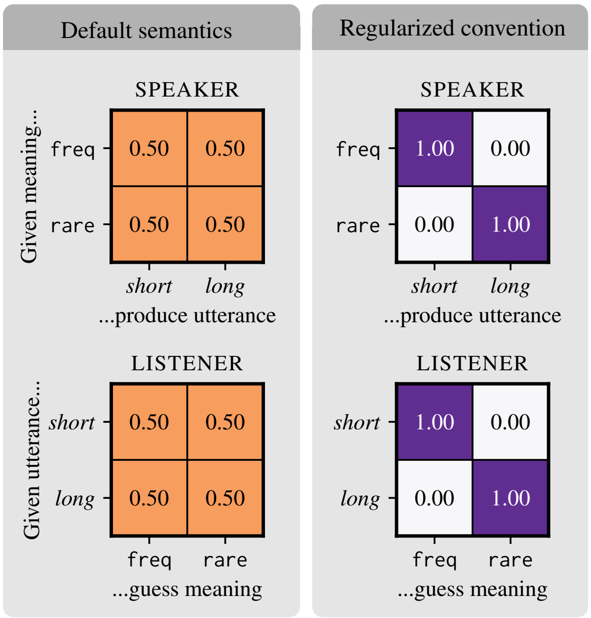

Figure 3: Manner implicatures in ReCo. (Left) Matrices representing conditional probabilities that represent the default semantics τ_S and τ_L. (Right) Matrices representing conditional probabilities that represent the resulting regularized conventions π_S and π_L. By incorporate prior probabilities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

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

	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.

.	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 that enables the LISTENER to distinguish the taraget color (in the black box) from others in the context.

	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. RECO performs best on both validation and test sets.

5 Probabilistic 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 probability that particular utterances might carry particular meanings. As RECO, like RSA-family models, captures probabilistic associations between utterances and meanings, we evaluate its predictions by measuring their *correlation* between human judgments. Specifically, for each task (and all tasks jointly), we compute the correlation between $p(\text{meaning} \mid \text{utterance})$ predicted by the model, and the average $p(\text{meaning} \mid \text{utterance})$ predicted by humans (with one data point for each (meaning, utterance, context) triple). We refer the reader to Frank (2016) for more details about the experimental setup.

Comparisons between RECO, RSA, BR SPEAKER (i.e., best-response to a literal speaker) and RD-RSA (Zaslavsky et al., 2021a) are shown in Table 1, with additional information about parameters in Figure 4. In these figures, ALL denotes correlations computed across all four tasks. RECO modestly improves upon the best predictions of RSA-family methods, both overall and on 3/4 tasks individually. In addition, it is robust across a wide range of speaker hyperparameters.

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 80% / 10% / 10% 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.

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. Model and hyperparameter details can be found in Appendix B.

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 RECO, a model of pragmatic language understanding based on game-theoretic equilibrium search. In this model, speakers and

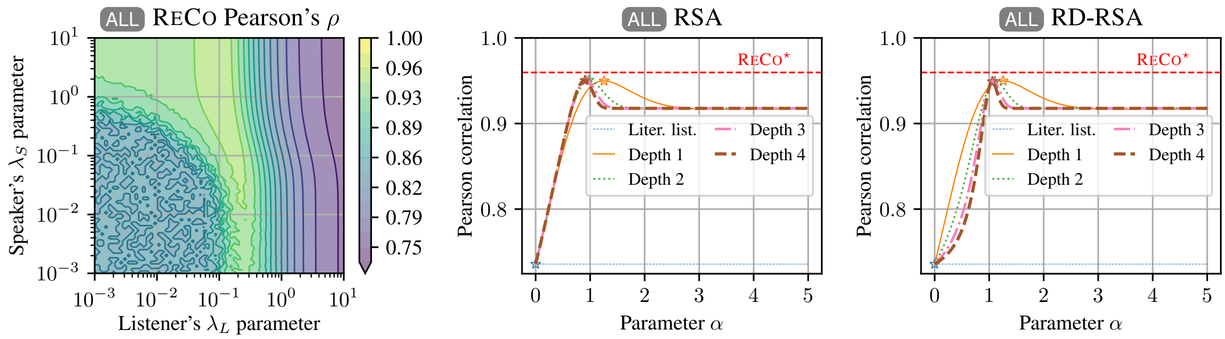

Figure 4: Pearson's correlation ρ on the full dataset of graded human judgments from (Frank, 2016). (Left) Correlation for ReCo as a function of λ_L and λ_S represented as a contour plot. (Middle) Correlation between RSA at different levels of α and recursive depth (Right) Correlation between RD-RSA 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.

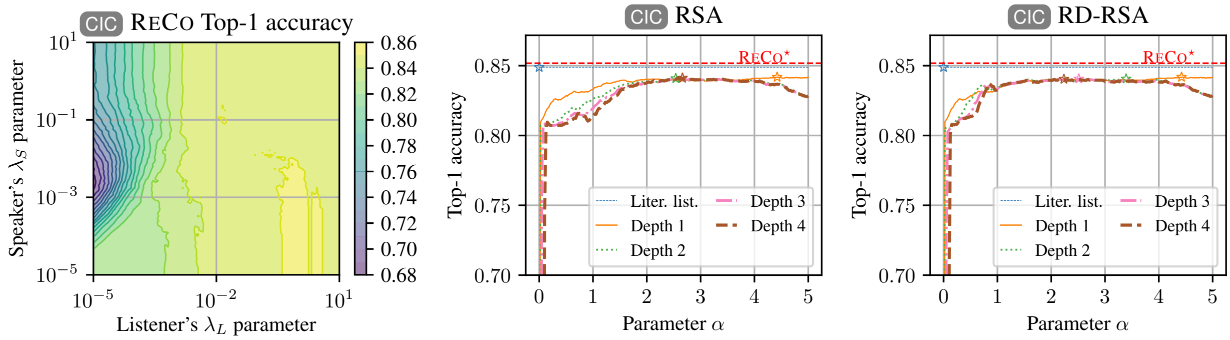

Figure 5: Top-1 accuracy of predicting meanings on the validation set of the Colors in Context task (Monroe et al., 2017). (Left) Accuracy for ReCo as a function of λ_L and λ_S represented as a contour plot. (Middle) Accuracy of RSA at different levels of α and recursive depth (Right) Accuracy of RD-RSA 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.

listeners solve communicative tasks by searching for utterance-meaning mappings that that simultaneously optimize reward and similarity to a distribution encoding default meanings.

Our work can be interpreted as a response to the observation by Jäger (2012) that *"it is not so clear whether the solution concept of a Nash equilibrium (or strengthenings thereof) is really appropriate to model the action of rational agents in one-shot [communication] games."* Equilibrium-finding models of language understanding have been widely studied in the context of iterated communication and language evolution, where there is a clear mechanism by which groups of language users might collectively arrive at one of many game-theoretically optimal equilibria (Jäger, 2007, 2008b,a; Trapa and Nowak, 2000). These studies characterize the stability of population-level conventions in game-theoretic terms. However, as noted in Jäger (2008a), such approaches have historically struggled to explain how language understanding occurs in one-shot settings. ReCo, by way of "regularized equilibrium" concepts, offers an explanation for these behaviors, and in doing so bridges the gap between past evolutionary work and the one-shot inference problems that are of special interest in pragmatics.

Looking ahead, ReCo can be used as a platform for studying related problems in context-dependent, multi-party communication. For example, it might be possible to study *iterated* conventions (Hawkins et al., 2017), established over multiple rounds of communication, by updating the default semantics $\boldsymbol{\tau}$ to the *equilibrium policy* at the previous round. While our experiments here have focused on single-turn interactions, tools for solving *extensive-form* games might similarly be used to model communicative strategies that play out over multiple turns of dialog. More generally, we hope these results highlight the effectiveness of game theoretic tools for understanding and enriching models of pragmatic language production and comprehension.

Acknowledgements

Thanks to Jennifer Hu for helpful suggestions about experiment design. We also thank the anonymous reviewers for their feedback. This work was supported by the National Science Foundation under grant IIS-2212310.

Limitations

The algorithms described in this paper assume that communication tasks are defined by a finite set of possible utterances and possible meanings. While tools exist for computing equilibria of games with combinatorial action spaces, additional work would be required to apply this method to open-ended text generation problems.

Ethics Statement

We do not anticipate any ethical concerns associated with methods described in this paper.

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 *Proceedings of the International Conference on Learning Representations*.

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

Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

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 Conference on Empirical Methods in Natural Language Processing*.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

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

Robert Hawkins, Minae Kwon, Dorsa Sadigh, and Noah Goodman. 2020. Continual adaptation for efficient machine communication. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 408–419.

Robert D Hawkins, Michael Franke, Michael C Frank, Adele E Goldberg, Kenny Smith, Thomas L Griffiths, and Noah D Goodman. 2023. From partners to populations: A hierarchical bayesian account of coordination and convention. *Psychological Review*, 130(4):977.

Robert XD Hawkins, Mike Frank, and Noah D Goodman. 2017. Convention-formation in iterated reference games. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

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

Athul Paul Jacob, Abhishek Gupta, and Jacob Andreas. 2023a. Modeling boundedly rational agents with latent inference budgets. *arXiv preprint arXiv:2312.04030*.

Athul Paul Jacob, Yikang Shen, Gabriele Farina, and Jacob Andreas. 2023b. 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 *Proceedings of the International Conference on Machine Learning*.

Gerhard Jäger. 2007. Game dynamics connects semantics and pragmatics. In *Game Theory and Linguistic Meaning*, pages 103–117.

Gerhard Jäger. 2008a. Applications of game theory in linguistics. *Language and Linguistics compass*, 2(3):406–421.

Gerhard Jäger. 2008b. Evolutionary stability conditions for signaling games with costly signals. *Journal of theoretical biology*, 253(1):131–141.

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*.

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 Annual Meeting of the Association for Computational Linguistics*.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Peter E Trapa and Martin A Nowak. 2000. Nash equilibria for an evolutionary language game. *Journal of mathematical biology*, 41(2):172–188.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

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

Noga Zaslavsky, Jennifer Hu, and Roger P. Levy. 2021b. A Rate–Distortion view of human pragmatic reasoning? In *Proceedings of the Society for Computation in Linguistics 2021*, pages 347–348, Online. Association for Computational Linguistics.

A Per-task results

In Figure 6, we compare ReCo, RSA, BR and RD-RSA (Zaslavsky et al., 2021b) across each of the four reference tasks based on graded human judgements that we consider in Section 5.

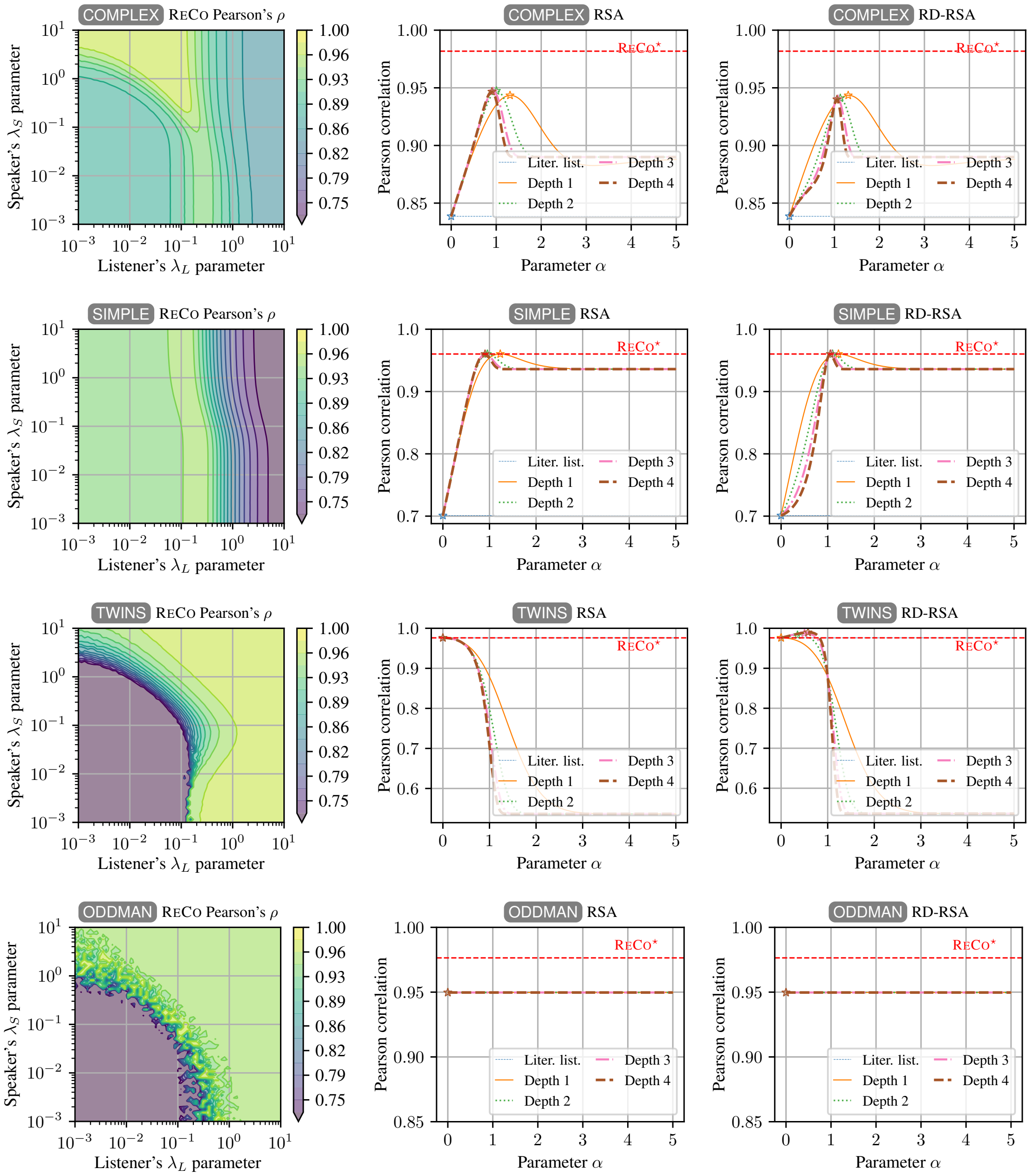

Figure 6: Pearson's correlation ρ on the each of the four reference tasks (SIMPLE, COMPLEX, TWINS and ODDMAN) of graded human judgments from (Frank, 2016). (First column) Correlation for ReCo as a function of λ_L and λ_S represented as a contour plot. (Second column) Correlation between RSA at different levels of α and recursive depth (Third column) Correlation between RD-RSA at different levels of α and recursive depth. (Second, Third columns) 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.

B Model, Training and Hyperparameter Details

The speaker and listener models from Section 6 are based on the transformer architecture. Following past work (Jacob et al., 2023a), the speaker model is based on the T5 model (Raffel et al., 2020) and the listener is based on BERT (Devlin et al., 2019). We use the hyperparameter settings used in Jacob et al. (2023a) for the speaker and listener models. The speaker model was trained with a batch size of 64 using the Adam optimizer with learning rate 10^{-4} for 25 epochs. We trained the models using PyTorch (Paszke et al., 2019) and Huggingface (Wolf et al., 2020) libraries. These models were trained using a single V100 GPU for 3-4 hours. All other experiments were performed on an 8-core Intel CPUs and M2 Macbook Pro. For experiments in Section 5, ReCo was run with 10 seeds and the run with the highest sum of regularized utilities of the SPEAKER and LISTENER was used.

The parameter η was set to 0.1 in all our experiments.

C Computational Complexity

For each player, each iteration of the algorithm runs in time linear in the product of the number of utterances $|U|$ and the number of meanings $|M|$. So, one iteration of ReCo is as expensive as one iteration of RSA. We note however, that we might need a higher number of iterations to get to convergence; unlike RSA (where increasing the number of iterations leads to undesirable behavior), in ReCo more iterations simply improve the approximation of the equilibrium point. We use 2000 iterations of ReCo in our experiments. As a practical implementation note, we remark that ReCo (just like RSA) is very easy to implement in `pytorch` using efficient vectorized tensor operations, and each of our experiments took only milliseconds to complete 2000 iterations and approximate the regularized equilibrium.

D Examples of predictions

In Figure 7, we show the predictions generated by ReCo (for $\lambda_S = 1, \lambda_L = 0.1$) and by RSA (for the best selection of hyperparameters: depth 3, alpha 0.95) in the Graded Human Judgements domain (`complex` dataset). As shown in Figure 6 in our paper, in this domain ReCo achieves substantially higher Pearson correlation than even the best RSA hyperparameters. Qualitatively, you can see that, given the utterance of *glasses*, RSA predicts meaning `target` with a significantly higher probability than what ReCo predicts.

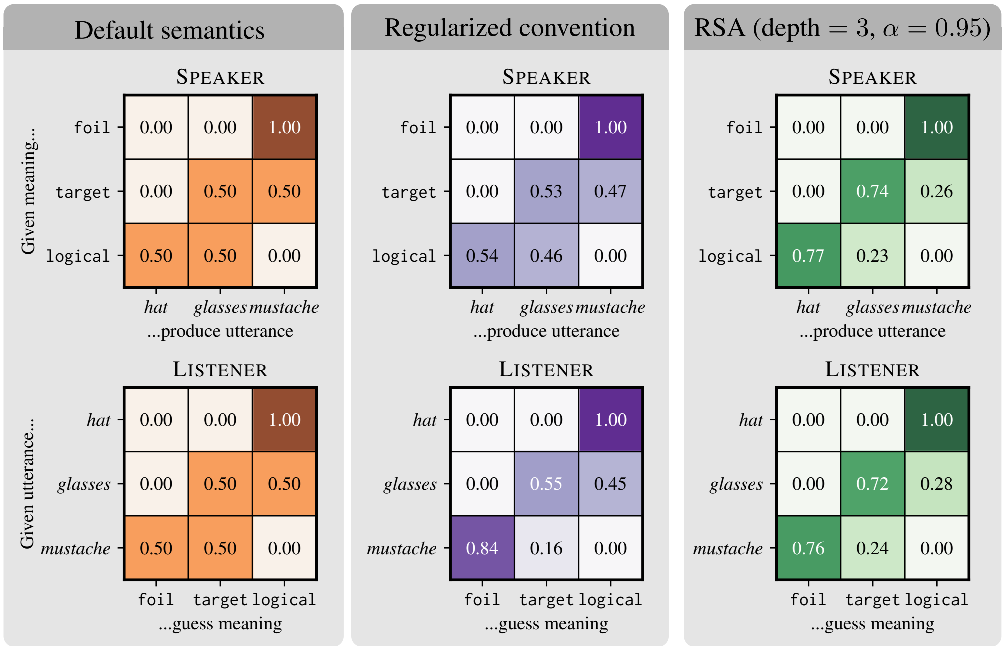

Figure 7: Predictions generated by ReCo (for $\lambda_S = 1, \lambda_L = 0.1$) and by RSA (for the best selection of hyperparameters: depth 3, alpha 0.95) in the Graded Human Judgements domain (`complex` dataset)

The figure shows well the effect the regularization parameters λ in ReCo. Indeed, because of the high regularization in the speaker ($\lambda_S = 1.0$), the ReCo strategy for the speaker has not moved much from the default semantics. In contrast, the listener's low regularization ($\lambda_L = 0.1$) enables the listener to deviate significantly, and pick a more counterspeculative strategy (i.e., one that is more tuned to the speaker).