Colors in Context: A Pragmatic Neural Model for Grounded Language Understanding

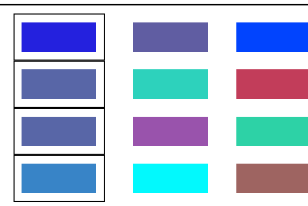
Abstract

We present a model of pragmatic referring expression interpretation in a grounded communication task (identifying colors from descriptions) that draws upon predictions from two recurrent neural network classifiers, a speaker and a listener, unified by a recursive pragmatic reasoning framework. Experiments show that this combined pragmatic model interprets color descriptions more accurately than the classifiers from which it is built. We observe that pragmatic reasoning helps primarily in the hardest cases: when the model must distinguish very similar colors, or when few utterances adequately express the target color. Our findings make use of a newly-collected corpus of human utterances in color reference games, which exhibit a variety of pragmatic behaviors. We also show that the embedded speaker model reproduces many of these pragmatic behaviors.

1 Introduction

Human communication is *situated*. In using language, we are sensitive to context and our interlocutors' expectations, both when choosing our utterances (as speakers) and when interpreting the utterances we hear (as listeners). Visual referring tasks exercise this complex process of grounding, in the environment and in our mental models of each other, and thus provide a valuable test-bed for computational models of production and comprehension.

Table 1 illustrates the situated nature of reference understanding with descriptions of colors from a task-oriented dialogue corpus we introduce in this paper. In these dialogues, the speaker is trying to identify their (privately assigned) target color for the listener. In context 1, the comparative *darker* implicitly refers to both the target (boxed) and one of the other colors. In contexts 2 and 3, the target color is the same, but the distractors led the speaker to choose different basic color terms. In context 4, *blue* is a pragmatic choice even though two colors are shades of blue, because the interlocutors assume about each other that they find the target color a more prototypical representative of blue and would prefer other descriptions (*teal*, *cyan*) for the middle color. The fact that *blue* appears in three of these four cases highlights the flexibility and context dependence of color descriptions.



	Context			Utterance
1.				darker blue
2.				Purple
3.				blue
4.				blue

Table 1: Examples of color reference in context, taken from our corpus. The target color is boxed. The speaker's description is shaped not only by this target, but also by the other context colors and their relationships.

In this paper, we present a scalable, learned model of pragmatic language understanding. At its core, the model is a version of the Rational Speech Acts (RSA) model (Frank and Goodman, 2012; Goodman and Frank, 2016), in which agents reason recursively about each other's expectations and intentions to communicate more effectively than literal semantic agents could. In most work on RSA, the literal semantic agents use fixed message sets and stipulated grammars, which is a barrier to experiments in linguistically complex domains. In our formulation, the literal semantic agents are recurrent neural networks (RNNs) that produce and interpret color descriptions in context. These models are learned from data and scale easily to large datasets containing diverse utterances. The RSA recursion is then defined in terms of these base agents: the *pragmatic speaker* produces utterances based on a literal RNN listener (Andreas and Klein, 2016), and the *pragmatic listener* interprets utterances based on the pragmatic speaker's behavior.

We evaluate this model with a new, psycholinguistically motivated corpus of real-time, dyadic reference games in which the referents are patches of color. The corpus includes 948 complete games with 53,365 utterances produced by human participants paired into dyads on the web. The linguistic behavior of the players exhibits many of the intricacies of language in general, including not just the context dependence and cognitive complexity discussed above, but also compositionality, vagueness, and ambiguity. While many previous data sets feature descriptions of individual colors (Cook et al., 2005; Munroe, 2010; Kawakami et al., 2016), situating colors in a communicative context elicits greater variety in language use, including negations, comparatives, superlatives, metaphor, and shared associations.

We focus on accuracy in a listener task (i.e., at language understanding). However, our most successful model integrates speaker and listener perspectives, combining predictions made by a system trained to understand color descriptions and one trained to produce them. Experiments on the data in our corpus show that this combined pragmatic model improves accuracy in interpreting human-produced descriptions over the basic RNN listener alone. Moreover, the improvements from pragmatic reasoning come primarily in the hardest cases: (1) contexts with colors that are very similar, thus requiring the interpretation of descriptions that convey fine distinctions; and (2) target colors that most referring expressions fail to identify, whether due to a lack of adequate descriptive terms or a consistent bias against the color in the RNN listener.

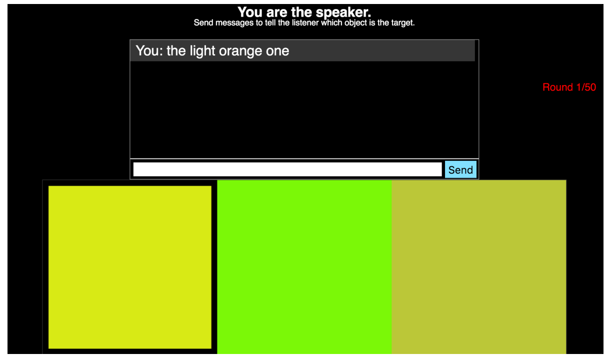


Figure 1: Example trial in corpus collection task, from speaker's perspective. The target color (boxed) was presented among two distractors on a neutral background.

2 Task and data collection

We evaluate our agents on a task of language understanding in a dyadic reference game (Rosenberg and Cohen, 1964; Krauss and Weinheimer, 1964; Paetzel et al., 2014). Unlike traditional natural language processing tasks, in which participants provide impartial judgements of language in isolation, reference games embed language use in a goal-oriented communicative context (Clark, 1996; Tanenhaus and Brown-Schmidt, 2008). Since they offer the simplest experimental setup where many pragmatic and discourse-level phenomena emerge, these games have been used widely in cognitive science to study topics like common ground and conventionalization (Clark and Wilkes-Gibbs, 1986), referential domains (Brown-Schmidt and Tanenhaus, 2008), perspective-taking (Hanna et al., 2003), and overinformativeness (Koolen et al., 2011).

To obtain a corpus of natural color reference data across varying contexts, we recruited 967 unique participants from Amazon Mechanical Turk to play 1,059 games of 50 rounds each, using the open-source framework of Hawkins (2015). Participants were sorted into dyads, randomly assigned the role of speaker or listener, and placed in a game environment containing a chat box and an array of three color patches (Figure 1). On each round, one of the three colors was chosen to be the target and highlighted for the speaker. They were instructed to communicate this information to the listener, who could then click on one of the colors to advance to the next trial. Both participants were free to use the chat box at any point.

To ensure a range of difficulty, we randomly interspersed an equal number of trials from three different conditions: (1) *close*, where colors were all within a distance of θ from one another but still perceptible,[1] (2) *split*, where one distractor was within a distance of θ of the target, but the other distractor was farther than θ, and (3) *far*, where all colors were farther than θ from one another. Colors were rejection sampled uniformly from RGB (red, green, blue) space to meet these constraints.

After excluding extremely long messages,[2] incomplete games, and games whose participants self-reported confusion about the instructions or non-native English proficiency, we were left with a corpus of 53,365 speaker utterances across 46,994 rounds in 948 games. The three conditions are equally represented, with 15,519 *close* trials, 15,693 *split* trials, and 15,782 *far* trials. Participants were allowed to play more than once, but the modal number of games played per participant was one (75%). The modal number of messages sent per round was also one (90%). We release the filtered corpus we used throughout our analyses alongside the raw, pre-filter data collected from these experiments (see Footnote 11).

3 Behavioral results

Our corpus was developed not only to facilitate the development of models for grounded language understanding, but also to provide a richer picture of human pragmatic communication. The collection effort was thus structured like a large-scale behavioral experiment, closely following experimental designs like those of Clark and Wilkes-Gibbs (1986). This paves the way to assessing our model not solely based on the listener's classification accuracy, but also in terms of how qualitative features of the speaker's production compare to that of our human participants. Thus, the current section briefly reviews some novel findings from the human corpus that we use to inform our model assessment.

3.1 Listener behavior

Since color reference is a difficult task even for humans, we compared listener accuracy across conditions to calibrate our expectations about model performance. While participants' accuracy was close to ceiling (97%) on the *far* condition, they made significantly more errors on the *split* (90%) and *close* (83%) conditions (see Figure 4).

3.2 Speaker behavior

For ease of comparison to computational results, we focus on five metrics capturing different aspects of pragmatic behavior displayed by both human and artificial speakers in our task (Table 2). In all cases, we report test statistics from a mixed-effects regression including condition as a fixed effect and game ID as a random effect; except where noted, all test statistics reported correspond to p-values $< 10^{-4}$ and have been omitted for readability.

Words and characters We expect human speakers to be more verbose in *split* and *close* contexts than *far* contexts; the shortest, simplest color terms for the target may also apply to one or both distractors, thus incentivizing the speaker to use more lengthy descriptions to fully distinguish it. Indeed, even if they *know* enough simple color terms to distinguish all the colors lexically, they might be unsure their listeners will and so resort to modifiers anyway. To assess this hypothesis, we counted the average number of words and characters per message. Compared to the baseline *far* context, participants used significantly longer messages both in the *split* context ($t = 45.85$) and the *close* context ($t = 73.06$). Similar results hold for the character metric.

Comparatives and superlatives As noted in Section 1, comparative morphology implicitly encodes a dependence on the context; a speaker who refers to the target color as *the darker blue* is presupposing that there is another (lighter) blue in the context. Similarly, superlatives like *the bluest one* or *the lightest one* presuppose that all the colors can be compared along a specific semantic dimension. We thus expect to see this morphology more often where

[1] We used the most recent CIEDE standard to measure color differences, which is calibrated to human vision (Sharma et al., 2005). All distances were constrained to be larger than a lower bound of $\epsilon = 5$ to ensure perceptible differences, and we used a threshold value of $\theta = 20$ to create conditions.

[2] Specifically, 627 messages with greater than 4σ of the mean number of words per message; these often included meta-commentary about the game rather than color terms.

	human			S_0			S_1		
	far	split	close	far	split	close	far	split	close
# Chars	7.8	12.3	14.9	9.0	12.8	16.6	9.0	12.8	16.4
# Words	1.7	2.7	3.3	2.0	2.8	3.7	2.0	2.8	3.7
% Comparatives	1.7	14.2	12.8	3.6	8.8	13.1	4.2	9.0	13.7
% High Specificity	7.0	7.6	7.4	6.4	8.4	7.6	6.8	7.9	7.5
% Negatives	2.8	10.0	12.9	4.8	8.9	13.3	4.4	8.5	14.1
% Superlatives	2.2	6.1	16.7	4.7	9.7	17.2	4.8	10.3	16.6

Table 2: Corpus statistics and statistics of samples from artificial speakers (rates per utterance). S_0: RNN speaker; S_1: pragmatic speaker derived from RNN listener (see Section 4.4). The human and artificial speakers show many of the same correlations between language use and context type.

two or more of the colors are comparable in this way. To test this, we used the Stanford CoreNLP part-of-speech tagger (Toutanova et al., 2003) to mark the presence or absence of comparatives (JJR or RBR) and superlatives (JJS or RBS) for each message.

We found two related patterns across conditions. First, participants were significantly more likely to use both comparatives ($z = 37.39$) and superlatives ($z = 31.32$) when one or more distractors were close to the target. Second, we found evidence of an asymmetry in the use of these constructions across the *split* and *close* contexts. Comparatives were used significantly more often in the *split* context ($z = 4.4$), where only one distractor was close to the target, while superlatives were much more likely to be used in the *close* condition ($z = 32.72$).[3]

Negatives In our referential contexts, negation is likely to play a role similar to that of comparatives: a phrase like *not the red or blue one* singles out the third color, and *blue but not bright blue* achieves a more nuanced kind of comparison. Thus, as with comparatives, we expect negation to be more likely where one or more distractors are close to the target. To test this, we counted occurrences of the string 'not' (by far the most frequent negation in the corpus). Compared to the baseline *far* context, we found that participants were more likely to use negative constructions when one ($z = 27.36$) or both ($z = 34.32$) distractors were close to the target.

WordNet specificity We expect speakers to prefer basic color terms wherever they suffice to achieve the communicative goal, since such terms are most likely to succeed with the widest range of listeners. Thus, a speaker might choose *blue* even for a clear periwinkle color. However, as the colors get closer together, the basic terms become too ambiguous, and thus the risk of specific terms becomes worthwhile (though lengthy descriptions might be a safer strategy, as discussed above). To evaluate this idea, we use WordNet (Fellbaum, 1998) to derive a specificity hierarchy for color terms, and we hypothesized that *split* or *close* conditions will tend to lead speakers to go lower in this hierarchy.

For each message, we transformed adjectives into their closest noun forms (e.g. 'reddish' → 'red'), filtered to include only nouns with 'color' in their hypernym paths, calculated the depth of the hypernym path of each color word, and took the maximum depth occurring in a message. For instance, the message "deep magenta, purple with some pink" received a score of 9. It has three color terms: "purple" and "pink," which have the basic-level depth of 7, and "magenta," which is a highly specific color term with a depth of 9. Finally, because there weren't meaningful differences between words at depths of 8 ("rose", "teal") and 9 ("tan," "taupe"), we conducted our analyses on a binary variable thresholded to distinguish "high specificity" messages with a depth greater than 7. We found a small but reliable increase in the likelihood of "high specificity" messages from human speakers in the *split* ($z = 2.84, p = 0.005$) and *close* ($z = 2.33, p = 0.02$) contexts, compared to the baseline *far* context.

[3]We used Helmert coding to examine this pair of orthogonal contrasts, compared to dummy coding in the other analyses.

blue	**1**	**1**	0
teal	0	**1**	0
dull	**1**	0	**1**

(a) The lexicon $\mathcal{L}$ defines utterances' truth values. Our neural listener skips $\mathcal{L}$ and models l_0's probability distributions directly.

blue	**50%**	**50**	0
teal	0	**100**	0
dull	**50**	0	**50**

(b) The literal listener l_0 chooses colors compatible with the literal semantics of the utterance; other than that, it guesses randomly.

blue	**50**	33	0
teal	0	**67**	0
dull	**50**	0	**100**

(c) The pragmatic speaker s_1 soft-maximizes the informativity of its utterances. (For simplicity, $\alpha = 1$ and $\kappa(u) = 0$.)

blue	**60**	40	0
teal	0	**100**	0
dull	33	0	**67**

(d) The pragmatic listener l_2 uses Bayes' rule to infer the target using the speaker's utterance as evidence.

Figure 2: RSA applied to a reference task (literal semantics and alternative utterances simplified for demonstration).

4 Models

We first define the basic RSA model as applied to the color reference games introduced in Section 2; a worked example is shown in Figure 2. The starting point of RSA is a model of a *literal listener*:

$$l_0(t \mid u, \mathcal{L}) \propto \mathcal{L}(u, t)P(t) \tag{1}$$

where t is a color in the context set C, u is a message drawn from a set of possible utterances U, P is a prior over colors, and $\mathcal{L}(u, t)$ is a semantic interpretation function that takes the value 1 if u is true of t, else 0. Figure 2a shows the values of $\mathcal{L}$ defined for a very simple context in which $U = \{\textit{blue}, \textit{teal}, \textit{dull}\}$, and $C = \{\ ,\ ,\ \}$; Figure 2b shows the corresponding literal listener l_0 if the prior P over colors is flat. (In our scalable extension, we will substitute a neural network model for l_0, bypassing $\mathcal{L}$ and allowing for non-binary semantic judgments.)

RSA postulates a model of a *pragmatic speaker* (Figure 2c) that behaves according to a distribution that soft-maximizes a utility function rewarding informativity and penalizing cost:

$$s_1(u \mid t, \mathcal{L}) \propto e^{\alpha \log(l_0(t \mid u, \mathcal{L})) - \kappa(u)} \tag{2}$$

Here, κ is a real-valued cost function on utterances, and $\alpha \in [0, \infty)$ is an inverse temperature parameter governing the "rationality" of the speaker model. A large α means the pragmatic speaker is expected to choose the most informative utterance (minus cost) consistently; a small α means the speaker is modeled as choosing suboptimal utterances frequently.

Finally, a *pragmatic listener* (Figure 2d) interprets utterances by reasoning about the behavior of the pragmatic speaker:

$$l_2(t \mid u, \mathcal{L}) \propto s_1(u \mid t, \mathcal{L})P(t) \tag{3}$$

The α parameter of the speaker indirectly affects the listener's interpretations: the more reliably the speaker chooses the optimal utterance for a referent, the more the listener will take deviations from the optimum as a signal to choose a different referent.

The most important feature of this model is that the pragmatic listener l_2 reasons not about the semantic interpretation function $\mathcal{L}$ directly, but rather about a speaker who reasons about a listener who reasons about $\mathcal{L}$ directly. The back-and-forth nature of this interpretive process mirrors the cooperative principle of conversational implicature (Grice, 1975) and reflects more general ideas from Bayesian cognitive modeling (Tenenbaum et al., 2011). The model and its variants have been shown to capture a wide range of pragmatic phenomena in a cognitively realistic manner (Goodman and Stuhlmüller, 2013; Smith et al., 2013; Kao et al., 2014; Bergen et al., 2014), and the central Bayesian calculation has proven useful in a variety of communicative domains (Tellex et al., 2014; Vogel et al., 2013).

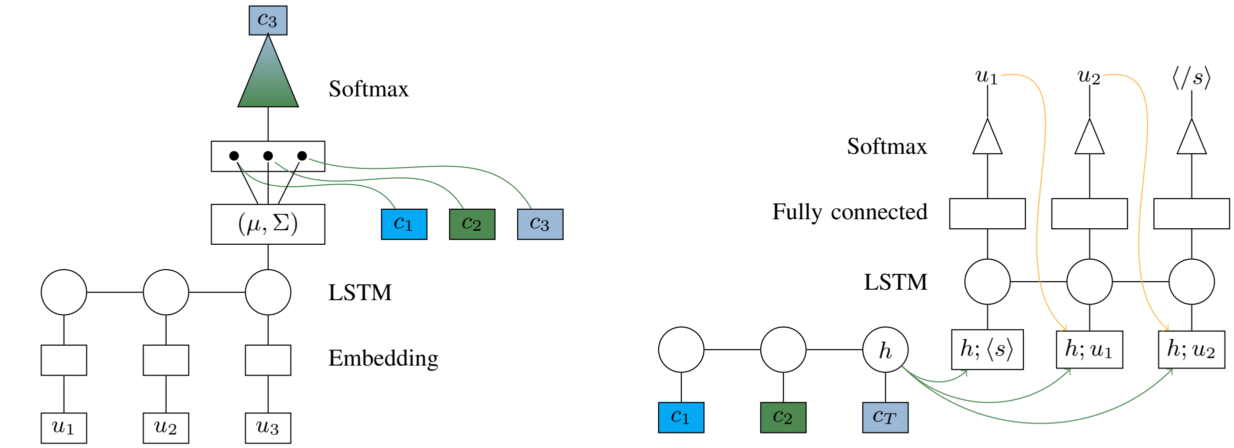
The definitions of s_1 (2) and l_2 (3) give a general method of deriving a speaker from a listener and vice versa. This suggests an alternative formulation of a pragmatic listener, starting from a literal speaker:

$$s_0(u \mid t, \mathcal{L}) \propto \mathcal{L}(u, t)e^{-\kappa(u)} \tag{4}$$

$$l_1(t \mid u, \mathcal{L}) \propto s_0(u \mid t, \mathcal{L})P(t) \tag{5}$$

Here, it is the speaker that reasons about the semantics, while the listener reasons about this speaker.

Both of these versions of RSA pose problems with scalability, stemming from the set of messages U



(a) The L_0 agent processes a color description sequentially. The final representation is transformed into a Gaussian distribution in color space, which is used to score the context colors.

(b) The S_0 agent processes the target color c_T in context and produces a color description sequentially. Each step in production is conditioned by the final contextual representation h and the previous word produced.

Figure 3: The neural base speaker and listener agents.

and the interpretation function $\mathcal{L}$. In most versions of RSA, these are specified by hand (but see Monroe and Potts 2015). This presents a serious practical obstacle to applying RSA to large data sets containing realistic utterances. The set U also raises a more fundamental issue: if this set is not finite (as one would expect from a compositional grammar), then in general there is no exact way to normalize the s_1 scores, since the denominator must sum over all messages. The same problem applies to s_0, unless $\mathcal{L}$ factorizes in an unrealistically clean way.

Over the next few subsections, we overcome these obstacles by replacing l_0 and s_0 with RNN-based listener agents, denoted with capital letters: L_0, S_0. We use the S_0 agent both as a base model for a pragmatic listener analogous to l_1 in (5) and to acquire sample utterances for approximating the normalization required in defining the s_1 agent in (2).

4.1 Base listener

Our base listener agent L_0 (Figure 3a) is an LSTM encoder model that predicts a Gaussian distribution over colors in a transformed representation space. The input words are embedded in a 100-dimensional vector space. Word embeddings are initialized to random normally-distributed vectors ($\mu = 0$, $\sigma = 0.01$) and trained. The sequence of word vectors is used as input to an LSTM with 100-dimensional hidden state, and a linear transformation is applied to the output representation to produce the parameters μ and Σ of a quadratic form[4]

$$\text{score}(f) = -(f-\mu)^T \Sigma (f-\mu)$$

where f is a color in Fourier representation space (Section 4.2). The values of $\text{score}(f)$ for each of the K context colors are normalized in log space to produce a probability distribution over the context colors. We denote this distribution by $L_0(t \mid u, C; \theta)$, where θ represents the vector of parameters that define the trained model.

4.2 Representing colors

Each color is represented in its simplest form as a three-dimensional vector in RGB space; we transform this to HSV (hue, saturation, value) as a preprocessing step. These HSV vectors are then Fourier-transformed as in Monroe et al. (2016) before being fed as input to the models. The Fourier transformation maps a color (h, s, v) to a higher-dimensional[5] vector f:

$$\begin{aligned} \hat{f}_{jk\ell} &= \exp\left[-2\pi i \left(jh^* + ks^* + \ell v^*\right)\right] \\ f &= \left[\Re \hat{f} \quad \Im \hat{f}\right] \qquad j, k, \ell \in \{0, 1, 2\} \end{aligned}$$

[4] The quadratic form is not guaranteed to be negative definite and thus define a Gaussian; however, it is for $> 95\%$ of inputs. The distribution over context colors is well-defined regardless.

[5] For the listener, we restrict j, k to $\{0, 1\}$ and ℓ to 0, as we found this improved performance slightly.

where $(h^*, s^*, v^*) = (h/360, s/200, v/200)$. The Fourier transformation is meant to help the models identify non-convex components of denotations of color language, particularly periodic components.

4.3 Base speaker

We also employ an LSTM-based speaker model $S_0(u \mid t, C; \phi)$. This speaker serves two purposes: 1) it is used to define a pragmatic listener akin to l_1 in (5), and 2) it provides sets of alternative utterances for each context, to avoid enumerating the intractably large space of possible utterances.

The speaker model consists of an LSTM context encoder and an LSTM description decoder (Figure 3b). In this model, the colors of the context $c_i \in C$ are transformed into Fourier representation space, and the sequence of color representations is passed through an LSTM with 100-dimensional hidden state. The context is reordered to place the target color last, minimizing the length of dependence between the most important input color and the output (Sutskever et al., 2014) and eliminating the need to represent the index of the target separately. The output of this recurrent neural network is concatenated with a 100-dimensional embedding for the previous token at each time step in decoding. The resulting vector is input along with the previous cell state to the LSTM cell, and an affine transformation and softmax function are applied to the output to produce a probability distribution predicting the following token of the description. The model is substantively similar to well-known models for image caption generation (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015), which use the output of a convolutional neural network as the representation of an input image and provide this representation to the RNN as an initial state or first word (we represent the context using a second RNN and concatenate the context representation onto each input word vector).

4.4 Pragmatic agents

Using the above base agents, we define a pragmatic speaker S_1 and a pragmatic listener L_2:

$$S_1(u \mid t, C; \theta) = \frac{L_0(t \mid u, C; \theta)^\alpha}{\sum_{u'} L_0(t \mid u', C; \theta)^\alpha} \quad (6)$$

$$L_2(t \mid u, C; \theta) = \frac{S_1(u \mid t, C; \theta)}{\sum_{t'} S_1(u \mid t', C; \theta)} \quad (7)$$

These definitions mirror those in (2) and (3) above, with $\mathcal{L}$ replaced by the learned weights θ.

Just as in (2), the denominator in (6) should consist of a sum over the entire set of potential utterances, which is exponentially large in the maximum utterance length and might not even be finite. As mentioned in Section 4.3, we limit this search by taking m samples from $S_0(u \mid i, C; \phi)$ for each target index i, adding the actual utterance from the testing example, and taking the resulting multiset as the universe of possible utterances, weighted towards frequently-sampled utterances.[6] Taking a number of samples from S_0 for each referent in the context gives the pragmatic listener a variety of informative alternative utterances to consider when interpreting the true input description. We have found that m can be small; in our experiments, it is set to 8.

To reduce the noise resulting from the stochastically chosen alternative utterance sets, we also perform this alternative-set sampling n times and average the resulting probabilities in the final L_2 output. We again choose $n = 8$ as a satisfactory compromise between effectiveness and computation time.

A second pragmatic listener L_1 can be formed in a similar way, analogous to l_1 in (5):

$$L_1(t \mid u, C; \theta) = \frac{S_0(u \mid t, C; \theta)}{\sum_{t'} S_0(u \mid t', C; \theta)} \quad (8)$$

We expect L_1 to be less accurate than L_0 or L_2, because it is performing a listener task using only the outputs of a model trained for a speaker task. However, this difference in training objective can also give the model strengths that complement those of the two listener-based agents. One might also expect a realistic model of human language interpretation to lie somewhere between the "reflex" interpretations of the neural base listener and the "reasoned" interpretations of one of the pragmatic models. This has an intuitive justification in people's uncertainty about whether their interlocutors are speaking pragmatically: "should I read more into that statement, or take it at face value?" We therefore also evaluate models defined as a weighted average of L_0

[6] An alternative would be to enforce uniqueness within the alternative set, keeping it a true set as in the basic RSA formulation; this could be done with rejection sampling or beam search for the highest-scoring speaker utterances. We found that doing so with rejection sampling hurt model performance somewhat, so we did not pursue the more complex beam search approach.

and each of L_1 and L_2, as well as an "ensemble" model that combines all of these agents. Specifically, we consider the following blends of neural base models and pragmatic models, with $\mathbf{L}_i$ abbreviating $L_i(t \mid u, C; \theta)$ for convenience:

$$\mathbf{L}_a \propto \mathbf{L}_0{}^{\beta_a} \cdot \mathbf{L}_1^{1-\beta_a} \tag{9}$$

$$\mathbf{L}_b \propto \mathbf{L}_0{}^{\beta_b} \cdot \mathbf{L}_2^{1-\beta_b} \tag{10}$$

$$\mathbf{L}_e \propto \mathbf{L}_a{}^{\gamma} \cdot \mathbf{L}_b^{1-\gamma} \tag{11}$$

The hyperparameters in the exponents allow tuning the blend of each pair of models—e.g., overriding the neural model with the pragmatic reasoning in L_b. The value of the weights β_a, β_b, and γ can be any real number; however, we find that good values of these weights lie in the range $[-1, 1]$. As an example, setting $\beta_b = 0$ makes the blended model L_b equivalent to the pragmatic model L_2; $\beta_b = 1$ ignores the pragmatic reasoning and uses the base model L_0's outputs; and $\beta_b = -1$ "subtracts" the base model from the pragmatic model (in log probability space) to yield a "hyperpragmatic" model.

4.5 Training

We split our corpus into approximately equal train/dev/test sets (15,665 train trials, 15,670 dev, 15,659 test), ensuring that trials from the same dyad are present in only one split. We preprocess the data by (1) lowercasing; (2) tokenizing by splitting off punctuation as well as the endings *-er*, *-est*, and *-ish*;[7] and (3) replacing tokens that appear once or not at all in the training split[8] with `<unk>`. We also remove listener utterances and concatenate speaker utterances on the same context. We leave handling of interactive dialogue to future work (Section 7).

We use ADADELTA (Zeiler, 2012) and Adam (Kingma and Ba, 2014), adaptive variants of stochastic gradient descent (SGD), to train listener and speaker models. The choice of optimization algorithm and learning rate for each model were tuned with grid search on a held-out tuning set consisting of 3,500 contexts.[9] We also use a fine-grained grid search on this tuning set to determine the values of the pragmatic reasoning parameters α, β, and γ. In our final ensemble L_e, we use $\alpha = 0.544$, base weights $\beta_a = 0.492$ and $\beta_b = -0.15$, and a final blending weight $\gamma = 0.491$. It is noteworthy that the optimal value of β_b from grid search is *negative*. The effect of this is to amplify the difference between L_0 and L_2: the listener-based pragmatic model, evidently, is not quite pragmatic enough.

5 Model results

5.1 Speaker behavior

To compare human behavior with the behavior of our embedded speaker models, we performed the same behavorial analysis done in Section 3.2. Results from this analysis are included alongside the human results in Table 2. Our pragmatic speaker model S_1 did not differ qualitatively from our base speaker S_0 on any of the metrics, so we only summarize results for humans and the pragmatic model.

Words and characters We found human speakers to be more verbose when colors were closer together, in both number of words and number of characters. As Table 2 shows, our S_1 agent shows the same increase in utterance length in the *split* ($t = 18.07$) and *close* ($t = 35.77$) contexts compared to the *far* contexts.

Comparatives and superlatives Humans used more comparatives and superlatives when colors were closer together; however, comparatives were preferred in the *split* contexts, superlatives in the *close* contexts. Our pragmatic speaker shows the first of these two patterns, producing more comparatives ($z = 14.45$) and superlatives ($z = 16$) in the *split* or *close* conditions than in the baseline *far* condition. It does not, however, capture the peak in comparative use in the *split* condition. This suggests that our model is simulating the human strategy at some level, but that more subtle patterns require further attention.

Negations Humans used more negations when the colors were closer together. Our pragmatic speaker's use of negation shows the same relationship to the context ($z = 8.55$ and $z = 16.61$ respectively).

[7] We only apply this heuristic ending segmentation for the listener; the speaker is trained to produce words with these endings unsegmented, to avoid segmentation inconsistencies when passing speaker samples as alternative utterances to the listener.

[8] 1.13% of training tokens, 1.99% of dev/test.

[9] For L_0: ADADELTA, learning rate $\eta = 0.2$; for S_0: Adam, learning rate $\alpha = 0.004$.

model	accuracy (%)	perplexity
L_0	83.30	1.73
$L_1 = L(S_0)$	80.51	1.59
$L_2 = L(S(L_0))$	83.95	1.51
$L_a = L_0 \cdot L_1$	84.72	1.47
$L_b = L_0 \cdot L_2$	83.98	1.50
$L_e = L_a \cdot L_b$	**84.84**	**1.45**
human	*90.40*	
L_0	85.08	1.62
L_e	**86.98**	**1.39**
human	*91.08*	

Table 3: Accuracy and perplexity of the base and pragmatic listeners and various blends (weighted averages, denoted $A \cdot B$). Top: dev set; bottom: test set.

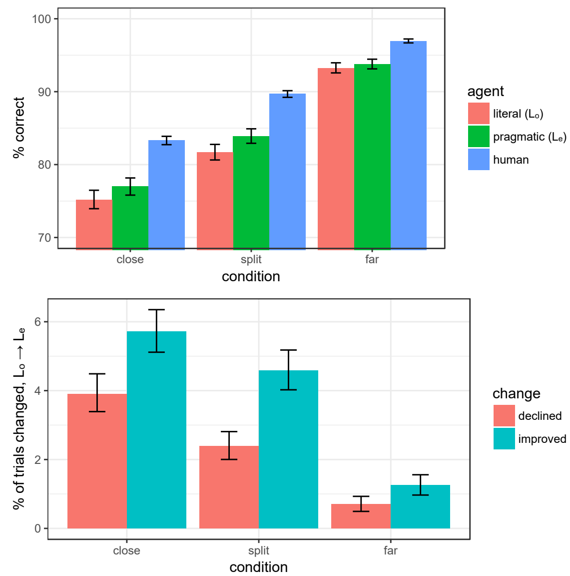


Figure 4: Human and model reference game performance (top) and fraction of examples improved and declined from L_0 to L_e (bottom) on the dev set, by condition.

WordNet specificity Humans used more "high specificity" words (by WordNet hypernymy depth) when the colors were closer together. Our pragmatic speaker showed a similar effect ($z = 2.65, p = 0.008$ and $z = 2.1, p = 0.036$ respectively).

5.2 Listener accuracy

Table 3 shows the accuracy and perplexity of the base listener L_0, the pragmatic listeners L_1 and L_2, and the blended models L_a, L_b, and L_e at resolving the human-written color references. As we expected, the speaker-based L_1 alone performs the worst of all the models. However, blending it with L_0 doesn't drag down L_0's performance but rather produces a considerable improvement compared to both of the original models, consistent with our expectation that the listener-based and speaker-based models have complementary strengths.

We observe that L_2 substantially outperforms its own base model L_0, showing that pragmatic reasoning on its own contributes positively. Blending the pragmatic models with the base listener also improves over both individually. Finally, the most effective listener combines both pragmatic models with the base listener. Plotting the number of examples changed by condition on the dev set (Figure 4) reveals that the primary gain from including the pragmatic models is in the *close* and *split* conditions, when the model has to distinguish highly similar colors and often cannot rely only on basic color terms. On the test set, the final ensemble improves significantly[10] over the base model on both metrics.

Examining the full probability tables for various dev set examples on which L_2 is superior to L_0 reveals a general pattern. In most of these examples, the alternative utterances sampled from S_0 for one of the referents i fail to identify their intended referent to L_0. The pragmatic listener interprets this to mean that referent i is inherently difficult to refer to, and it compensates by increasing referent i's probability. This is beneficial when i is the true target but harmful when i is a distractor.

Figure 5 shows one such example: a context consisting of a somewhat prototypical blue, a bright cyan, and a purple-tinged brown, with the utterance *blue*. The base listener interprets this as referring to the cyan with 91% probability, perhaps due to the extreme saturation of the cyan maximally activating certain parts of the neural network. However, when the pragmatic model takes samples from S_0 to probe the space of alternative utterances, it becomes apparent that indicating the more ordinary blue to the listener is difficult: for the utterances chosen by S_0 intending this referent (*true blue*, *light blue*), the listener also chooses the cyan with >89% confidence.

[10] $p < 0.001$, approximate permutation test (Padó, 2006), 10,000 samples

L_0	(target)		
blue	9%	**91**	<1
true blue	$\boxed{11}$	**89**	<1
light blue	$\boxed{<1}$	>99	<1
brightest	<1	$\boxed{\mathbf{>99}}$	<1
bright blue	<1	$\boxed{\mathbf{>99}}$	<1
red	<1	1	$\boxed{\mathbf{99}}$
purple	<1	2	$\boxed{\mathbf{98}}$
S_1	(target)		
blue	41	19	<1
true blue	$\boxed{\mathbf{47}}$	19	<1
light blue	$\boxed{5}$	**20**	<1
brightest	<1	$\boxed{\mathbf{20}}$	<1
bright blue	2	$\boxed{\mathbf{20}}$	<1
red	1	2	$\boxed{\mathbf{50}}$
purple	5	1	$\boxed{\mathbf{50}}$
L_2	(target)		
blue	**68**	32	<1

Figure 5: Conditional probability tables used to calculate L_2 for one dev set example. The true target color is boxed, and "blue" is the human utterance. Boxed cells for alternative utterances indicate the intended target; highest-probability referents (listeners) and utterances (speaker) are in **bold**. Sample sizes are reduced to save space; here, $m = 2$ and $n = 1$ (see Section 4.4).

Pragmatic reasoning overcomes this difficulty. Only two utterances in the alternative set (the actual utterance *blue* and the sampled alternative *true blue*) result in any appreciable probability mass on the true target, so the pragmatic listener's model of the speaker predicts that the speaker would usually choose one of these two utterances for the prototypical blue. However, if the target were the cyan, the speaker would have many good options. Therefore, the fact that the speaker chose *blue* is interpreted as evidence for the true target. This mirrors the back-and-forth reasoning behind the definition of conversational implicature (Grice, 1975).

6 Related work

Prior work combining machine learning with probabilistic pragmatic reasoning models has largely focused on the speaker side, i.e., generation. Golland et al. (2010) develop a pragmatic speaker model, $S(L_0)$, that reasons about log-linear listeners trained on human utterances containing spatial references in virtual-world environments. Tellex et al. (2014) apply a similar technique, under the name *inverse semantics*, to create a robot that can informatively ask humans for assistance in accomplishing tasks.

The closest work to ours that we are aware of is that of Andreas and Klein (2016), who also combine neural speaker and listener models in a reference game setting. They propose a pragmatic speaker, $S(L_0)$, sampling from a neural S_0 model to limit the search space and regularize the model toward human-like utterances. We show these techniques help in listener (understanding) tasks as well. Approaching pragmatics from the listener side requires either inverting the pragmatic reasoning (i.e., deriving a listener from a speaker), or adding another step of recursive reasoning, yielding a two-level derived pragmatic model $L(S(L_0))$. We show both approaches contribute to an effective listener.

7 Conclusion

In this paper, we present a newly-collected corpus of color descriptions from reference games, and we show that a pragmatic reasoning agent incorporating neural listener and speaker models interprets color descriptions in context better than the listener alone.

An important next step is to pursue multi-turn dialogue. As noted in Section 2, both participants in our reference game task could use the chat window at any point, and more than half of dyads had at least one two-way interaction. Dialogue agents are more challenging to model than isolated speakers and listeners, requiring long-term planning, remembering previous utterances, and (for the listener) deciding when to ask for clarification or commit to a referent (Lewis, 1979; Brown and Yule, 1983; Clark, 1996; Roberts, 1996). We release our dataset[11] with the expectation that others may find interest in these challenges as well.

[11] http://see-supplementary-material/

References

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *EMNLP*.

Leon Bergen, Roger Levy, and Noah D. Goodman. 2014. Pragmatic reasoning through semantic inference. Ms., MIT, UCSD, and Stanford.

Gillian Brown and George Yule. 1983. *Discourse analysis*. Cambridge University Press.

Sarah Brown-Schmidt and Michael K. Tanenhaus. 2008. Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. *Cognitive science*, 32(4):643–684.

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

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

Richard S. Cook, Paul Kay, and Terry Regier. 2005. The World Color Survey database. *Handbook of Categorization in Cognitive Science*, pages 223–241.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

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

Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *EMNLP*.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818 – 829.

Noah D. Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1):173–184.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics*, volume 3: Speech Acts, pages 43–58. Academic Press, New York.

Joy E. Hanna, Michael K. Tanenhaus, and John C. Trueswell. 2003. The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, 49(1):43–61.

Robert X. D. Hawkins. 2015. Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, 47(4):966–976.

Justine T. Kao, Jean Y. Wu, Leon Bergen, and Noah D. Goodman. 2014. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Kazuya Kawakami, Chris Dyer, Bryan Routledge, and Noah A. Smith. 2016. Character sequence models for colorful words. In *EMNLP*.

Diederik P. Kingma and Jimmy Lei Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ruud Koolen, Albert Gatt, Martijn Goudbeek, and Emiel Krahmer. 2011. Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, 43(13):3231–3250.

Robert M. Krauss and Sidney Weinheimer. 1964. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1(1-12):113–114.

David Lewis. 1979. Scorekeeping in a language game. *Journal of philosophical logic*, 8(1):339–359.

Will Monroe and Christopher Potts. 2015. Learning in the Rational Speech Acts model. In *Proceedings of the 20th Amsterdam Colloquium*.

Will Monroe, Noah D. Goodman, and Christopher Potts. 2016. Learning to generate compositional color descriptions. In *EMNLP*.

Randall Munroe. 2010. Color survey results. Online at http://blog.xkcd.com/2010/05/03/color-survey-results.

Sebastian Padó, 2006. *User's guide to `sigf`: Significance testing by approximate randomisation*. `http://www.nlpado.de/~sebastian/software/sigf.shtml`.

Maike Paetzel, David Nicolas Racca, and David DeVault. 2014. A multimodal corpus of rapid dialogue games. In *LREC*.

Craige Roberts. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, pages 91–136.

Seymour Rosenberg and Bertram D. Cohen. 1964. Speakers' and listeners' processes in a word communication task. *Science*, 145:1201–1203.

Gaurav Sharma, Wencheng Wu, and Edul N. Dalal. 2005. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1):21–30.

Nathaniel J. Smith, Noah D. Goodman, and Michael C. Frank. 2013. Learning and using language via recursive pragmatic reasoning about other agents. In *NIPS*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Michael K. Tanenhaus and Sarah Brown-Schmidt. 2008. Language processing in the natural world. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1493):1105–1122.

Stefanie Tellex, Ross A. Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. 2014. Asking for help using inverse semantics. In *Robotics: Science and Systems*.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL-HLT*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Adam Vogel, Christopher Potts, and Dan Jurafsky. 2013. Implicatures and nested beliefs in approximate Decentralized-POMDPs. In *ACL*.

Matthew D. Zeiler. 2012. ADADELTA: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.