Analyzing Speaker Strategy in Referential Communication

Brian McMahan
Rutgers University
brian.c.mcmahan@gmail.com

Matthew Stone
Rutgers University
mdstone@rutgers.edu

Abstract

We analyze a corpus of referential communication through the lens of quantitative models of speaker reasoning. Different models place different emphases on linguistic reasoning and collaborative reasoning. This leads models to make different assessments of the risks and rewards of using specific utterances in specific contexts. By fitting a latent variable model to the corpus, we can exhibit utterances that give systematic evidence of the diverse kinds of reasoning speakers employ, and build integrated models that recognize not only speaker reference but also speaker reasoning.

1 Introduction

Language users are able to work together to identify objects in the world (Clark and Wilkes-Gibbs, 1986, among others). This ability involves formulating creative utterances, assessing their meaning in context, and anticipating listeners' understanding and response (Dale and Reiter, 1995; Clark and Schaefer, 1989, among others). Despite long study, fundamental questions remain unanswered about how people manage this complex problem solving. This paper explores one question in particular: how speakers establish that references are likely to be successful. In general, such expectations can be underwritten either linguistically, by reasoning about the meanings and denotations of candidate linguistic expressions, or cooperatively, by reasoning about and anticipating their interlocutors' collaborative problem solving. Both kinds of reasoning are undoubtedly common, and both play a significant role in the psychological and computational literature on referential communication.

In this paper, we use quantitative cognitive models, fit to naturalistic corpora, to characterize the contributions of linguistic and cooperative reasoning in the spontaneous strategies of human interlocutors in referential communication. Our research offers a number of contributions for the SIGDIAL community.

- In Section 2, we provide a catalogue of phenomena and examples to distinguish linguistic reasoning and cooperative reasoning in reference. This analysis shows that linguistic reasoning and cooperative reasoning attribute different risks and rewards to utterances, and so explains why formalizations of linguistic reasoning, such as traditional plan-based approaches to generating referring expressions, and formalizations of cooperative reasoning, as often realized in machine learning approaches, can lead to different predictions about utterance choice.
- In Section 3, we refine approaches from the literature to capture the key phenomena we associate with different aspects of linguistic and cooperative reasoning. This modeling effort allows us to explore different inferences on an equal footing, using learned meanings with open-ended vocabulary and probabilistic, vague denotations.
- In Section 4, we evaluate the predictions of the models on human utterances in dialogue. By fitting a latent variable model to the corpus, we find strong evidence that while speakers often offer safe, conservative references, a sizeable fraction take risks that are only explained either by linguistic reasoning or by cooperative reasoning; these risky choices are broadly successful.

Our findings give new detail to the received understanding of collaborative problem solving in dialogue. Interlocutors often improvise, using risky strategies, in problematic situations; in these cases, they may have to work together interactively to achieve mutual understanding.

We believe that the researchers working on computational discourse and dialogue are uniquely positioned to take up these results to build more powerful models of the reasoning of human speakers, and to use analogous models in the choices of automated systems. At the same time, we argue that appreciating the diversity of dialogue is necessary to build interactive systems that understand and respond appropriately to their human users. Our work culminates in a mixture model that, given a description, predicts not only what the likely referent is, but also what reasoning the speaker was likely to have used to produce it.

2 Linguistic and Cooperative Reasoning

Our work is motivated by a distinction between reasoning linguistically, about meanings and denotations, and reasoning cooperatively, about understanding and collaboration. We begin by reviewing the theoretical and practical literature behind the two different approaches. To be clear, many reference problems have simple, good solutions that any reasoning will find. Differences arise in more complicated cases, when speakers need to exploit the flexibility of linguistic meaning or the ability of the listener to recognize implicatures, and when speakers need to trade off between specific and general referring expressions.

For clarity, our discussion illustrates these effects with concrete examples, even though this requires us to anticipate some results from later in the paper. In particular, we draw on attested examples from the Colors in Context (CIC) dataset of Monroe et al. (2017), where a director must signal one target in a display of three color patches. An example is shown in Figure 1.

We characterize the examples in terms of the quantitative predictions of models (described in full detail in Section 3), which formalize linguistic reasoning and cooperative reasoning. These models adopt a decision-theoretic approach. Utterances achieve various outcomes with different probabilities. For example, we may be uncertain whether an utterance will be judged appropriate to the context, whether it will be understood correctly—and so whether it will be successful in advancing referential communication. Safer utterances, with a higher probability of success, contrast with riskier utterances, with lower probability of success.



director:	the very dull red
matcher:	lighter or darker
director:	lighter

Figure 1: An example from the Colors in Context (CIC) dataset (Monroe et al., 2017) of the director and matcher coordinating so that the matcher can click on the correct color patch (x_0).

In tandem, each utterance has a cost (fixed across models), which determines the utility obtained when the utterance is successful. A rare utterance, like *chartreuse*, is modeled as having a higher cost than a more frequent utterance with the same meaning, like *yellow-green*. In fact, general terms, like *blue-gray*, which describe a comparatively large subset of color space, are typically assigned a lower cost than more specific terms, like *slate*, which describe a narrower subset. This is because, in situations where general terms and specific terms both offer equal prospects of task success, human speakers tend to prefer the general ones. This preference is particularly strong for basic-level terms (Rosch, 1978; Berlin, 1991), like *blue*.

Overall then, the models assign each utterance an EXPECTED UTILITY, which combines risk and cost in a single preference ranking. As is common in empirical models of human choice (Luce, 1959; McFadden, 1973), speakers are modeled as stochastic, approximate utility maximizers. The greater the utility advantage of the best choice, the more likely speakers are to use it; less advantageous choices are unlikely but not impossible. This assumption translates the model of expected utility into a distribution over potential descriptions (w) conditioned on the target color patch (x_0) and context of all three color patches (C).

2.1 Linguistic Reasoning

For linguistic models of referential communication, reference is a matter of meaning. The referent of a definite referring expression must be the unique entity from the contextually salient set of candidates that satisfies the expression's descriptive content. If there is no such unique entity, the referent is undefined.[1] Semantic reference is a proxy for successful communication. A speaker who establishes uniqueness can generally be confident that the listener will

[1] In formal semantics and pragmatics, this requirement is typically modeled as a grammatically-encoded presupposition, with the contextually salient set derived via the general process of quantifier domain restriction (Roberts, 2003).

Linguistic Reasoning		
$P(*\text{bright green} \mid x_0, C)$	$= 0.65$	x_0
$P(\text{neon green} \mid x_0, C)$	$= 0.16$	
$P(\text{green} \mid x_0, C)$	$= 0.12$	x_1
Cooperative Reasoning		x_2
$P(\text{green} \mid x_0, C)$	$= 0.33$	
$P(*\text{bright green} \mid x_0, C)$	$= 0.18$	C
$P(\text{lime green} \mid x_0, C)$	$= 0.11$	

Figure 2: Speakers can make their referring expressions more specific to come up with a description that's true of the target and false of the distractors. The observed description is marked with $*$.

Linguistic Reasoning		
$P(*\text{yellow} \mid x_0, C)$	$= 0.69$	x_0
$P(\text{mustard} \mid x_0, C)$	$= 0.06$	
$P(\text{greenish yellow} \mid x_0, C)$	$= 0.03$	x_1
Cooperative Reasoning		x_2
$P(*\text{yellow} \mid x_0, C)$	$= 0.30$	
$P(\text{yellow green} \mid x_0, C)$	$= 0.13$	C
$P(\text{lime green} \mid x_0, C)$	$= 0.06$	

Figure 3: Linguistic flexibility. Speakers can tailor a denotation for vague predicates that distinguishes their target from its distractors. The observed description is marked with $*$.

identify the same referent—without simulating the listener's perspective or interpretive reasoning.[2]

Planning-based approaches to generating referring expressions in the tradition of Dale and Reiter (1995) implement linguistic reasoning: the fundamental task is to come up with a description that characterizes to the target object but excludes its distractors. Such uniquely identifying descriptions are successful; alternative descriptions that fail to characterize the target or fail to exclude distractors are not. See van Deemter (2016) for a recent survey. A consequence of this model is to favor more specific vocabulary when it is necessary to avoid ambiguity, as demonstrated in Figure 2 where the linguistic reasoning model heavily favors the attested description *bright green* that a human speaker uttered when presented with the context. Although individual items offer only anecdotal evidence, when human speakers reliably choose to use semantically-identifying descriptions (*bright green*) with higher costs than alternatives that cooperative reasoning predicts to be successful (*green*), we find systematic evidence that speakers do use linguistic reasoning to identify targets.

The vagueness of color terms complicates the story. The natural way to extend linguistic reasoning to vague descriptions is to follow Kennedy (2007) in defining vague predicates in terms of a contextually-determined threshold of applicability. Vague predicates apply to those items that meet the threshold and exclude those that do not. On this theory, vagueness arises because, in any real context, a range of thresholds (of indeterminate extent) will typically be in play. There may be borderline cases that are neither clearly above all the thresholds in play nor clearly below them.

Speakers can exploit vagueness to communicate effectively (van Deemter, 2012). In particular, a speaker can implicitly choose to adopt further constraints on the threshold, leading to a more specific interpretation for the vague word. Once we take this possibility into account, a vague description refers uniquely as long as there are some (contextually-appropriate) thresholds where it identifies the target and none where it identifies a distractor (van Deemter, 2006; Meo et al., 2014). As an example, consider the attested utterance of *yellow* in Figure 3, where the target x_0 is a borderline case. Because x_0 is clearly a better yellow than the alternatives, there's a natural specific interpretation for *yellow* (with threshold ranging from the yellowness of x_1 to that of x_0) that uniquely identifies the target. In contrast, if we do not track the specialized interpretations that arise from a semantic requirement of uniqueness, we predict that the term might still apply to the distractor objects, and create potential disambiguation problems even for a cooperative listener.

In using a vague description, the speaker may be uncertain about whether its interpretation as uniquely identifying is appropriate for the context. If this interpretation is too specific, meaning that the word draws a contrast between similar and salient items on either side of its threshold, the listener may judge it to be infelicitous (Graff Fara, 2000). This is a matter of degree; in evaluating descriptions that require relatively unusual or precise interpretations to uniquely identify the target (e.g., *blue* in Figure 4 below), linguistic reasoning predicts that they will be less likely to be contextually appropriate and so less likely to be used.

[2]Of course, where the listener's knowledge of language or the world is unexpectedly incomplete, linguistic reasoning may result in an expression that characterizes the referent uniquely but in a way the listener may not recognize (Clark and Marshall, 1981).

In short, when human speakers reliably exploit the flexibility of vague meanings to produce low-cost, linguistically-identifying descriptions, in ways that look comparatively risky on purely cooperative reasoning, as in Figure 3, we find evidence for linguistic reasoning.

2.2 Cooperative Reasoning

Cooperatively, meanwhile, listeners approach interpretation with preferences and expectations that efficient speakers can and should meet and exploit (Schelling, 1960; Clark, 1996, among others). A description doesn't have to characterize the target uniquely—or even correctly—for the speaker to be confident that the listener will successfully retrieve the intended referent. Such cooperative effects are visible in the implicit strengthening of scalar implicatures (Horn, 1984; Frank and Goodman, 2012), where the listener naturally excludes a candidate interpretation that is technically possible but that the speaker could have been expected to signal differently. They are also visible in "loose talk" and exaggeration (Sperber and Wilson, 1986; Carston, 2002), where the description, while strictly speaking false, fits the target close enough to leave no doubt in the listener's mind. These "inaccurate references" can even include cases of outright falsehood (Perrault and Cohen, 1981), if there's a unique basis to link the false description with the intended target. Cooperative reasoning thus accommodates a diverse catalogue of non-unique descriptions that nevertheless succeed—what you might call, following Grice (1975), referential implicatures. A range of recent computational work has combined machine learning models of listener inference with probabilistic planning with the goal of generating such referential implicatures (Frank and Goodman, 2012; Monroe and Potts, 2015, among others).

Figure 4 shows an attested case, which we describe following Horn (1984). Interlocutors understand that *blue* can and will refer to the bright blue target in this context because it wouldn't be rational to try to use *blue* to refer to either of the dull blue alternatives. Quantitatively, the linguistic judgment that the target but not the alternatives is in fact blue represents a very specific and unlikely interpretation of *blue*. By contrast, the cooperative speaker sees *blue* as a likely choice, because of the low cost of the expression being used, on the one hand, and the good likelihood of being (cooperatively and correctly) understood, on the other.

Linguistic Reasoning		
$P(\text{bright blue} \mid x_0, C)$	$=$	0.34
$P(*\text{blue} \mid x_0, C)$	$=$	0.24
$P(\text{royal blue} \mid x_0, C)$	$=$	0.23
Cooperative Reasoning		
$P(*\text{blue} \mid x_0, C)$	$=$	0.67
$P(\text{bright blue} \mid x_0, C)$	$=$	0.07
$P(\text{royal blue} \mid x_0, C)$	$=$	0.07

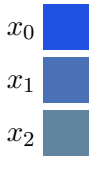


Figure 4: Referential implicature. A speaker who anticipates the listener's cooperative reasoning can use a potentially ambiguous description if the intended target is the most salient fit. The observed description is marked with $*$.

Linguistic Reasoning		
$P(\text{orange} \mid x_0, C)$	$=$	0.64
$P(*\text{red} \mid x_0, C)$	$=$	0.13
$P(\text{red orange} \mid x_0, C)$	$=$	0.08
Loose Talk		
$P(\text{orange} \mid x_0, C)$	$=$	0.12
$P(*\text{red} \mid x_0, C)$	$=$	0.12
$P(\text{peach} \mid x_0, C)$	$=$	0.10

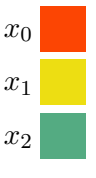


Figure 5: Loose talk. Even if this speaker judged the target x_0 to be orange, rather than red, she could be confident that her audience would resolve *red* to x_0. The observed description is marked with $*$.

In such cases, when speakers reliably move forward with general expressions backed up by referential implicatures while linguistic reasoning favors more specific expressions, as in Figure 4, we find evidence for cooperative reasoning.

The description of Figure 4 is true of the target. What of inaccurate but comprehensible references? We show one possible attested case in Figure 5. Linguistic reasoning predicts the target should be described as *orange* rather than *red*. However, *red*, though a stretch, is unambiguous.

Unfortunately, we cannot be sure that the speaker intended the description *red* to be false but recognizable. An alternative explanation is that the speaker did categorize the patch as *red* (in a weird and idiosyncratic way). Our current data and methods cannot rule out such individual differences. In any case, our analysis suggests such examples are comparatively rare in this dataset, so our key models are designed to avoid loose talk.

In summary, prior linguistic research and prior computational models appeal to heterogeneous kinds of reasoning to explain how speakers plan

referential expressions. These models make incompatible predictions, particularly about how to handle vagueness and implicature, which are visible in their predicted trade-offs between specific and general referring expressions. How do these differences actually play out in natural dialogue? What evidence is there for linguistic reasoning and cooperative reasoning in the utterances of human speakers? And what effects might utterances with different origins have on the dynamics of interaction? The increasing availability of corpus data and the increasing power of machine learning methods makes it possible to adopt a quantitative approach to answering such questions. The remainder of this paper offers an initial experiment in this direction.

3 Learning Speaker Reasoning

We formulate computational models of speaker reasoning in two steps. First, as described in Section 3.1, we build the XKCD model based on the applicability and cost of color terms, following McMahan and Stone (2015); Monroe et al. (2016); McDowell and Goodman (2019). Second, we describe the linguistic and cooperative reasoning choices of speakers as a function of these learned parameters. As described in Section 3.2, our models of linguistic reasoning use probabilistic models of vagueness to formulate low-cost descriptions that denote the target uniquely (van Deemter, 2006; Meo et al., 2014). Meanwhile, as described in Section 3.3, our models of cooperative reasoning use probabilistic planning to find low-cost utterances likely to be understood by the listener, following the Rational Speech Acts approach (Frank and Goodman, 2012).

3.1 The XKCD Model

The linguistic and cooperative reasoning models depend on a shared model of meaning and cost which we name the XKCD model. We fit the XKCD model using a corpus of color patch descriptions that were freely labeled by volunteer crowd workers then cleaned in previous work (McMahan and Stone, 2015) resulting in 1.5M training, 108K development, and 544K testing examples.

Our assumption, in line to previous work (McMahan and Stone, 2015; McDowell and Goodman, 2019), is that speaker choices in this dataset can be attributed to two factors. The first is the APPLICABILITY of the description w_k to color patch x_i, denoted $\phi_{w_k}(x_i)$, which is a probabilistic measure of the degree to which a color description naturally fits a color patch. Applicability serves as a shared model of meaning for the models (which the models enrich pragmatically in different ways). The second factor is the AVAILABILITY, denoted α_{w_k}, which is a measure of the intrinsic frequency of a color description w_k. Availability inverts the intuitive notion of cost; descriptions with lower cost have higher availability (and higher utility).

We treat the XKCD model as a "literal speaker" in the specific sense that no referential implicatures factor in ϕ_{w_k}, since the speaker is not presented with alternative referential candidates and does not have the goal of identifying an intended target. As defined in Equation 1, the probability that the literal XKCD speaker uses the description w_k to describe patch x_i in context C is proportional both to w_k's applicability to x_i and to w_k's availability, and doesn't depend on context.

$$S_0(w_k|x_i, C) = \frac{\phi_{w_k}(x_i)\alpha_{w_k}}{\sum_l \phi_{w_l}(x_i)\alpha_{w_l}} \quad (1)$$

In addition, we define a "literal listener" L_0 that leverages the applicability functions of the XKCD model in Equation 2:

$$L_0(x_i|w_k, C) = \frac{\phi_{w_k}(x_i)}{\sum_j \phi_{w_k}(x_j)} \quad (2)$$

L_0 quantifies the preference for interpretation x_i based on how appropriate the description w_k is for color patch x_i relative the other color patches.[3]

We implement the model as a neural network using the PyTorch deep learning framework (Paszke et al., 2019).[4] Neural networks learn data-driven representations of color space and color categories, which leads to more flexible and accurate meanings (Monroe et al., 2016) compared to models that use handcrafted parameterizations for color space and color meanings as in McMahan and Stone (2015).

Starting from a Fourier feature representation of color patches (Monroe et al., 2016), we use a 3-layer Multilayer Perceptron (MLP) with layers of size 32 and ELU intermediate activation functions to map the features of a color patch x to an intermediate scalar value, $\hat{x}$. Next, we use a sigmoid function on $\hat{x}$ to compute the applicability. The

[3] McMahan and Stone (2015) argue that S_0 and L_0 so defined represent an equilibrium, where naive interlocutors and strategic interlocutors converge on their interpretations.

[4] All code and data is available at https://go.rutgers.edu/ugycm1b0.

series of computations from an HSV color patch to applicability are shown in Equation 3.

$$\phi_{w_k}(x_i) = \sigma(\text{MLP}(\text{FFT}(x_i^{\text{HSV}})) \quad (3)$$

We implement availability as a vector that is transformed to probability values using the sigmoid function and fit during the training routine.

The model is fit in a two-stage approach. The first stage uses a conditioned language modeling objective: minimize the negative log likelihood of $S_0(w_k|x_i, C)$ in Equation 1. In the second stage, we define a CALIBRATION technique so that the rates of applicability for a description do not encode its frequency in the training dataset. The technique, inspired by work on knowledge distillation (Hinton et al., 2015), forces the applicabilities for each description be close to 1 for at least one training data point. Calibration begins by using the model trained in the first stage to compute applicability values for every training data point. Each description's vector of applicability values is normalized by their 99th percentile value and bounded in the 0-1 range.

The final step of the calibration technique trains a second model to minimize both the original language modeling objective and a binary cross entropy between the second model's applicability predictions and the first model's normalized applicabilities. Both models are trained using the RAdam optimization algorithm (Liu et al., 2019) with a learning rate of 0.01 and a learning rate annealing which decreases the learning rate by 75% if the perplexity of the validation set does not improve for 2 epochs. Training is terminated if the validation perplexity does not improve for 4 epochs.

3.2 The Linguistic Reasoning Model (RGC)

Our linguistic reasoning model extends the XKCD model to enable vague predicates that distinguish a target from its competing alternatives. Recall that, in the XKCD model, the applicability calculation for each description w_k concludes with a sigmoid operation. We conceptualize this as the cumulative distribution function over a random variable τ_{w_k} representing a contextual threshold: the probability w_k applies to color patch x is the probability that x exceeds the contextual threshold τ_{w_k}. Following Meo et al. (2014), a description w_k can then distinguish between the target x_0 and its competing alternatives x_1 and x_2 by committing to the thresholds that distinguish them. The goal of referring to x_0 and not x_1 or x_2 with w_k requires corresponding comparisons to bound the cumulative distribution τ_{w_k}, shown in Equation 4 and simplified in Equation 5.

$$P(\max(x_1, x_2) < \tau_{w_k} < x_0) \quad (4)$$

$$\phi_{w_k}(x_0) - \max(\phi_{w_k}(x_1), \phi_{w_k}(x_2)) \quad (5)$$

$$:= \psi_{w_k}(x_0, \neg x_1, \neg x_2) \quad (6)$$

To compute the linguistic speaker's probability distribution over descriptions, we utilize ψ_{w_k} in Equation 6 to replace ϕ_{w_k} in Equations 1 and 2. We refer to this model as REFERENTIAL GOAL COMPOSITION (RGC), reflecting the fact that it decomposes the goal of identifying the target to sub-goals of describing the target and excluding the alternatives.

3.3 The Cooperative Reasoning Model (RSA)

Our cooperative reasoning model extends the XKCD model by adapting the Rational Speech Acts (RSA) model of Monroe and Potts (2015). The basic idea is that the strategic speaker S_1^{RSA} chooses a description w_k for x_i in proportion to the probability that the literal listener, when presented with w_k, will recover the intended referent x_i.

$$S_1^{RSA}(w_k|x_i, C) = \frac{L_0(x_i|w_k, C)\alpha_{w_k}}{\sum_l L_0(x_i|w_l, C)\alpha_{w_l}} \quad (7)$$

Although RSA generates and interprets scalar implicatures, which assume that the listener will take salience into account in resolving reference, nothing in Equation 7 privileges descriptions that are more naturally applicable to the target referent. The literal listener L_0's interpretations can easily stray from literal meaning—recovering the target object from utterances that fit the target poorly but fit alternative objects worse, as in the case of loose talk considered in Section 2. To model the data of Monroe et al. (2017), it's important to stay closer to literal meaning and penalize utterances that are poor fits for the target object.

We do this by modifying the RSA formulation so the listener entertains the possibility that they are unfamiliar with or cannot identify the speaker's intended referent. When the listener adopts this *out-of-context* interpretation (as they will if the speaker's description is sufficiently unlikely to fit the target), the speaker has not communicated successfully. This gives a pragmatic speaker a reason not to rely on loose talk.

More formally, we define the *out-of-context* interpretation which the listener assigns a probability

$\psi_{w_k}(\neg x_0, \neg x_1, \neg x_2)$ that the description does not apply to any of the potential targets. This leads to a revised listener $L_{0+}(x_i|w_k, C)$ defined as in Equation 8:

$$\frac{\phi_{w_k}(x_i)}{\psi_{w_k}(\neg x_0, \neg x_1, \neg x_2) + \sum_j \phi_{w_k}(x_j)} \quad (8)$$

and a correspondingly revised speaker S_{1+}^{RSA}.

3.4 A Conservative Baseline Model (CB)

In addition to the linguistic reasoning model (RGC) and cooperative reasoning model (RSA), we evaluate a conservative baseline which prioritizes simple, unambiguous referring expressions. When speakers use such expressions, they don't show any evidence of relying on linguistic flexibility or referential implicatures. In fact, key recent results in modeling referential communication use models that exclusively use conservative referring expressions (McDowell and Goodman, 2019).

A conservative speaker uses a description w_k to identify x_i in context C by striking a balance between the literal listener and literal speaker:

$$S_1^{CB(\lambda)}(w_k|x_i, C) \propto L_0(x_i|w_k, C)^{\lambda} S_0(w_k|x_i, C) \quad (9)$$

The "rationality parameter" exponent λ is typically set to a value substantially greater than 1, which gives the model slim confidence that the listener will do cooperative reasoning to disambiguate. Consequently, w_k will be heavily penalized unless the literal meaning clearly indicates that the distractors do not fit the description. Using the XKCD model, $S_0(w_k|x_i, C)$ simplifies to $\phi_{w_k}(x_i)\alpha_{w_k}$. The difference with S_{1+}^{RSA} in Equation 7 is the additional factor $\phi_{w_k}(x_i)$, which says that w_k should be true of the target, and so penalizes both loose talk and linguistic flexibility.[5]

4 Experiments

Having presented mathematical abstractions that identify linguistic reasoning, cooperative reasoning, and the conservative baseline, we now evaluate how well they fit natural utterances in interactive referential communication. We approach this question in two ways. Section 4.2 takes the naive approach of measuring how well each approach explains speaker choices on its own. Ultimately, however, we believe this is somewhat misleading. It's more instructive, we argue, to hypothesize that speakers can use different strategies in different situations. Section 4.3 uses a mixture model to provide evidence that the different models fit different aspects of speakers' language use.

4.1 The Colors in Context Dataset

The data we use to evaluate our models of speaker reasoning comes from Monroe et al. (2017), who asked participants to talk about items in a visual display using a free-form chat interface. On each round of interaction, one human subject, designated the director, was provided privately with a target item from a randomized display of three colors and tasked with instructing the other human subject, designated the matcher, to click on the correct item. The displays varied the relationship between the target and the distractors: in the FAR condition, all three colors were visually dissimilar; in the SPLIT condition, the target had a single visually similar distractor; and in the CLOSE condition, all three colors were visually similar. Overall, 775 subjects participated in 948 games with 50 rounds per game for a total of 47,041 rounds. As shown in Figure 1, some rounds have multiple utterances, resulting in 57,946 utterances in total. To eliminate any confounds of processing complex utterances, our experiments focus on a 23,801 utterance subset created by selecting rounds where the director made a single utterance before the matcher clicked a target and where the director's utterance matched an item from the XKCD lexicon.[6]

4.2 Analyzing Strategies Independently

Our first analysis measures how well each model predicts speaker choices in the filtered dataset. To start, we gathered predictions from the three strategies for every data point. RGC aggressively rules out descriptions which have a higher applicability for one of the alternate objects, resulting in 0 probabilities for 530 examples (6.8%) in the training data. To handle the 0 probabilities, we use Jelinek-Mercer smoothing (Jelinek, 1980) for each strategy's predictions, which uses a tuned hyper parameter to interpolate between the strategy's predictions and the relative frequencies of descriptions

[5] This factor was originally proposed by Andreas and Klein (2016) in the context of adding pragmatic reasoning to systems whose fundamental computational operation was sampling true descriptions.

[6] A regular expression approach was used to allow for descriptions like "the blue square" or "the red one".

	Dataset Split		
Model	train	val/dev	test
S_0	15.50	14.88	13.28
RGC	16.15	15.03	13.32
S_{1+}^{RSA}	14.62	14.05	12.49
$S_1^{CB(2)}$	14.14	13.50	11.84
$S_1^{CB(15)}$	20.76	18.83	16.36
S^{EM}	13.47	12.75	11.30

Table 1: Perplexity scores on the CIC dataset for the linguistic (RGC), cooperative (RSA), conservative (CB), and mixture (EM) models. $S^{CB(2)}$ has the lowest independent perplexity and $S^{CB(15)}$ is selected in the mixture analysis. A Wilcoxon Signed-Ranks Test indicated all differences are significant ($p < 10^{-4}$).

in the colors-in-isolation dataset. This is the only parameter here that's estimated from the Monroe et al. (2017) training set. We show the perplexity of each model using the interpolated probabilities in Table 1. Overall, cooperative reasoning S_{1+}^{RSA} and the conservative baseline with a small rationality parameter $S_1^{CB(2)}$ better predict what people say on average.

4.3 Analyzing Strategies as a Mixture

Ultimately, the different models all represent plausible reasoning for speakers. There is no reason to think all speakers are the same. We therefore use a mixture analysis (also known as a latent variable analysis) to understand the predictions for individual items (Zeigenfuse and Lee, 2010; Lee, 2018). Overall, the optimization goal is to maximize the likelihood of the data under a posterior distribution where an observed utterance w_i for color patch x_i is generated by a mixture of each model M_j:

$$P(w_i|x_i, C) = \sum_j P(w_i|x_i, C, M_j)P(M_j)$$

The posterior distribution is maximized using an Expectation-Maximization (EM) routine that iteratively computes the probability of a model conditioned on each data point (the "expectation" step) and the prior probabilities for each model (the "maximization" step) (Bishop, 2006, p. 430). The probabilities for observed items $P(w_i|x_i, C, M_j)$ are the non-smoothed probabilities from the independent model analysis in Section 4.2 and were not updated in the EM routine. We repeat the procedure until convergence.[7] The result is a set of inferred prior probabilities for the models as well as overall perplexity for the dataset. The prior probabilities are computed for the training set only and used to evaluate the perplexity on the development and test portions of the dataset.

When generating referential expressions, speakers could be using linguistic or cooperative reasoning, they could be acting more conservatively, or they could even be behaving randomly. We structure the mixture analysis to evaluate these options by pitting the RGC model, the RSA model, a CB model, and two random baselines against each other. For the CB model, we set $\lambda = 15$ by evaluating the mixture analyses for the range $1 \leq \lambda \leq 26$ and selecting the λ that results in lowest perplexity on the development set. For the random baselines, we use both a uniform distribution and the normalized frequency distribution of the XKCD corpus.

Because model predictions typically overlap, EM mixture weights are highly sensitive to outlier predictions where models give low probabilities. Nevertheless, the inferred prior probabilities in Table 2 provide evidence that a heterogeneous mixture of speaker choices do exist in the dataset. We can see this clearly in particular utterances. For example, the EM analysis allowed us to find the divergent cases presented in Figures 2–5.

To better understand what each model explains and how the dialogue evolves, we partition the dataset by both difficulty condition as manipulated by Monroe et al. (2017) and by which model best predicted the speaker's utterance. For each partition, Table 3 reports the number of cases and the matcher success rate. Additionally, we further break out the cases where the RGC model gave 0 probability to the speaker's utterance. To test for significance, we use the Mann-Whitney U Test for matcher success and the Wilcoxon Signed-Rank

[7] Since models are not updated during EM, we define convergence to be when the sum over absolute differences in priors was less than 10^{-6}.

Model M_j	$P(M_j)$	Model M_j	$P(M_j)$
RGC	0.33	$freq^{\text{XKCD}}$	0.006
S_{1+}^{RSA}	0.46	Uniform	0.004
$S_1^{CB(15)}$	0.19		

Table 2: The EM-fit prior probabilities for linguistic reasoning (RGC), cooperative reasoning (RSA), conservative baseline (CB), and two random baselines, normalized XKCD frequencies ($freq^{\text{XKCD}}$) and a uniform distribution (Uniform). We show the perplexities using these priors as S^{EM} in Table 1.

Winning Model	Matcher Success By Condition		
	FAR (4055)	SPLIT (2657)	CLOSE (1889)
RGC (2437)	98.15% (971)	92.88% (839)	87.44% (637)
S_{1+}^{RSA} (582)	88.73% (222)	79.20% (226)	75.54% (134)
$S_1^{CB(15)}$ (5064)	99.46% (2803)	98.14% (1396)	97.46% (865)
RGC = 0 (518)	44.07% (59)	56.80% (206)	58.50% (253)

Table 3: Matcher success rates in the test data by difficulty condition and best-explaining model. Counts are shown in parenthesis. The cases where RGC gave 0 probability to the utterance are counted separately (RSA is the overwhelming winner for these cases).

Test for utterance probabilities.[8]

Although RSA has the largest mixture weight, it actually doesn't score the speaker's utterance as highly as the other models most of the time (in all conditions, $p < 10^{-4}$), which suggests that cooperative reasoning predicts a wider range of descriptions (each with lower probability). By contrast, CB has a lower mixture weight, but scores higher on more data points than RSA in all conditions ($p < 10^{-4}$), RGC in the FAR condition ($p < 10^{-4}$), and RGC in the SPLIT condition ($p < 10^{-2}$); CB puts strong weight on a subset of likely descriptions that covers most, but not all cases. Indeed, CB seems to choosing precise, unambiguous descriptions, while the matcher success rates for linguistic and cooperative reasoning are lower ($p < 10^{-3}$), suggesting that these models do embody risky choices. Linguistic reasoning, as embodied by RGC, seems to be somewhat more successful than cooperative reasoning, as embodied by RSA ($p < 10^{-2}$). Finally, cases where RGC was not able to give a probability to the speaker utterance have far lower matcher success rates ($p < 10^{-2}$); it seems in these cases the matcher was genuinely confused.

5 Discussion and Conclusion

This paper has argued that human speakers in collaborative reference dialogues take diverse strategies: they can stick with clear, precise descriptions; alternatively, they can create innovative interpretations for words; alternatively, they can count on their audience to fill in the gaps in what they say. While computational models often focus on one specific kind of reasoning, we believe that our findings are broadly consonant with the psycholinguistics literature, with its evidence of the psychological difficulty of semantically identifying targets (Clark and Wilkes-Gibbs, 1986), its evidence of the psychological difficulty of taking the audience's perspective into account (Keysar et al., 2000), and its concepts of "least collaborative effort" (Clark and Schaefer, 1989) in characterizing interaction as fundamental to success in conversation. We are optimistic that future work can continue to develop precise data-driven models that integrate these different explanations to understand and respond to user utterances in dialogue systems.

Our work has a number of limitations that we leave for future research. Even within the simple domain of identifying color patches, we see the utterances that RGC cannot explain—utterances where a speaker seems to refer to a target object with a description that fits the target less well than a distractor—as a strong indication of variability in meaning across individuals. This needs to be accounted for. In addition, it would be good to explore models of reference to colors in context that generalize from colors in isolation data using more flexible machine-learned models of choice.

What about more complex domains and interactions? The challenges of providing fine-grained and wide-ranging analyses of interlocutors' referential problem-solving strategies remain substantial. Neverthless, we do see promising directions. One is to follow Elsner et al. (2018) in conceptualizing reference production in terms of a high-level choice of strategy followed by detailed content choices, and build a corresponding probabilistic model of reference production. Another is cover more complex interactions, by including additional interactive strategies for framing alternatives, excluding wrong interpretations, asking clarification questions, and answering them.

Acknowledgments

This research was supported by NSF awards IIS-1526723 and CCF-1934924. Preliminary versions were presented at NYU and Bochum. Thanks to audiences there, the anonymous reviewers, Malihe Alikhani, and Baber Khalid for comments.

[8] To accommodate multiple comparisons, we adjust the reported significance levels using Bonferonni correction.

References

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182.

Brent Berlin. 1991. *Basic Color Terms: Their Universality and Evolution*. Univ of California Press.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

Robyn Carston. 2002. *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Blackwell.

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

Herbert H. Clark and Catherine R. Marshall. 1981. Definite reference and mutual knowledge. In Arivind Joshi, Bonnie Webber, and Ivan Sag, editors, *Elements of Discourse Understanding*, pages 10–63. Cambridge University Press.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

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

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 18:233–263.

Kees van Deemter. 2006. Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222.

Kees van Deemter. 2012. *Not exactly: In praise of vagueness*. Oxford University Press.

Kees van Deemter. 2016. *Computational Models of Referring: A Study in Cognitive Science*. MIT Press.

Micha Elsner, Alasdair Clarke, and Hannah Rohde. 2018. Visual complexity and its effects on referring expression generation. *Cognitive science*, 42:940–973.

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

Delia Graff Fara. 2000. Shifting sands: An interest-relative theory of vagueness. *Philosophical Topics*, 28(1):45–81.

H. P. Grice. 1975. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics III: Speech Acts*, pages 41–58. Academic Press.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Laurence R. Horn. 1984. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In Deborah Schiffrin, editor, *Meaning, Form, and Use in Context: Linguistic Applications*, pages 11–42. Georgetown University Press.

Frederick Jelinek. 1980. Interpolated estimation of markov source parameters from sparse data. In *Proc. Workshop on Pattern Recognition in Practice, 1980*.

Christopher Kennedy. 2007. Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30(1):1–45.

Boaz Keysar, Dale J. Barr, Jennifer A. Balin, and Jason S. Brauner. 2000. Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11(1):32–38.

Michael D Lee. 2018. Bayesian methods in cognitive modeling. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, 5:1–48.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019. On the variance of the adaptive learning rate and beyond.

R. Duncan Luce. 1959. *Individual Choice Behavior: A Theoretical analysis*. Wiley.

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

D. McFadden. 1973. Conditional logit analysis of qualitative choice behaviour. In P. Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press.

Brian McMahan and Matthew Stone. 2015. A Bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*, 3:103–115.

Timothy Meo, Brian McMahan, and Matthew Stone. 2014. Generating and resolving vague color references. In *SEMDIAL 2014: The 18th Workshop on the Semantics and Pragmatics of Dialogue*, pages 107–115.

Will Monroe, Noah D Goodman, and Christopher Potts. 2016. Learning to generate compositional color descriptions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2243–2248.

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.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

C. Raymond Perrault and Philip R. Cohen. 1981. It's for your own good: a note on inaccurate reference. In Aravind K. Joshi, Bonnie Lynn Webber, and Ivan Sag, editors, *Elements of Discourse Understanding*, pages 217–230. Cambridge University Press.

Craige Roberts. 2003. Uniqueness in definite noun phrases. *Linguistics and Philosophy*, 26(3):287–350.

Eleanor Rosch. 1978. Principles of categorization. In Eleanor Rosch and Barbara B. Lloyd, editors, *Cognition and Categorization*, pages 27–48. Erlbaum.

Thomas C. Schelling. 1960. *The Strategy of Conflict*. Harvard University Press.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and Congition*. Harvard University Press.

Matthew D Zeigenfuse and Michael D Lee. 2010. A general latent assignment approach for modeling psychological contaminants. *Journal of Mathematical Psychology*, 54(4):352–362.