

Why do you ask? Good questions provoke informative answers.

Robert X.D. Hawkins, Noah D. Goodman
Stanford University

Author Note

This research took place while the first author was a student at Indiana University. He is now at Stanford University, Stanford, CA 94305, USA. This work was supported by National Institute of Mental Health Grant 05771710 awarded to J. T. T. We thank Devin Burns for helpful comments on the experiment reported, and Patricia Knapp for her assistance in data collection. Correspondence concerning this article should be addressed to Robert X.D. Hawkins, e-mail: rxdh@stanford.edu

Abstract

What makes a question useful? What makes an answer appropriate? In this paper, we formulate a family of increasingly sophisticated models of question-answer behavior within the Rational Speech Act framework. We compare these models based on three different pieces of evidence: first, we demonstrate how our answerer models capture a battery of four classic effects in psycholinguistics, which show that an answerer’s level of informativeness varies with the inferred questioner goal. Second, we jointly test the questioner and answerer components of our model based on empirical evidence from a simple question-answer reasoning game. Third, we designed a real-time, multi-player version of this game with a wider range of conditions, which allows us to distinguish among the questioner models. We find that sophisticated pragmatic reasoning is needed to account for some critical aspects of the data. People can use questions to provide cues to the answerer about their interest, and can select answers that are informative about inferred interests.

Keywords: pragmatics, computational modeling

Why do you ask? Good questions provoke informative answers.

Introduction

Q: “Are you gonna eat that apple?”

A: “Oh, go ahead!”

In this exchange, Q wants to know whether she can eat A’s apple. Instead of directly asking this question, Q strategically chooses a question that differs from her true interest, avoiding an impolite question, yet still manages to send a signal to A about her true intentions. A, in turn, reasons beyond the overt question and provides an answer that addresses Q’s true interests. This subtle interplay highlights the prevalence of social reasoning tasks in everyday dialogue and raises two specific questions for formal models of language: What makes a question useful? And what makes an answer appropriate?

A number of psycholinguistic studies have provided evidence that answerers are both sensitive to a questioner’s goals and attempt to be informative with respect to those goals. For instance, in the classic study of Clark (1979), researchers called liquor merchants and opened the conversation with one of two sentences to set context: “I want to buy some bourbon” (the *uninformative* condition) or “I’ve got \$5 to spend” (the *five dollar* condition). They then asked, “Does a fifth of Jim Beam cost more than \$5?” Merchants gave a literal yes/no answer significantly more often in the latter condition than the former, where an exact price was more common. When provided with the five dollar context, the merchant inferred that the questioner’s goal was literally to find out whether or not they could afford the whiskey, hence a simple ‘yes’ sufficed. In the uninformative context, however, the merchant inferred that the questioner’s goal was just to buy whiskey, so the exact price was the most relevant response (Clark, 1979).

Context and questioner goals have also been implicated in accounts of answers to identification questions like “who is X?” (Boër & Lycan, 1975), questions like “Do you have the time?” that permit answers with different degrees of approximation to the true value (Der Henst, Carles, & Sperber, 2002; Gibbs Jr & Bryant, 2008), and to questions like “where are you?” that permit answers at many levels of abstraction (Potts, 2012). While most of this work has focused on *answerer* behavior, it suggests that the question itself is

important in prompting a relevant answer.

Recent work on Rational Speech Act (RSA) models (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013) has mathematically formalized pragmatic language understanding as a form of recursive Bayesian inference, where listeners reason about speakers who choose utterances that maximize information gained by an imagined listener. In this paper we extend the RSA framework to address simple question-answer dialogs. The immediate challenge in doing so is that the speaker utility in RSA is based on direct information provided by an utterance—since questions don’t provide direct information, we must say what utility they do have.

We suggest, following Van Rooy (2003), that the value of a question is the extent to which it can be expected to elicit information relevant to the questioner later in the dialogue. More specifically, for the questioner, the value of a question is the expected information gained about her interests, given the set of likely answers it may provoke. This diverges from regular RSA in that the value of a question depends on information gained by the speaker (rather than listener), and that this information comes later in the (very short) conversation.

To fully specify this questioner we need a model of the answerer, which can serve as both the model assumed by a questioner, and as a model of answer behavior itself. We explore three, increasingly sophisticated, answerer models. The simplest answerer provides a literal answer to the question (without attempting to be informative); the explicit answerer attempts to be informative with respect to the explicit question asked (without inferring the questioner’s underlying interests); the pragmatic answerer infers the most likely true interests of the questioner, and then informatively addresses those interests. The latter model extends RSA to reason about the topic of conversation, as proposed by Kao, Wu, Bergen, and Goodman (2014) to explain hyperbole; it goes beyond previous work by using the explicit question as a (potentially indirect) cue to this topic.

The rest of this paper is structured as follows. First, we lay out the details of our computational questioner and answer models. We then present a set of computational experiments demonstrating how this set of models captures four classic answerer-sensitivity effects from the psycholinguistics literature. Because data on *questioner* behavior is

relatively sparse, we collected data in a simple communication game paradigm allowing us to manipulate private goals, potential questions, and potential answers. After testing the predictions of the different models on data from a simple single-player version of the task, we scale up this paradigm to a more natural real-time, multi-player experiment using a wider variety of goal sets, question sets, and answer sets. We dedicate particular attention to one condition in which the questioners models make different predictions. We close with a brief discussion of related models and future directions.

A Rational Speech Act model of question and answer behavior

How should a questioner choose between questions? We start by assuming that the questioner aims to *learn information relevant to a private goal*. In order to choose a question that results in useful information, the questioner reasons about how the answerer would respond, given different possible states of the world; she selects a question that results in an answer that tends to provide goal-relevant information.

More formally, suppose there is a set of world states \mathcal{W} , a set of possible goals \mathcal{G} , a set of possible questions \mathcal{Q} , and a set of possible answers \mathcal{A} . These sets are taken to be in common ground between the questioner and the answerer. An informational goal $g \in \mathcal{G}$ is a projection function that maps a world state to a particular feature or set of features that the questioner cares about; this is similar to the notion of a question-under-discussion (Roberts, 1996). We will use the notation $P_g(w)$ to indicate the probability $\hat{P}(g(w))$ of the g -relevant aspect of w under the projected distribution $\hat{P}(v) = \int_{\mathcal{W}} \delta_{v=g(w)} P(w) dw$.

The **questioner** takes a goal $g \in \mathcal{G}$ as input and returns a distribution over questions $q \in \mathcal{Q}$:

$$P(q|g) \propto e^{\mathbb{E}_{P(w^*)}[D_{KL}(P_g(w|q, w^*) \| P_g(w))] - C(q)}$$

It trades off the cost of asking a question, $C(q)$, and expected information gain. The cost likely depends on question length, among other factors. Information gain is measured as the Kullback-Leibler divergence between the prior distribution over g -relevant worlds, $P_g(w)$, and the posterior distribution one would expect after asking a question q whose

answer reflected true world state w^* :

$$P_g(w|q, w^*) = \sum_{a \in \mathcal{A}} P_g(w|q, a) P(a|q, w^*)$$

This distribution has two components: First, it depends on $P(a|q, w^*)$, a model of the answerer which we will explore shortly. Second, it depends on (the goal projection of) $P(w|q, a)$, an ‘interpreter’ that specifies the likelihood assigned to different worlds given question and answer pairs.

To define the interpreter function, which all agents use to compute the literal interpretation of a question-answer pair, we must assign questions a semantic meaning. We assume that a question is an informational goal that projects from worlds to the answer set \mathcal{A} . This is equivalent to the more common partition semantics of Groenendijk and Stokhof (1984), as can be seen by considering the pre-image of such a projection; an answer picks out an element of the partition via $q^{-1}(a)$. The **interpreter** constrains the prior on worlds to the subset of its support that is consistent with the semantics of a question-answer pair¹:

$$P(w|q, a) \propto P(w) \delta_{q(w)=a}$$

We next describe three different answerer models; the questioner could assume any one of them, leading to three corresponding versions of the questioner model. All answerers take a question $q \in \mathcal{Q}$ and a true world state $w^* \in \mathcal{W}$ as input and return a distribution over answers $a \in \mathcal{A}$. The **literal answerer** simply chooses answers by trading off prior answer probability and how well a question-answer pair conveys the true state of the world to an interpreter:

$$P(a|q, w^*) \propto P(a) P(w^*|q, a)$$

For a fixed question, this is equivalent to the speaker in previous RSA models. The question enters only in specifying the literal meaning of an answer. The **explicit answerer** additionally evaluates answers with respect to how well they address the explicit question q :

$$P(a|q, w^*) \propto P(a) P_q(w^*|q, a)$$

¹We should also have a semantic evaluation function that maps an answer utterance to its value in \mathcal{A} . For clarity we assume this is a trivial mapping and suppress it.

The **pragmatic answerer** also evaluates answers with respect to how well they address the informational goal, but doesn't take the question's explicit meaning at face value. Instead, the pragmatic answerer reasons about which goals g are likely given that a question q was asked, and chooses answers that are good on average:

$$P(a|q, w^*) \propto p(a) \sum_{g \in \mathcal{G}} P(g|q) P_g(w^*|q, a)$$

Reasoning backwards from questions to goals is a simple Bayesian inversion of the (explicit) questioner using a prior on goals:

$$P(g|q) \propto P(q|g)P(g)$$

For all of the questioner and answerer models, we can vary how strongly optimizing they are—that is, to what extent they are sampling from the distributions defined above, and to what extent they deterministically choose the most likely element. For any such distribution over utterances, we introduce an optimality parameter α and transform it by $P'(x) \propto P(x)^\alpha$.

This concludes our specification of the model space, giving a set of three answerers and three corresponding questioners that reason about them. We have implemented these models in WebPPL, a probabilistic programming language (Goodman & Stuhlmüller, electronic), and runnable code for all reported simulations will be available online at [http:// hawkrobe.github.io/Q_and_A/](http://hawkrobe.github.io/Q_and_A/). The model predictions shown throughout the rest of the paper are computed using this implementation.

Related models

The above probabilistic model of question and answer behavior bears some resemblance to recent decision theoretic (Vogel, Bodoia, Potts, & Jurafsky, 2013) and game theoretic (Franke, 2013) models of pragmatic reasoning. In particular, all these approaches emphasize *reasoning* and *inference* about a partner's underlying mental state. In this respect, these theories contrast with the popular interactive alignment model (Pickering & Garrod, 2004) and other dynamical systems models of dialogue (e.g. Fusaroli & Tylén, 2015), where coordination occurs through a process of priming and adjusting to a partner's syntactic, lexical, and phonological choices.

While there is a growing literature on dialogue models in general, far less attention has been devoted to the specifics of question and answer dynamics. The primary theoretical work in this area derives from formal linguistic theories, which focused on the notion of informativeness. In Groenendijk and Stokhof's (1984) foundational work on question and answer semantics, asking a question induces a partition over the space of possible worlds, where each cell of the partition corresponds to a possible answer. An answer, then, consists of eliminating cells in this partition, and the most useful answers are those that eliminate all relevant alternatives to the true world. However, as van Rooy (2003) and others (Ginzburg, 1995) have pointed out, this predicts that *wh*-questions like "Where can I buy an Italian newspaper?" can only be fully resolved by exhaustively mentioning whether or not such a newspaper can be bought at each possible location. Clearly, this is not the case: a single nearby location would suffice. These theories also cannot account for contextual variation in what counts as a useful answer.

More recent theories have tried to fix these problems by introducing some consideration of the questioner's goals. Van Rooy (2003), for instance, formalizes these goals as a decision problem faced by the questioner. A useful answer under this decision theoretic account is one that maximizes the expected value of the questioner's decision problem. A useful question is one that induces a sufficiently fine-grained partition, optimally distinguishing the worlds relevant to the decision problem. While this framework elegantly accounts for the context-dependence and relevance-maximization of question and answer behavior, it assumes that the questioner's decision problem is known a priori by the answerer. If this were the case, the act of asking questions would seem irrelevant: why wouldn't the answerer directly tell the questioner which action to take? Nevertheless, while empirical work has almost exclusively focused on answerer behavior, van Rooy's approach suggests that the question itself is important in prompting a relevant answer.

Our models are an effort to expand on this core idea in a probabilistic framework that also provides an inferential mechanism for the answerer to infer the 'decision problem' instead of assuming it. We now proceed to show how these models address four classic examples of question and answer pragmatics from the psycholinguistics literature.

Four case studies

Clark (1979), Experiment 4

First, we show that our model can provide different—sometimes over- or under-informative—answers to the same explicit question, depending on context. For this illustration, we model the whiskey-pricing study presented in the Introduction (Clark, 1979). Recall that liquor merchants were more likely to give over-informative answers (specifying exact price) to the question “Does a fifth of Jim Beam cost more than \$5?” in the uninformative context (“I want to buy some bourbon”) than in the five dollar context (“I’ve got \$5 to spend”).

Our world state is simply the whiskey’s price (\$1, \$2, \dots , \$10). There are two possible goals: learning the price of whiskey and learning whether the price is greater than \$5. The set of answers includes exact prices as well as “yes” and “no”, with lower cost for “yes” and “no” than the price statements. We model the context sentence as affecting the questioner’s goal prior. When the context is “I’d like to buy some whiskey,” we assume that the two goals are equally likely. When it is “I only have \$5 to spend,” we assume that it is 9:1 in favor of learning whether the price is greater than \$5.

Results. When the question is “Does Jim Beam cost more than \$5?”, the correct Boolean answer is the most probable choice (at probability .44 and .49). Critically, there is a context-dependence for answers to this question: when prefaced with “I’d like to buy some whiskey,” the correct exact price answer is favored more strongly (at probability .18) than when the context is “I only have \$5 to spend.” (probability .11). By contrast, the literal and explicit answerers (which have no natural way to account for context) do not make differential predictions in the two situations. The literal model predicts that the answerer is equally likely to say the true Boolean answer and the true numerical answer, and the explicit model predicts that the answerer will always give the true Boolean answer, since it is the explicit question being asked. This suggests that our pragmatic *answerer* is consistent with human behavior in psychologically interesting situations, passing a first, qualitative, test.

Groenendijk and Stokhof (1984)

For a slightly more complex example, we consider the classic puzzle of *mention-some* and *mention-all* readings of wh-questions (Groenendijk & Stokhof, 1984; Schulz & Van Rooij, 2006). Some questions, like “Who is coming to dinner tonight?” are intended to elicit an exhaustive list of the entities that are answers. For other questions, like “Where can I find a bathroom in this building?”, a single answer would be sufficient.

The question “Where can one buy an Italian newspaper?” can intuitively be ambiguous between these meanings depending on who is asking: if it is a tourist, they probably just want to know the nearest place, but if it is a businessperson trying to build a newspaper distribution network in town, they likely want the whole list. The puzzle is determining how the same question can take on different semantics in different contexts: according to our account, this happens via an inference about the questioner’s underlying goal.

Our world state is an object consisting of four cafes in town. Each cafe is assigned two properties: its distance from the speaker and whether or not it sells Italian newspapers. There are two possible goals: learning the identity of the *nearest* cafe selling a newspaper and learning the identity of *all* cafes selling a newspaper. The set of answers includes all 16 combinations of different cafes (e.g. “cafe 1 and cafe 3” or “cafe 2, cafe 3, and cafe 4”), as well as the answer “none”.

The prior over answer utterances is constructed as follows: there is a 10% chance of saying “none,” . Otherwise, the agent selects one of the four cafes and terminates with probability .5. If the agent does not terminate, they pick another cafe from the list and continue until either terminating or running out of possible cafes. This naturally gives longer answers lower probability, reflecting their higher cost of utterance. We model the context sentence as affecting the questioner’s goal prior: if they say “I’m new here,” there is a 9:1 chance that they are interested in the closest location with a newspaper; if they say “I’m a businessperson. . .,” there is a 9:1 chance that they are interested in all of the newspaper locations.

Results. For concreteness, we set the world to be the following :

```

world = { 'cafe1 ' : [3, false],
          'cafe2 ' : [1, true],
          'cafe3 ' : [3, true],
          'cafe4 ' : [3, true]}

```

and enumerated over all model executions for both contexts. We find that the highest probability response given the “I’m new here” context is the single location “cafe2,” with probability 0.56. Given the “I’m a businessperson. . .” context, however, the highest probability response is the conjunction “cafe2 and cafe3 and cafe4,” with probability 0.82. Note that cafe 1 was not assigned any probability in either context because it did not sell Italian newspapers (demonstrating that the pragmatic answerer will not lie), and that the nearest cafe was given precedence in the “tourist” context.

The literal and explicit answerers, as in the Clark scenario, do not have the means of making inferences about the questioner’s underlying goals. Thus, both models incorrectly predicted that the context would not affect the preferred response. The literal answerer predicted that all combinations of cafes 2, 3, and 4 would be given in proportion to their prior likelihood (with no special preference given to the closest one). The explicit answerer predicted that cafe2 would be preferred in all contexts. Crucially, the only difference between our model of this scenario and the Clark scenario is the set of QUDs, the structure of the world, and the meanings of the answers. The questioner and answerer functions stayed exactly the same.

Gibbs Jr. & Bryant (2008): Experiment 3

It has been shown in previous work that people typically round their answers to the nearest 5 or 10 minute interval when asked ‘Do you have the time?’, even when they’re wearing a digital watch (Der Henst et al., 2002). Gibbs Jr and Bryant (2008) replicated this result, and then performed a follow-up study where they preceded their question by the context “I have a meeting at 4:00.” They found that the tendency to round times decreased as a function of the time remaining until the stated deadline: when people were asked at 3:40, they would say “It’s 3:45,” but when people were asked at 3:53, they would make their response precise to the minute. They explain this result by appealing to the

questioner’s goals: while an approximate time is sufficiently informative with respect to most goals, a questioner who is running late to an appointment may have the goal of judging whether to rush, in which case a precise time is needed.

We take the world state to be the true current time. Unlike in the previous two scenarios, where there are only two QUDs and two contexts that make each QUD more or less likely, we now let there be a *family of QUDs* and a fixed context stating the time of the appointment. This family of QUDs is parameterized by a threshold time, below which times are rounded to the nearest 5 minute increment and above which times are given exactly. For example, the QUD with the threshold set at 3:50 corresponds to a goal of wanting to learn the true time if it is higher than 3:50 and only caring about the approximate time below 3:50. The prior over thresholds ($\tau = 3:30, 3:35, \dots, 3:55$), given a context sentence, weights each threshold by its distance to the appointment time, reflecting the intuition that the questioner will be more interested in the exact time as it approaches the appointment time.

The set of answers is the set of times that could be given (e.g. 3:30, 3:31, \dots , 4:00), with uniform probability. Critically, the meaning of an answer is not exact: it introduces a factor in the “interpreter” component of the model, upweighting world states that are close to the answer utterance. In other words, we specify an “approximate” semantics for number references, which is necessary for rounded answers to be produced in the first place.

Results. To test the behavior of this model, we compare two simulations. In both, the context is “I have an appointment at 4:00.” In the first, the true world state is 3:34 (the ‘appointment far’ condition); in the second, the true world state is 3:54 (the ‘appointment near’ condition). Gibbs Jr. and Bryant (2008) found that the answerer is most likely to round to 3:35 in the ‘appointment far’ condition, but most likely to give the exact time, 3:54, in the ‘appointment near’ condition. Indeed, in the ‘appointment far’ condition, the pragmatic questioner model, which infers which threshold is likely given the context and threshold, replies the rounded time ‘3:35’ with probability 0.40 compared to the true time ‘3:34’ with probability 0.27. In the ‘appointment near’ condition, this pattern reverses: the pragmatic questioner model replies the rounded time ‘3:55’ with probability 0.16 and the exact time ‘3:54’ with probability 0.68 (see Figure 1).

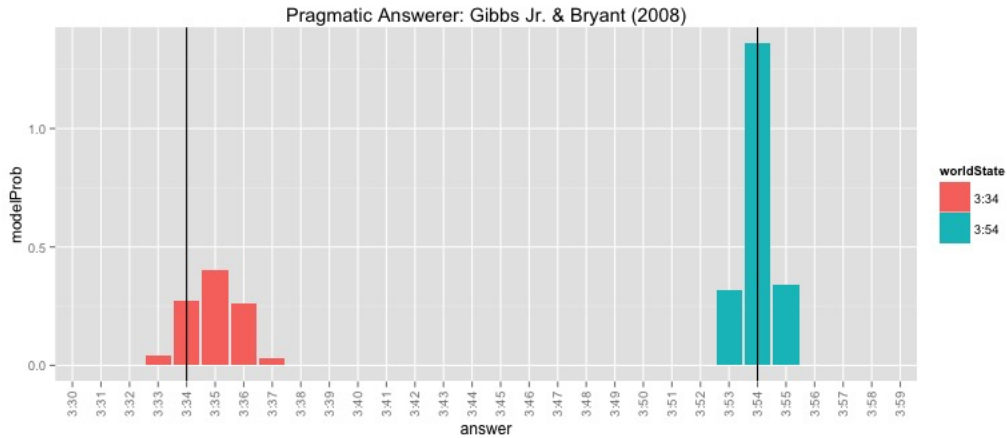


Figure 1. Results for our computational experiments replicating Gibbs Jr. and Bryant (2008). Vertical lines represent the true world state.

Clark (1979): Experiment 5

Our final scenario is a critical test for the questioner component of our model: if our RSA model is correct, the questioner’s choice of utterance itself should guide a pragmatic answerer’s inferences about likely underlying goals. In a follow-up to the first experiment we modeled in this paper, Clark (1979) tested the inferences the merchant made about the questioner’s goals on the basis of their question utterance, without hearing an explicit context sentence. The goal of the questioner was to find out all the kinds of credit cards the merchant accepted, and they could ask questions at different levels of specificity, e.g. “Do you accept Master Charge cards?” or “Do you accept credit cards?” or “Do you accept any kinds of credit cards?” He found that as the noun used in the question became less specific, literal *yes/no* responses were less likely and the alternative (over-informative) response *We accept x, y, z cards.* were preferred.

We have also not yet implemented this model, but the plan is to represent a world state as an object mapping each possible kind of credit card (Visa, Discover, MasterCard) to a boolean value representing whether the establishment will accept that card. There are four possible goals: learning whether or not the establishment accepts (1) Visa, (2) Discover, (3) MasterCard, or (4) learning the exact list of cards that the establishment accepts. The set of answers includes “yes,” “no,” and all possible lists of the three cards

listed above, with “yes” and “no” slightly preferred for their lower cost. Because there is now a decision between different questions that the questioner must entertain, we must also specify a *questioner* prior, which we set to uniform for simplicity.

This suggests that our pragmatic *answerer* is consistent with human behavior in psychologically interesting situations, passing a first, qualitative, test. However, we have not yet shown that the *questioner* behavior matches that of humans. Indeed, the questioner has been largely neglected in studies of answering (but see, e.g., Potts, 2012), even though, as our opening example illustrates, the choice of question is important for understanding answers. In the next section we introduce an experimental paradigm that allows us to jointly explore quantitative behavior of both questioners and answerers.

Exp. 1: Hierarchical questions and answers

In order to simultaneously test how questioners choose questions when faced with a particular goal and how answerers respond under uncertainty about this goal, we used a guessing-game task played by two players: a questioner and an answerer. In this game, 4 animals (a dalmatian, a poodle, a cat, and a whale) were hidden behind 4 gates. These animals corresponded to different levels in a class hierarchy (see Fig. 2). The questioner received a private goal of finding one of the objects (e.g. ‘find the poodle’), and the answerer (but not the questioner) knew the location of each object. Before choosing a gate, the questioner asked the answerer a single question, chosen from a restricted set of options, and the answerer responded by revealing the object behind a single gate. This restriction was motivated by one of the key features of our opening example: when the most direct question (“can I eat your food?”) is suppressed due to politeness, utterance length, complexity, or some other intervening factor, questioners must rely instead on an indirect question.

This set of restricted options was critical to distinguishing between the pragmatic and explicit variants of our model. If all questions were equally available, both our ‘explicit’ and ‘pragmatic’ questioner models would prefer the most direct one. To see how they make different predictions in the presence of restrictions, suppose ‘poodle?’ was not available the questioner. If the questioner asked about a ‘dog?’, the poodle and dalmatian would be

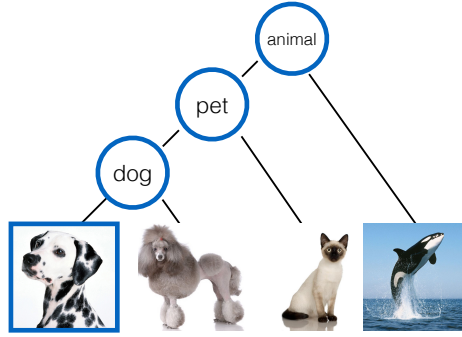


Figure 2. Stimulus hierarchy used in Exp. 1. The goal space and answer space contained the four leaves. The question space, however, was restricted to the highlighted nodes, proceeding up the hierarchy, allowing for indirect questions.

considered equally good options by an explicit answerer because they are both dogs. However, the pragmatic answerer could reason that if the questioner was truly interested in the location of the dalmatian, he would have *asked* about the dalmatian. Because he didn’t, he must be interested in the other valid response that he lacks a direct question for: the poodle.

Participants. We recruited 125 participants from Amazon’s Mechanical Turk to participate in this task. Eleven participants were excluded due to self-reported confusion about the task instructions or due to being non-native English speakers.

Stimuli & Procedure. In terms of our model specification, the world space \mathcal{W} was the set of $4! = 24$ possible assignments of four objects to four gates. The goal space \mathcal{G} was the set of four objects that the questioner could be trying to find (the leaves of the tree in Fig. 2). The answer space \mathcal{A} was the set of four gates that the answerer could reveal. The restricted question space \mathcal{Q} contained the set of highlighted nodes in the hierarchy: ‘dalmatian?’, ‘dog?’, ‘pet?’, and ‘animal?’.

Each participant provided responses for four trials in the role of the questioner (corresponding to the four goals), and four trials in the role of the answerer (corresponding to the four possible questions). In the questioner block, players were presented with a private goal from \mathcal{G} , like “find the poodle!” and were prompted to select a question from a drop-down menu containing elements of \mathcal{Q} that would best help them find it. In the

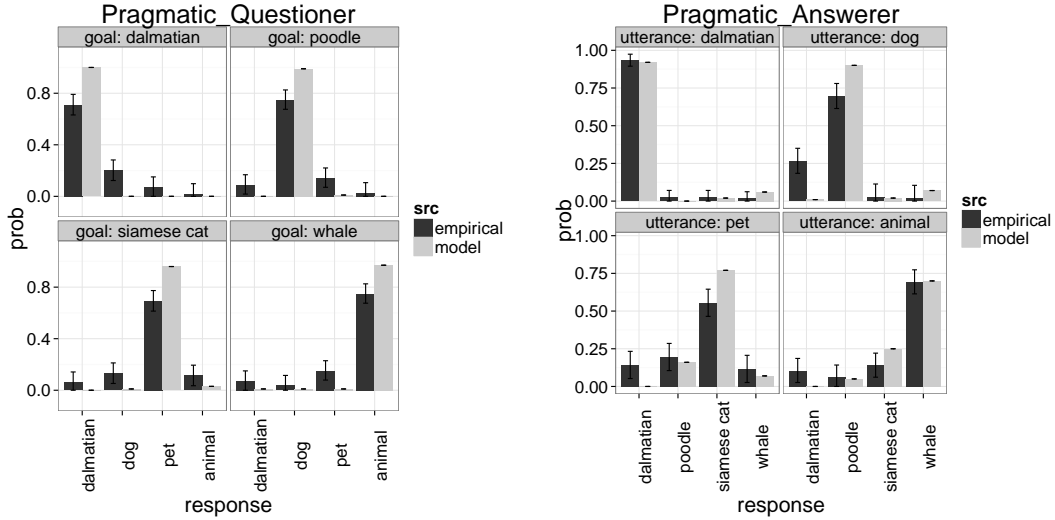


Figure 3. Exp. 1 results, compared with the predictions of the best-performing model for questioner (left) and answerer (right). The explicit and pragmatic questioner models do not make different predictions in this task, but the pragmatic answerer better accounts for the qualitative patterns in the response data than the explicit answerer.

answerer block, players were shown which items were behind which gates and were told that the other player had asked a particular question from \mathcal{Q} . They were prompted to select a gate from a drop-down menu that would be most helpful for the questioner, keeping in mind his or her constraints. (To minimize learning effects, questioners did not receive answers and neither role saw the outcome of the game.) In order to collect responses for all elements of \mathcal{G} and \mathcal{Q} , the order of the questioner and answerer blocks was randomly assigned for each participant, and the order of stimuli within these blocks was also randomized

Results. Results for the questioner role are shown alongside model predictions in Fig. 3 (left). We find that questioners systematically prefer to ask different questions given different goals, even as those questions become more indirect. χ^2 tests over each of the four response distributions show a significant divergence from uniform. Questioners preferentially ask about the ‘dalmatian’ given the dalmatian goal, $\chi^2(3) = 137, p < .001$, about the ‘dog’ given the poodle goal, $\chi^2(3) = 152, p < .001$, about the ‘pet’ given the cat goal, $\chi^2(3) = 120, p < .001$, and about the ‘animal’ when given the whale goal,

$$\chi^2(3) = 150, p < .001.$$

Results for the answerer role are shown in Fig. 3 (right). Answerers are highly sensitive to the constraints of the questioner, giving information about the dalmatian when asked about a ‘dalmatian’, $\chi^2(3) = 281, p < .001$, about the poodle when asked about a ‘dog’, $\chi^2(3) = 137, p < .001$, about the cat when asked about a ‘pet’, $\chi^2(3) = 57, p < .001$, and about the whale when asked about an ‘animal’, $\chi^2(3) = 121, p < .001$. Note that, under an explicit interpretation of the question, revealing the dalmatian and the poodle would both be perfectly acceptable answers to a question about a ‘dog’, but answerers strongly prefer to give the location of the poodle. In the next section, we compare these results to the predictions of our family of models (Fig. 4).

Model comparison. Each model was run with uniform prior probability over worlds, goals, questions, and answers, and with equal cost for all utterances. For each model, a single optimality parameter, which applied to all agents as described above, was fit to maximize correlation with the data.

We can rule out both the literal answerer and literal questioner. The *literal answerer* yields a uniform distribution over the four answers. This has consequences for the corresponding *literal questioner* model: when this questioner reasons about which question would generate the most helpful answer from the literal answerer, it finds no differences in response probabilities, and therefore has no preference for which question to ask. The predictions of these model, plotted against our empirical results, are shown in the left-hand column of Fig. 4.

The two remaining questioner models make roughly the same predictions for this task, and we are not able to distinguish them on the basis of these data. We found a model-data correlation of $r = 0.96$ for the explicit questioner and correlation of $r = 0.99$ for the pragmatic questioner. Although the pragmatic model has a slightly better fit, the two models only differ slightly in the magnitude of predictions, not in qualitatively important ways such as the rank ordering of response. The pragmatic questioner model’s predictions for each response distribution are shown in Fig. 3 (left). Although the magnitude of its predictions are not in perfect alignment with the magnitude of the empirical data (because it is strongly optimizing), it captures most of the interesting qualitative patterns of the

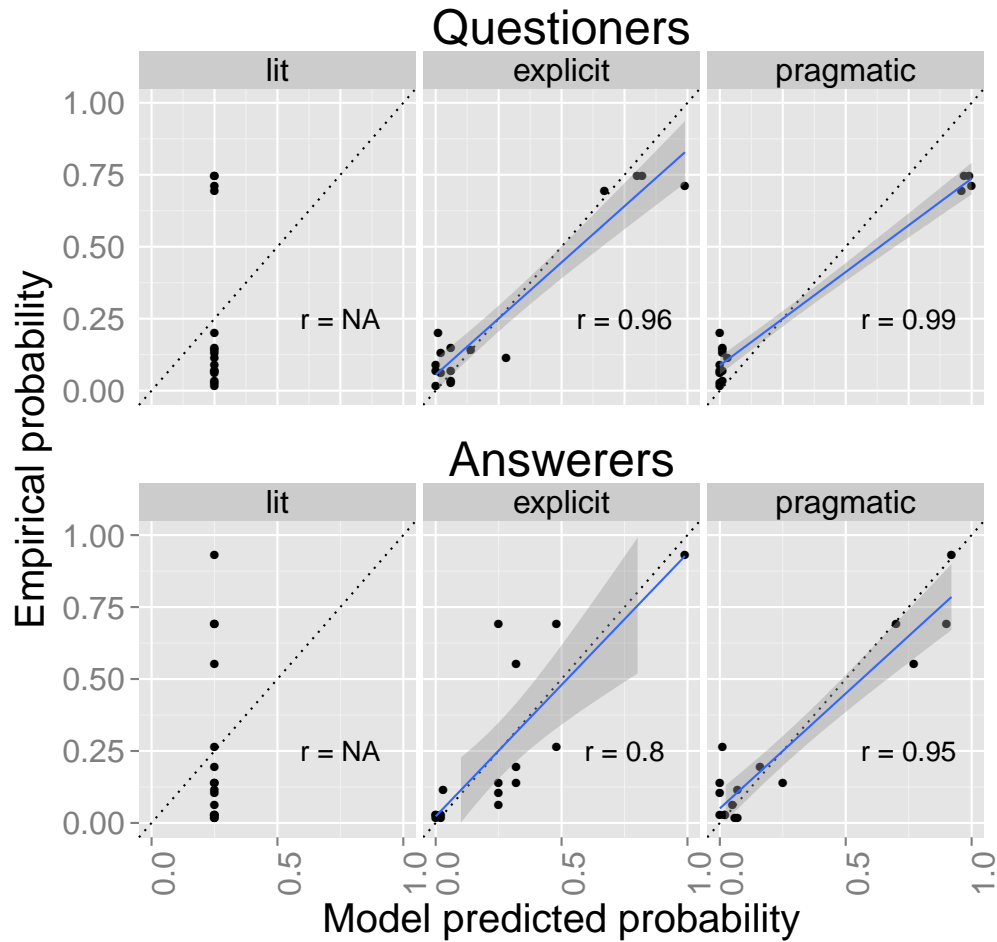


Figure 4. Full space of models, and their correlations with the data from Exp. 1.

Questioner models in the first row reason about the answerers directly below them, and the pragmatic answerer reasons about the explicit questioner.

data, particularly the modal responses.

The pragmatic answerer provides a much better fit to the data than the explicit answerer: a model-data correlation of $r = 0.8$ for the explicit answerer and $r = 0.95$ for the pragmatic answerer.

Discussion. Only the pragmatic answerer can account for essential qualitative features of the response data. For example, the explicit answerer predicts that participants will be equally likely to show the ‘dalmatian,’ ‘poodle,’ and ‘cat’ when asked about a pet. Instead, the data show a significant preference for revealing the cat, leaving ‘dalmatian’

and ‘poodle’ at the same level as the other alternative. The pragmatic answerer correctly predicts this pattern (see Fig. 3 (right)). Even more dramatically, the explicit answerer predicts a uniform distribution over responses to the ‘animal?’ question. However, the empirical distribution was significantly different from uniform. Thus, the pragmatic answerer is necessary to account for these data.

These data provide strong evidence for a pragmatic answerer, but are more equivocal with respect to the explicit and pragmatic questioner. Because the two models did not make significantly different predictions for this experiment (and both work quite well), we ran a follow-up study on a special case of the guessing-game paradigm in which the explicit and pragmatic questioners make different predictions.

Exp. 2: A Critical Test of Questioner Models

Participants. We recruited 50 participants to participate only in the questioner scenario of the guessing game presented above. Ten participants were excluded on the basis of having a non-English native language, or reporting confusion about the instructions.

Stimuli & Procedure. The procedure was the same as before with some changes to the stimuli. The world space \mathcal{W} consisted of possible assignments of the three pets to three gates. The possible goals \mathcal{G} were the dalmatian and poodle (not the cat). The possible questions \mathcal{Q} were ‘dalmatian?’ or ‘cat?’. The possible answers \mathcal{A} were the three gates. Each participant was given the two goals in a random order

Results. When the goal was to find the dalmatian, participants were significantly more likely to ask about the dalmatian than the cat, $\chi^2(1) = 12, p < 0.001$. When the goal was to find the poodle, participants were marginally more likely to ask about the cat than the dalmatian, $\chi^2(1) = 3.6, p = 0.058$. When looking only at the first of the two trials, the dalmatian result held, $\chi^2(1) = 14.4, p < 0.001$, but participants’ preference for asking about the cat disappeared, $\chi^2(1) = 0.07, p = 0.79$. These results are shown in Fig. 5.

Model comparison. The explicit questioner predicts that participants should have no preference for a question given the ‘poodle’ goal, since an explicit answerer would be equally unlikely to give the desired answer for both. The pragmatic questioner model, however, predicts that participants should prefer to ask about the cat. This is because the

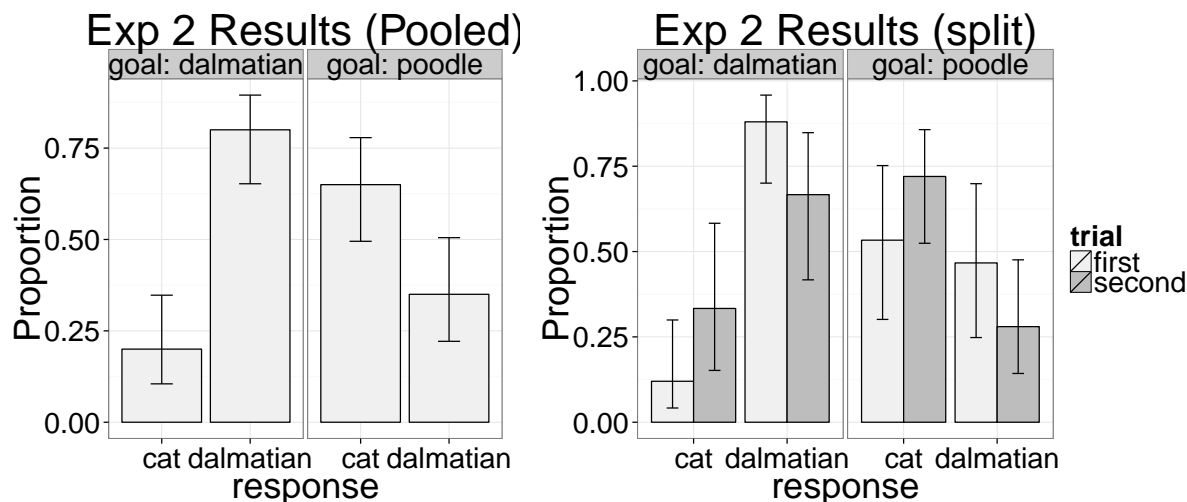


Figure 5. The overall response distribution in Exp. 2 (left) and the same distribution split into first- and second-trial data (right).

(internal) pragmatic answerer would reason that if the questioner was interested in the dalmatian, they would ask about the dalmatian; if they didn’t, they must be interested in the other possible goal.

It is again unclear which questioner model is best. Overall, the response distribution matches the predictions of the pragmatic model: questioners prefer to ask about the cat. However, participants don’t show this behavior if we look at only the first trial. This could be due to a number of reasons. Interestingly, the pragmatic model predicts a more explicit-like response distribution if the questioner does not take into account the constraint on possible goals: if participants thought the poodle was the only goal (counter to the instructions), then asking about the dog would be consistent with the pragmatic model as well. It is possible that participants only fully-processed the alternative (dalmatian) goal if they had first done the trial where that was the goal.²

Discussion. Experiment 1 established that the pragmatic answerer model is necessary to account for answerer behavior in our simple guessing-game task, complementing results from our computational experiments. Additionally, it established that the explicit and pragmatic questioner models could capture qualitatively important features of the questioner data, such as the systematic preference for under-informative

²This pattern was also replicated with a simpler set of stimuli containing red and blue squares and circles.

questions when the explicit label for an object was unavailable. However, both experiments 1 and 2 failed to distinguish between the explicit and pragmatic questioner models. In experiment 1, they made indistinguishable predictions; in experiment 2, the response data were confounded with participants’ beliefs about goal alternatives.

Certain aspects of the task are subject to other methodological concerns. For instance, participants only gave hypothetical judgements about what they *would* say given different goals or questions instead of playing out the full game, and participants did not believe that they were playing with a real partner. These issues contributed to overall confusion about the task, and potentially reflected abstract logic-based reasoning instead of the social and linguistic intuitions we intended to test.

In the following two experiments, we modified the task in two ways to address these concerns. In experiment 3, we replicated experiment 1 in a real-time, multi-player environment where participants are assigned to either the “questioner” role or the “answerer” role and interact directly with one another as they play the full game. In experiment 4, we additionally expanded the set of items beyond the single animal hierarchy used in the previous experiments. This item set includes four different domains (animals, plants, places, and artifacts), crossed with three different hierarchy structures eliciting a larger range of predictions from our models. One of these hierarchy structures was specifically designed to provide a critical test for distinguishing the explicit and pragmatic questioner models. Unlike experiment 2, this condition did not require participants to reason about unintuitive question alternatives that do not pick out any existing items in the world.

Exp. 3: Interactive Questions and Answers

Participants. We recruited 50 participants from Amazon’s Mechanical Turk to participate in this task: 25 were assigned to the questioner role and 25 to the answerer role, yielding 25 complete games.

Stimuli & Procedure. The world space \mathcal{W} , goal space \mathcal{G} , question space \mathcal{Q} , and answer space \mathcal{A} were the same as in experiment 1 (see Figure 2). The procedure was modified to accommodate real-time player-to-player interaction following Hawkins (2014).

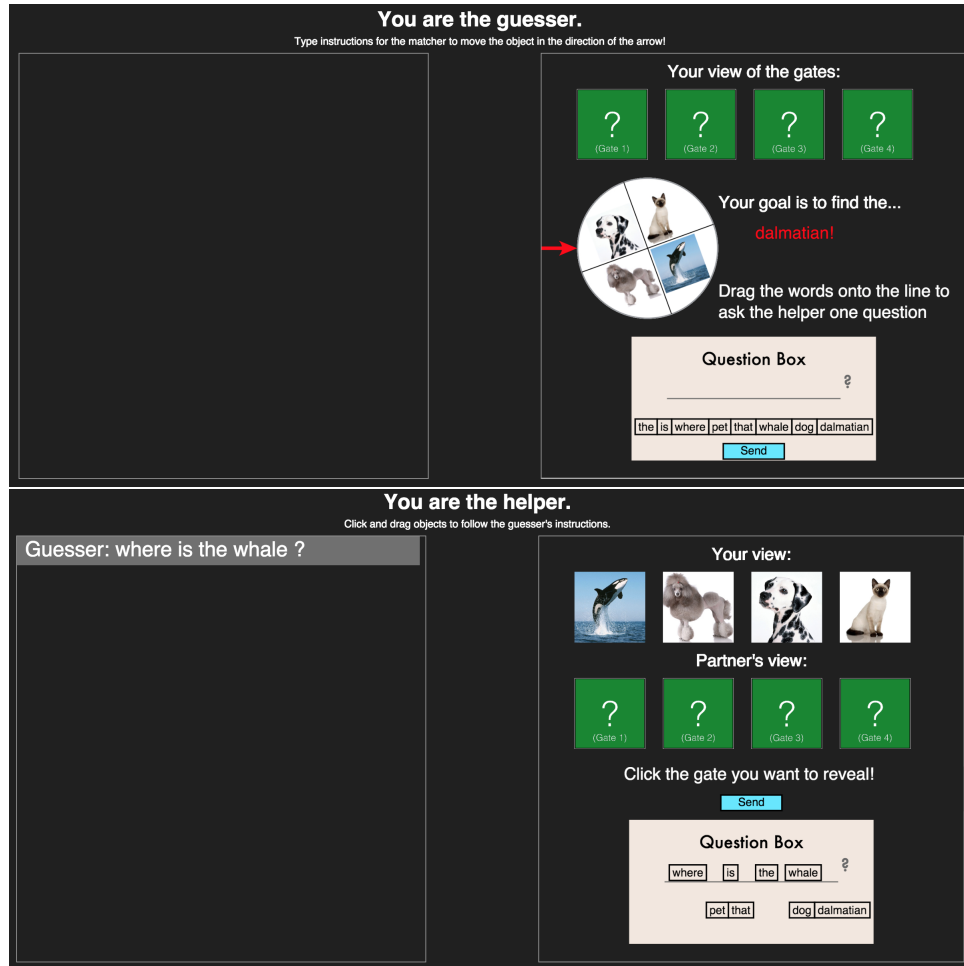


Figure 6. Exp. 3 interfaces, for the questioner (top) and answerer (bottom).

All players passed a short quiz on the game instructions and were immediately redirected to the game interface: the first player to join was assigned to be a “questioner” (which we called a “guesser” in the cover story) and told to wait until a second player was available. Once another player joined, they were assigned to be an “answerer” (or “helper”) and the game began.

The questioner and answerer interfaces are displayed in Figure 6. Chat messages were printed on the left side of the screen, and players used the right side as a workspace to view goals, ask questions, and respond with answers. At the beginning of each trial, the wheel on the questioner’s screen (Figure 6, top) would spin and select one of the four goals. The questioner then clicked and dragged words onto the line in the “Question box” to ask a question to help them find this goal. The answerer saw these words being dragged in real

time. Once the questioner clicked the ‘send’ button, the resulting question appeared in the chat log and control was passed to the answerer, who clicked on one of the four gates to send a response containing the location of the chosen object. Finally, the questioner was asked to guess which gate they believed the goal object was behind.

Each participant provided responses for four trials, where object locations were shuffled and a new goal was randomly selected for each trial. Thus, questioners could be given the same goal on multiple trials, preventing ‘process of elimination’ reasoning about what the goal may be on the part of the answerer.

Results. Results for the questioner role are shown alongside model predictions in Fig. 10 (left). We find that questioners systematically prefer to ask different questions given different goals, even as those questions become more indirect. χ^2 tests over each of the four response distributions show a significant divergence from uniform. Questioners preferentially ask about the ‘dalmatian’ given the dalmatian goal, $\chi^2(3) = 77, p < .001$, about the ‘dog’ given the poodle goal, $\chi^2(3) = 50, p < .001$, about the ‘pet’ given the cat goal, $\chi^2(3) = 47, p < .001$, and about the ‘animal’ when given the whale goal, $\chi^2(3) = 39, p < .001$.

Results for the answerer role are shown in Fig. 10 (right). Answerers are highly sensitive to the constraints of the questioner, giving information about the dalmatian when asked about a ‘dalmatian’, $\chi^2(3) = 102, p < .001$, about the poodle when asked about a ‘dog’, $\chi^2(3) = 47, p < .001$, about the cat when asked about a ‘pet’, $\chi^2(3) = 45, p < .001$, and about the whale when asked about an ‘animal’, $\chi^2(3) = 31, p < .001$. In the next section, we compare these results to the predictions of our family of models (Fig. 11).

Model comparison. Model comparison was conducted in the same way as in Experiment 1: a single optimality parameter, which applied to all agents, was fit for each of the six models to maximize correlation with the data.

We can again rule out both the literal answerer and literal questioner as they predict a uniform distribution of responses over the four questions and answers. The two remaining questioner models again make roughly the same predictions for this task: we found a model-data correlation of $r = 0.971$ for the explicit questioner and correlation of $r = 0.996$ for the pragmatic questioner. The difference between these correlations is

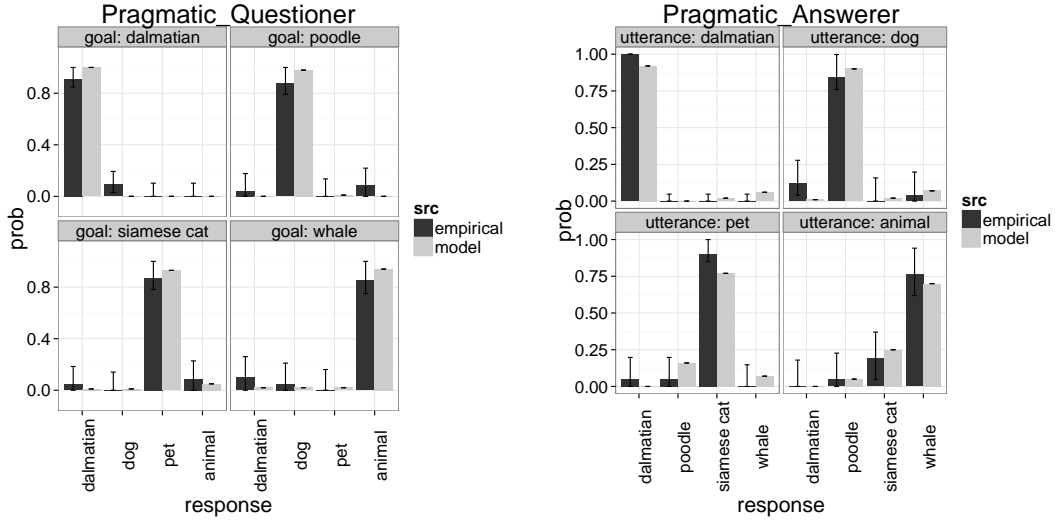


Figure 7. Exp. 3 results and model fits, for the best-performing questioner (left) and answerer (right) models.

statistically significant, accounting for their shared dependence on the empirical data (Zou’s confidence interval = $[-0.079, -0.009]$). However, they make nearly identical qualitative predictions; the pragmatic questioner model’s predictions for each response distribution are shown in Fig. 10 (left).

The pragmatic answerer provides a much better fit to the data than the explicit answerer: we find a model-data correlation of $r = 0.7$ for the explicit answerer and $r = 0.99$ for the pragmatic answerer. Taking into account the fact that these correlations are dependent and overlapping on the same empirical data, we find that the pragmatic answerer correlation is significantly larger than the explicit answerer correlation (Zou’s confidence interval = $[-0.676, -0.107]$).

Discussion. We replicated the results of experiment 1 in a real-time, interactive setting. Again, we found that both explicit and pragmatic questioner models provide an excellent fit to questioner behavior, and that the pragmatic answerer accounts for the data significantly better than the explicit answerer both quantitatively and qualitatively. In addition, the full, interactive game was designed to be more natural and less confusing to participants than the drop-down menu design from experiments 1 and 2. Because players received constant feedback from their partner, this task was framed as inherently social,

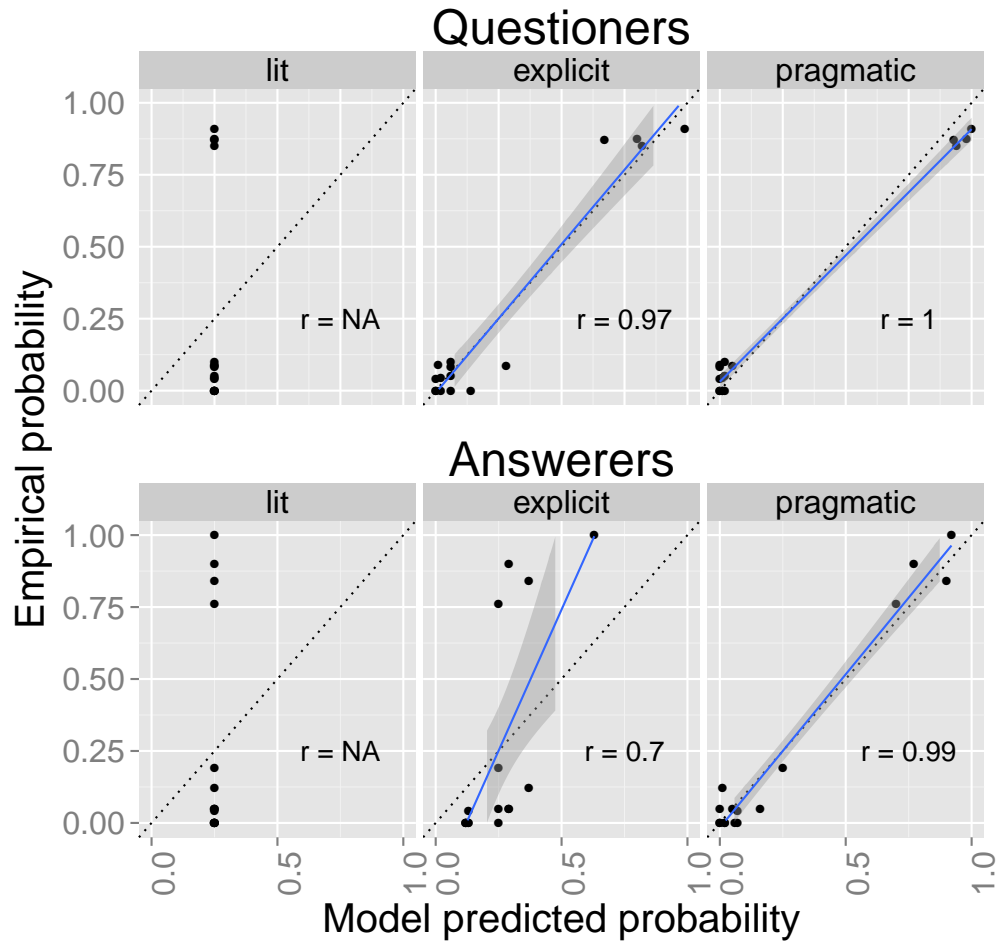


Figure 8. Full space of models, and their correlations with the data from Exp. 1.

Questioner models in the first row reason about the answerers directly below them, and the pragmatic answerer reasons about the explicit questioner.

and because answerers watched as questioners moved words to form questions, there was a convincing mechanism for people to believe they were playing with another human. This addresses some concerns raised with experiments 1 and 2.

Note that so far we have used the same animal hierarchy for the stimulus set in all our experiments, providing only 16 points of comparison between our models and empirical data. Furthermore, there exists a heuristic strategy for selecting questions given goals in the hierarchy structure we have been using which produces the same pattern of responses without requiring any social inference. Suppose questioners saw their goal on a given trial

and ruled out labels that do not apply (e.g. a ‘cat’ is neither a ‘dalmatian’ nor a ‘dog’), then picked the most specific of the remaining labels (‘pet’ picks out a smaller set of objects than ‘animal’).

In experiment 4, we test the generality of our model by expanding the stimulus set to encompass multiple stimulus domains and multiple hierarchy structures. This addresses the potential concern that the behavioral patterns we have been modeling are specific to the set of animals we decided to use or the tree-like hierarchy in which they were embedded. We also took care to include one simple but critical hierarchy structure where (1) the explicit and pragmatic questioner models make different predictions and (2) the heuristic strategy cannot explain these predictions.

Exp. 4: Generalizing Predictions

Participants. We recruited 199 participants from Amazon’s Mechanical Turk to participate in this task. Fifty participants were excluded due to a server crash that terminated the task before completion. Two additional participants were excluded because they were not non-native English speakers. This left 74 unique completed games.

Stimuli & Procedure. A set of twelve items were created by crossing four domains (animals, plants, places, and artifacts) with three hierarchy structures (“branching”, “overlapping”, and “equivocal”; see Figure ??). For each item, there was a set of four goal objects in \mathcal{G} , which appeared on the wheel for the questioners, four question labels in \mathcal{Q} that the questioner could use to ask about their assigned goal, and four answers in \mathcal{A} corresponding to the four items in \mathcal{G} that appear on the wheel.³

The procedure was the same as Experiment 3 with two major modifications. First, we addressed an asymmetry in the task: the answerer watched the questioner move words, but the questioner only saw template text responses in response. In pilot work, we found that answerers were more likely to believe they were playing with a human partner than questioners. To bring questioner and answerer beliefs closer together, we changed the mechanism by which the answerer responds. Instead of clicking on a gate and clicking the

³A document containing these mappings for all items, as well as their hierarchical relationships, is available online at https://github.com/hawkrobe/Q_and_A/blob/master/MultiExperiment2/stimuliLabels.pdf

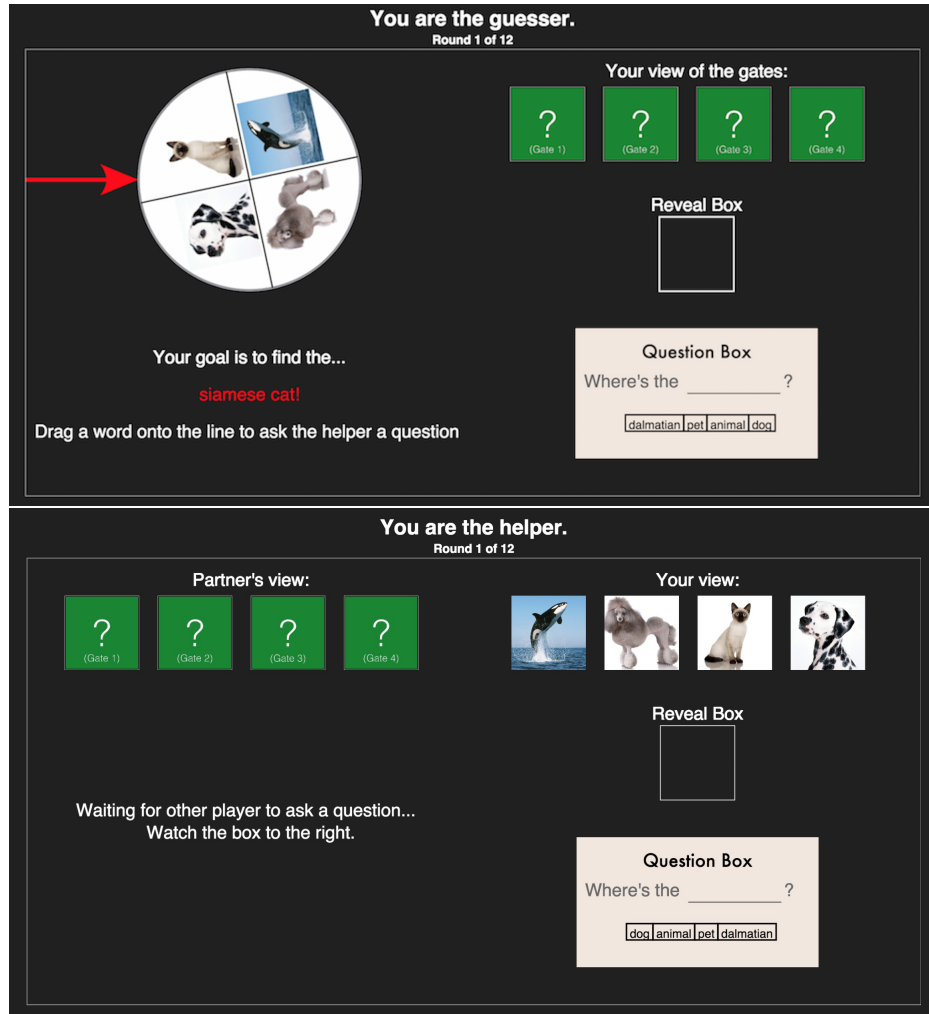


Figure 9. Exp. 4 interfaces, for the questioner (top) and answerer (bottom).

‘send,’ button, we supplied a “Reveal Box.” When the answerer was ready to reveal a gate, they would click and drag the object they wanted to reveal into this box. The questioner watched as the outline of the gate moved, in real time, and saw the image as soon as the answerer dropped it in the box.

Second, we noticed that some players began exploiting the freedom of the “question box” drag-and-drop procedure after playing several rounds. On difficult items, especially those in the equivocal hierarchy where our model predicts no preferred question for certain goals, players would use the words to form non-grammatical signals. This is interesting behavior in its own right, but as we are interested in testing models operating over a small set of question utterances, it was not appropriate for our purposes. We therefore pre-set a

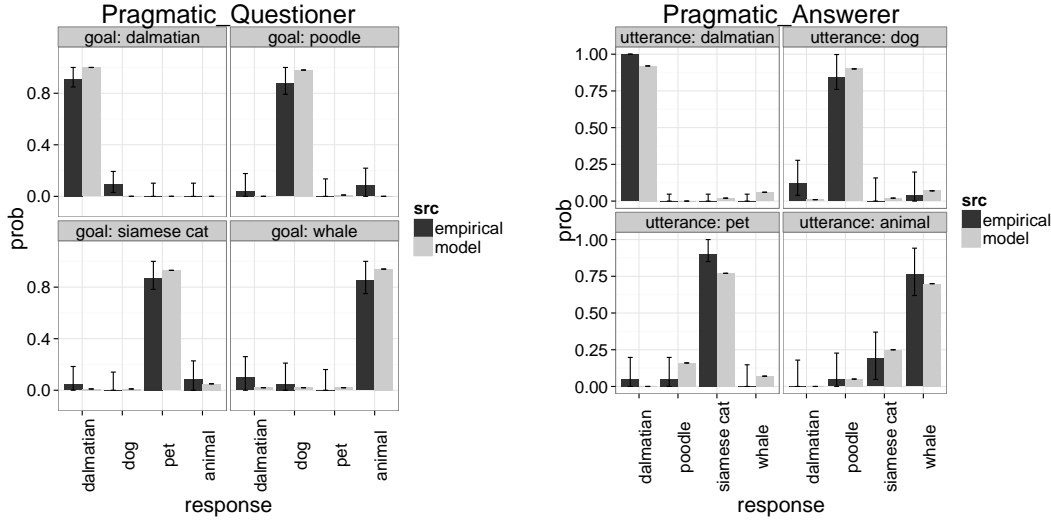


Figure 10. Exp. 3 results and model fits, for the best-performing questioner (left) and answerer (right) models.

frame for the question box and allowed participants to drag one and only one of the question labels into the blank.

The resulting questioner and answerer interfaces are displayed in Figure 9. Each participant provided one response for each of the twelve items, presented in random order.

Results.

Model comparison.

Discussion. We replicated the results of experiment 1 in a real-time, interactive setting. Again, we found that both explicit and pragmatic questioner models provide an excellent fit to questioner behavior, and that the pragmatic answerer accounts for the data significantly better than the explicit answerer both quantitatively and qualitatively. In addition, the full, interactive game was designed to be more natural and less confusing to participants than the drop-down menu design from experiments 1 and 2. Because players received constant feedback from their partner, this task was framed as inherently social, and because answerers watched as questioners moved words to form questions, there was a convincing mechanism for people to believe they were playing with another human. This addresses some concerns raised with experiments 1 and 2.

Note that so far we have used the same animal hierarchy for the stimulus set in all

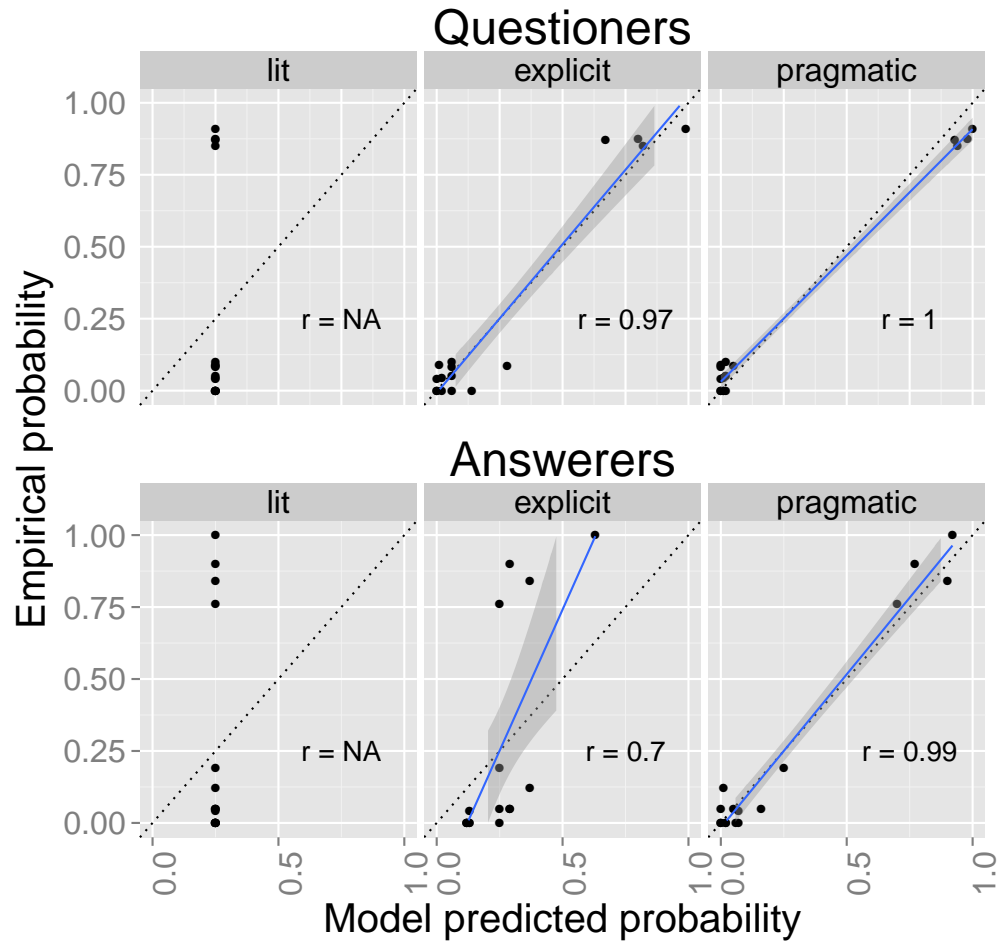


Figure 11. Full space of models, and their correlations with the data from Exp. 1.

Questioner models in the first row reason about the answerers directly below them, and the pragmatic answerer reasons about the explicit questioner.

our experiments, providing only 16 points of comparison between our models and empirical data. Furthermore, there exists a heuristic strategy for selecting questions given goals in the hierarchy structure we have been using which produces the same pattern of responses without requiring any social inference. Suppose questioners saw their goal on a given trial and ruled out labels that do not apply (e.g. a ‘cat’ is neither a ‘dalmatian’ nor a ‘dog’), then picked the most specific of the remaining labels (‘pet’ picks out a smaller set of objects than ‘animal’).

In experiment 4, we test the generality of our model by expanding the stimulus set to

encompass multiple stimulus domains and multiple hierarchy structures. This addresses the potential concern that the behavioral patterns we have been modeling are specific to the set of animals we decided to use or the tree-like hierarchy in which they were embedded. We also took care to include one simple but critical hierarchy structure where (1) the explicit and pragmatic questioner models make different predictions and (2) the heuristic strategy cannot explain these predictions.

General discussion

Perhaps the most important formal advance of the models considered here is to move the Rational Speech Act framework beyond interpretation of single utterances (in context), to consider the dynamics of simple dialogs (albeit consisting of a single question and its answer). Doing so requires replacing the immediate motive to convey true information with the more distant motive to provoke useful information from one’s interlocutor. On the answerer side, sophisticated inference was required to account for the implicit interests of the questioner. This provides a useful connection to current game-theoretic and decision-theoretic models (Vogel et al., 2013; Van Rooy, 2003), which also emphasize the importance of goals and speaker beliefs in communication but emphasize less the complex interplay of inference between questioner and answerer.

We have presented evidence that answerer behavior is best described by a pragmatic model that *does* reason about questioner intentions, using the question utterance as a signal. The superiority of pragmatic answerer predictions over the other answerer models was robust. Questioner behavior in Exp. 2, however, seemed to be much more dependent on experience. In another version of Exp. 1, we did not emphasize certain aspects of the game in the instructions, such as the fact that the answerer knows about the restricted answer set, which might prompt perspective-taking. Our data in this pilot experiment appeared to contain a mixture of explicit and pragmatic answerers and questioners (though other confounds were present in this version). Overall, it will be important to explore the mixture of explicit- and pragmatic-questioning across a larger range of situations: these issues may be a product of our artificial game paradigm, or they may be reflective of real tendencies in language use, raising novel questions about audience design in

question-answer behavior.

While the artificiality of our question-answer game may distance the behavior of participants from the natural use of language, there are also some benefits to this design. In particular, it is easy in this setting to control the exact space of questions, goals, and answers. While the restrictions on question space may seem peculiar, it is directly motivated by conversational scenarios in everyday usage which feature restrictions on the set of things one can ask about, due to politeness, salience, time cost, and other factors. In future work, we will explore the extent to which the proposed model can scale up to real-time, multiplayer games, extended dialogues, and other more naturalistic language settings. To deal with dialogues lasting longer than a single exchange, for instance, we must specify the way in which the contributions of questioner and answerer affect the *context* in which later utterances operate.

Humans are experts at inferring the intentions of other agents from their actions (Tomasello, Carpenter, Call, Behne, & Moll, 2005). Given simple motion cues, for example, we are able to reliably discern high-level goals such as chasing, fighting, courting, or playing (Barrett, Todd, Miller, & Blythe, 2005; Heider & Simmel, 1944). Experiments in psycholinguistics have shown that this expertise extends to speech acts. Behind every question lies a goal or intention. This could be an intention to obtain an explicit piece of information (“Where can I get a newspaper?”), signal some common ground (“Did you see the game last night?”), test the answerer’s knowledge (“If I add these numbers together, what do I get?”), politely request the audience to take some action (“Could you pass the salt?”), or just to make open-ended small talk (“How was your weekend?”). These wildly different intentions seem to warrant different kinds of answers. By formalizing the computational process by which answerers infer these different intentions, our model framework provides a unifying way to accommodate this diversity.

References

- Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, 26(4), 313–331.
- Boër, S. E., & Lycan, W. G. (1975). Knowing who. *Philosophical Studies*, 28(5), 299–344.
- Clark, H. H. (1979). Responding to indirect speech acts. *Cognitive psychology*, 11(4), 430–477.
- Der Henst, V., Carles, L., & Sperber, D. (2002). Truthfulness and relevance in telling the time. *Mind & Language*, 17(5), 457–466.
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), 998–998.
- Franke, M. (2013). Game theoretic pragmatics. *Philosophy Compass*, 8(3), 269–284.
- Fusaroli, R., & Tylén, K. (2015). Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science*.
- Gibbs Jr, R. W., & Bryant, G. A. (2008). Striving for optimal relevance when answering questions. *Cognition*, 106(1), 345–369.
- Ginzburg, J. (1995). Resolving questions, i. *Linguistics and Philosophy*, 18(5), 459–527.
- Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1), 173–184.
- Goodman, N. D., & Stuhlmüller, A. (electronic). *The design and implementation of probabilistic programming languages*. Retrieved 2015/1/16, from <http://dippl.org>
- Groenendijk, J., & Stokhof, M. (1984). On the semantics of questions and the pragmatics of answers. *Varieties of formal semantics*, 3, 143–170.
- Hawkins, R. X. D. (2014). Conducting real-time multiplayer experiments on the web. *Behavior research methods*, DOI 10.3758/s13428-014-0515-6.
- Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 243–259.
- Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33),

12002–12007.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue.

Behavioral and brain sciences, 27(02), 169–190.

Potts, C. (2012). Goal-driven answers in the cards dialogue corpus. In *Proceedings of the*

30th west coast conference on formal linguistics (pp. 1–20).

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal

theory of pragmatics. *Working Papers in Linguistics-Ohio State University*

Department of Linguistics, 91–136.

Schulz, K., & Van Rooij, R. (2006). Pragmatic meaning and non-monotonic reasoning:

The case of exhaustive interpretation. *Linguistics and Philosophy*, 29(2), 205–250.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and

sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*,

28(05), 675–691.

Van Rooy, R. (2003). Questioning to resolve decision problems. *Linguistics and*

Philosophy, 26(6), 727–763.

Vogel, A., Bodoia, M., Potts, C., & Jurafsky, D. (2013). Emergence of gricean maxims

from multi-agent decision theory. In *Hlt-naacl* (pp. 1072–1081).