

Why do you ask? To be informative.

Robert X. D. Hawkins (rxdh@stanford.edu)

Andreas Stuhlmüller (andreas@stuhlmüller.org)

Judith Degen (jdegen@stanford.edu)

Noah D. Goodman (ngoodman@stanford.edu)

Department of Psychology, 450 Serra Mall
Stanford, CA 94305 USA

Abstract

Keywords: questions; answers; computational pragmatics; theory of mind;

Introduction

NDG: there is a lot of good stuff in the intro below, but it is not globally coherent yet. let's discuss the flow of the argument we want to make...

NDG: this discussion of intention inference is good, but i think we should start with a few sentences more directly about questions and answers. i often like to start with an example, if there is a short and illustrative one...

RDH: this is an example I keep coming back to in my own thinking, but I'm not sure it's necessarily the best we can come up with...

How would you respond if someone asked “where are you?” There are a number of valid answers at different levels of coarseness: *in the kitchen*, *at home*, *in Chinatown*, *in San Francisco*, *on Earth*, and so on. Which one do you use? The answer will depend on context and, in particular, on the intentions of the questioner. If you had plans to meet a colleague at a coffee shop and you're running late, you might give a fairly specific answer to indicate you're close by. On the other hand, you might use a coarser answer if you just flew into a new city for a conference and were calling home to say that you're safe. That said, it doesn't seem entirely fair that the burden to make these inferences lies entirely on the answerer. Why couldn't the questioner be clearer about what level of information they care about, by asking “are you nearby?” or “what city did you fly into?”

In this paper, we present three progressively more sophisticated computational models of question and answer behavior, which formalize and probe the deep interaction between the way answerers infer intentions and the way questioners signal them. We show how these models handle different kinds of questions, and compare these models on the basis of two experiments in which participants must ask and answer questions given a fixed set of goals. We close by proposing that the purpose of questions in dialogue is to provide cues about questioner goals and intentions.

For example, when people are asked “Do you have the time?” they typically round their answers to the nearest 5 or

10 minute interval, even when they're wearing a digital watch (Der Henst, Carles, & Sperber, 2002). However, if the question is preceded by the statement “My watch stopped,” people make their response precise to the minute (Gibbs Jr & Bryant, 2008). This demonstrates sensitivity to questioner goals, and a pragmatic imperative to maximize relevance with respect to those goals.

Similar evidence comes from a classic study where researchers called liquor merchants and asked, “Does a fifth of Jim Beam cost more than \$5?” If this was preceded by the statement, “I want to buy some bourbon,” merchants gave the actual price significantly more frequently than when it was preceded by the statement, “I've got \$5 to spend.” In the former case, the merchant inferred that the questioner's goal was just to buy whiskey, so the exact price is the maximally relevant response. In the latter case, 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 (Clark, 1979). Context and questioner goals have also been implicated in accounts of identification questions like “who is X?” (Boër & Lycan, 1975), and to questions like “where are you?” that permit answers at many levels of abstraction (Potts, 2012).

NDG: i would set up the puzzle for (RSA style) models around here? say that recent formal models of pragmatics have made progress specifying the nature and role of information and goals for statements, but they have a problem when it comes to questions? then turn to other's answers?

What makes a question useful? What makes an answer to a question useful? In this paper, we propose an account of question and answer behavior in which the questioner's intentions are *not* known *a priori* by the answerer and instead must be inferred. We propose that a useful answer to a question is one that is maximally informative with respect to this inferred underlying decision problem. A useful question, then, is one that optimally *signals* the questioner's underlying decision problem and has a high probability of resulting in an answer that is maximally informative with respect to that decision problem. This places question and answer behavior in the larger class of social behavior governed by theory of mind.

The rest of this paper is structured as follows. First, we specify a family of questioner and answerer agents extending

the Rational Speech Act (RSA) framework (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013), highlighting some points of divergence from previous RSA models. We then individuate three particular models in this family, representing progressively more sophisticated hypotheses about how questioners and answerers reason about their task. In particular, we compare a pragmatic answerer making inferences about the questioner’s goals to two simpler models: one that takes into account only that an answerer wants to be maximally informative with respect to the explicit question asked (without inferring the questioner’s underlying decision problem) and one that provides a literal answer to the question (without attempting to be maximally informative).

To compare these models, we derive predictions for a pair of experiments using a novel guessing-game task, and compare these predictions to human performance. In one phase of the task, we require participants to ask a question (from a fixed set of possible questions), given a decision problem. In the second phase, we require participants to give an answer (from a fixed set of possible answers) to a question (from a fixed set of possible questions).

A Rational Speech Act model of question and answer behavior

NDG: i think it will make more sense to start with the top-level cognitive models for the Q and A, and then fill down to the sub-agents as increasingly complicated notional communicative partners.

Suppose there is a set of distinct 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 all taken to be in common ground. We begin by thinking about how an optimal questioner should choose a question. Critically, instead of trying to impart information about the state of the world, a questioner attempts to *learn information about a private goal*, sometimes called a QUD (or question under discussion) (Roberts, 1996). A goal $g \in \mathcal{G}$ is a function that transforms a complete world state to a particular feature or features of that world that the questioner cares about, thus serving as a utility function on the distribution of worlds consistent with the answer. In order to learn information about their private goal g , the questioner reasons about how an internal model of an answerer would respond given some true world.

- The **questioner** takes a goal $g \in \mathcal{G}$ as input and returns a distribution over questions \mathcal{Q} . It first computes a prior $P_g(w)$ over the features of the world relevant to their goal. For each question $q \in \mathcal{Q}$, it computes the expected information gain, averaged over all possible true worlds, and all possible responses the answerer could give. Information gain is measured as the Kullback-Leibler divergence between the prior distribution $P_g(w)$ and the posterior distribution over world states after hearing the answerer’s (hy-

pothetical) response:

$$P(q|g) \propto \sum_{a \in \mathcal{A}} \sum_{w^* \in \mathcal{W}} P(w^*) P(a|q, w^*) D_{KL}(P_g(w) \| P_g(w|q, a))$$

where $P(a|q, w^*)$ is the response distribution returned by reasoning about what an answerer would say in true world w^* after hearing questioner q . Note that $P(g(w)|q, a)$ is just an ‘interpreter’ function that gives the likelihood of different worlds given different question and answer pairs.

This questioner depends critically on the answer distribution $P(a|q, w^*)$, which will upweight questions that elicit useful answers and downweight questions that elicit vague answers. We now propose several different answerer agents that embody different assumptions about this distribution.

- The **literal answerer** takes a question utterance $q \in \mathcal{Q}$ and a true world state $w \in \mathcal{W}$ as input and returns a distribution over the answer space \mathcal{A} . It samples an answer a from \mathcal{A} with prior probability $P(a)$ and conditions on the likelihood of the questioner inferring the true world w from this answer, using only the interpreter function as a cue. Note that this agent ignores the question; it’s only concern is to convey the true state of the world. This makes it equivalent to the speaker in previous RSA models.

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

- The **explicit answerer** samples an answer a from \mathcal{A} with prior probability $P(a)$, then uses the explicit utterance q as a QUD when evaluating the likelihood of a world under an interpreter.

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

- The **pragmatic answerer** uses the questioner defined above as a generative model of questions given goals in order to estimate the likelihood of different goals given the question q . It samples an answer a from \mathcal{A} with prior probability $P(a)$, uses an internal model of the questioner to estimate the likelihood of different goals $g \in \mathcal{G}$ given their question q , and attempts to be informative with respect to this goal:

$$\begin{aligned} P(a|q, w^*) &\propto P(g|q) \times P_g(w = w^*|q, a) \times P(a) \\ &\propto P(q|g) P(g) \times P_g(w = w^*|q, a) \times P(a) \end{aligned}$$

Finally, we must define the interpreter function that all of these agents are using to compute the likelihood of a world given a question and an answer. This is where all semantic assumptions about the meaning of a question and answer are contained. For the purposes of this paper, we will use the standard Groenendijk & Stokhof semantics (1984), where a question induces a partition \mathcal{P}_q over the space of possible worlds, where each cell is an equivalence class corresponding to a different answer. An answer, then, selects a cell of this partition, denoted by $\mathcal{P}_q(a)$, which is a set.

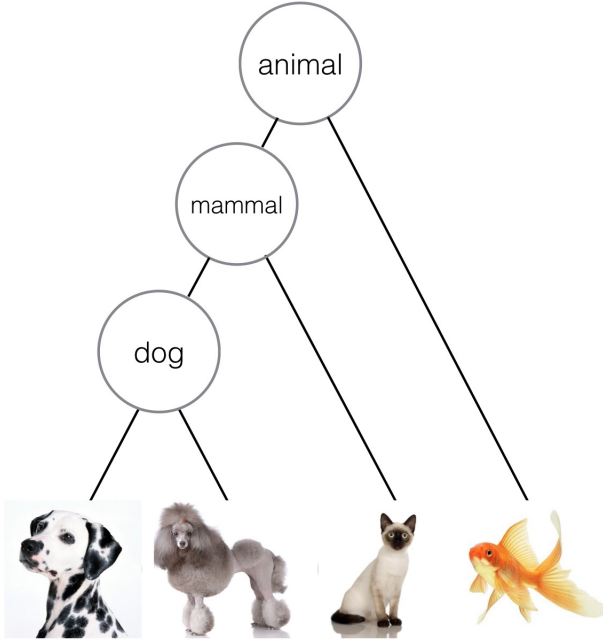


Figure 1: Hierarchy of objects used in experiment 1

- The **interpreter** takes a question q and an answer a and returns a distribution of worlds that are consistent with this pair:

$$P(w|q, a) = P(w) \mathbb{I}_{P_q(a)}(w)$$

where $\mathbb{I}_A(w)$ is the indicator function returning 1 if $w \in A$ and 0 otherwise.

This concludes our specification of the model space. Within this computational framework, different theories of question and answer behavior can be formalized and compared on the basis of their predictions. Assumptions about what is held in common ground are made transparent, and we can systematically manipulate individual elements of the model to test how they affect overall predictions. Because these are probabilistic models, we can succinctly write them down and evaluate their predictions in a probabilistic programming language (Goodman, 2013). The model predictions shown throughout the rest of the paper were computed using a language called WebPPL (Goodman & Stuhlmüller, electronic). Note that this specification diverges from previous work in the RSA framework in a few key ways: (1) for the questioner, we replace the goal of imparting information with the goal of learning information about the specified QUD and (2) ...**XXX there are probably other differences I can't think of Robert.**

Experiment 1:

Questions and answers in a hierarchical world

In order to test how questioners choose questions when given a decision problem, and how answerers choose answers un-

der uncertainty about this decision problem, we designed a guessing-game task played by two players: a guesser and a helper. In this game, 4 animals (a dalmatian, a poodle, a siamese cat, and a goldfish) were hidden behind 4 gates. Note that these animals were arranged in a class hierarchy (see Fig. 1). The guesser was given a private goal of finding one of the objects, and the helper had privileged information about the exact locations of each object. Before choosing a gate, the guesser could ask the helper a single question, and the helper could respond to this question by revealing the object behind a single gate.

In terms of our model specification, the world space \mathcal{W} is the set of $4! = 24$ possible assignments of four objects to four gates and the goal space \mathcal{G} is the set of four objects that the guesser could possibly be trying to find, and the answer space \mathcal{A} is the set of four gates that the helper could possibly reveal. The key constraint in the task, however, is that the guesser must choose from a *restricted set of questions*: they may be trying to find the goldfish, but cannot directly ask ‘where is the goldfish?’ Instead, the question space \mathcal{Q} is the set of highlighted nodes in the hierarchy, including higher order nodes that are consistent with multiple answers. ??

We recruited 25 participants from Amazon Mechanical Turk to participate in this task. Each participant provided responses for four trials in the role of the questioner (corresponding to each goal in \mathcal{G}), and four trials in the role of the answerer (corresponding to each question in \mathcal{Q}). 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, and the order of stimuli within these blocks was also randomized¹. Two participants were excluded due to self-reported confusion about the task.

Results for the guesser role are shown in Figure 2(a). There are two primary trends to note in these data. First, questioners tend to choose the indirect question node closest to their goal when the direct question is unavailable. It is unsurprising that they ask about the ‘dalmatian’ when looking for the dalmatian, but interesting that they strongly prefer asking about a ‘dog’ when looking for the poodle, about a ‘mammal’ when looking for the cat, and so on. **XXX need statistics here. Probably will need more than 25 participants to get enough power... Robert**

Results for the helper role are shown in Figure 2(b). The most striking feature of these data is how closely match the questioner distribution **XXX Possibly just because people did both... It might be the second-timers driving the effect, so we should separate out the different orders, or run it again between-subjects Robert.** If the helper is asked about a ‘dog’, the dalmatian and poodle would be equally good literal answers to this question, but they strongly prefer to give the location of the poodle. Similarly, the dalmatian, poodle, and siamese cat are all mammals, but helpers prefer to respond to the ‘mammal’ question by revealing the location of the cat.

We now compare these results to the predictions of our

¹ All materials are available on Github

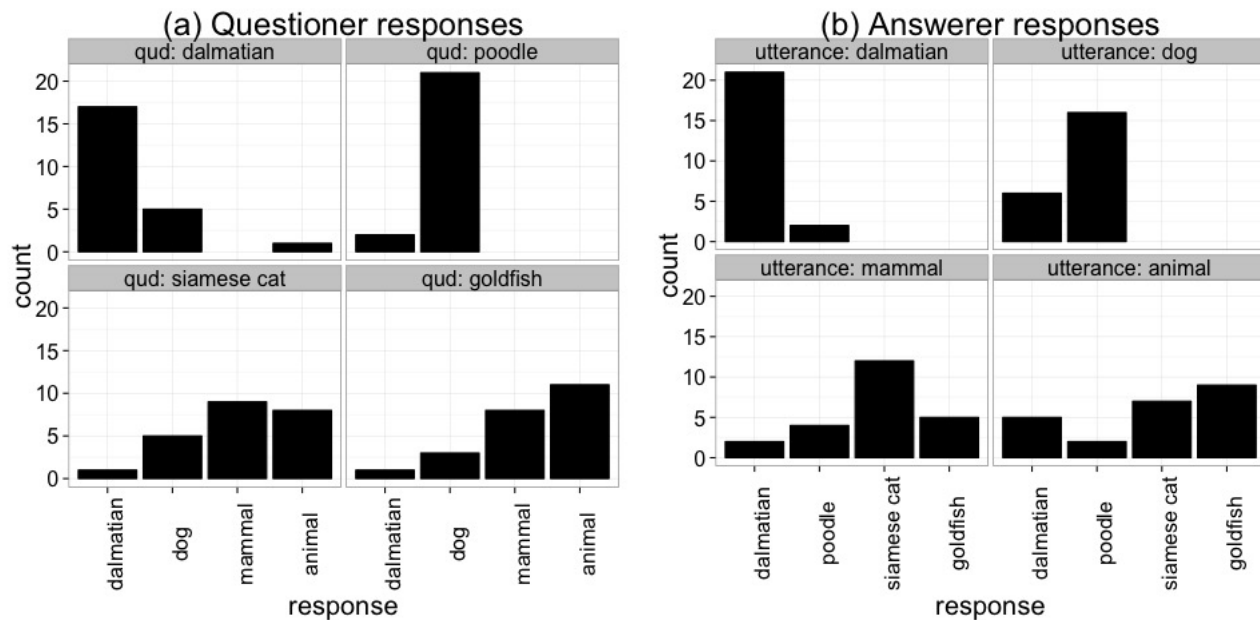


Figure 2: Experiment 1 results

family of models. At the very simplest level, our *literal answerer* yields a uniform distribution over the four answers that are the case in the given world. This has important consequences for the corresponding questioner model: when this questioner reasons about which question would generate the most helpful answer from the literal answerer, they find no difference across questioners, and therefore have no preference over which question to ask. This is clearly not the case and we will not consider these options further.

At a slightly higher level of sophistication, we have an *explicit answerer*. It up-weights answers that result in a distribution of worlds where as many as possible of the leaf nodes picked out by the question are in the correct location. Similarly, it down-weights answers that would mislead the questioner into the thinking that the leaf nodes mentioned in the question are in the wrong locations. This model produces the answerer predictions depicted in Figure 3(b), which successfully down-weights the least relevant alternatives but cannot break the symmetries between the explicitly relevant ones, and is therefore unable to satisfactorily account for our data.

While our ‘explicit answerer’ model cannot account for answerer data, it does provide a foundation for a reasonably well-performing questioner model. When the questioner selects questions that maximize the expected information gain of the explicit answerer’s response, it produces the distribution of questions shown in Figure 3(a) ...**XXX Will fill in once we have better questioner predictions Robert**

Finally, we examine a pragmatic answerer that begins by estimating the likelihood of the questioner having a goal g given the question being asked. It then applies this distribution of goals as the criterion for whether an answer is use-

ful: it prefers answers for which the distribution of consistent worlds have the item specified by the particular goal is in the right position. The results for this answerer are depicted in Figure 3(c).

To address concerns that the preceding results were due to particular features of the design such as the one-to-one mapping from goals to questions and from questions to answers, which gives the task the sense of an ‘elimination game,’ we ran a second experiment using a larger hierarchy ...

Experiment 2:

XXX TODO

Related work

Our account of question and answer behavior ultimately converges on a similar solution as contemporary decision theoretic or game theoretic accounts in linguistics. These theories were a response to early work on question and answer semantics, which focused on the notion of informativeness. In Groenendijk & Stokhof’s (1984) theory of 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 (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

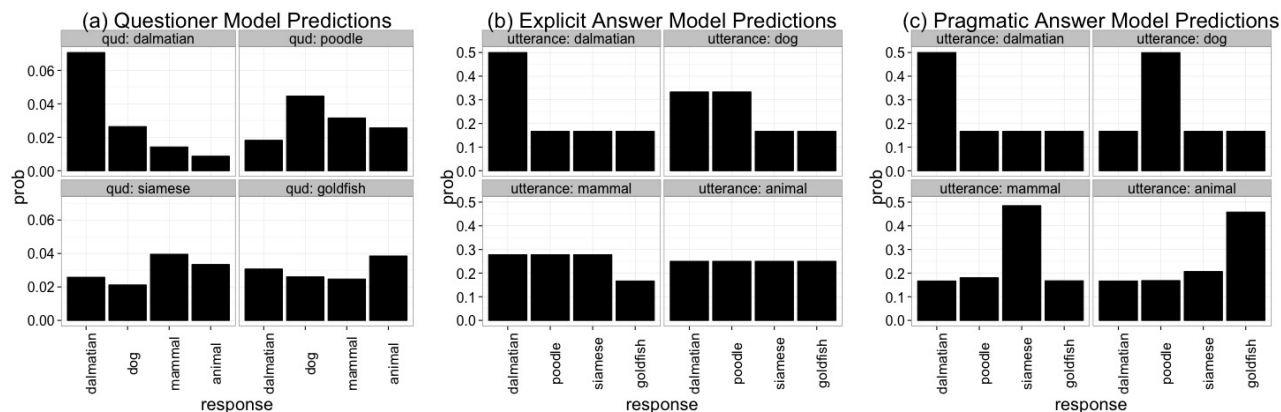


Figure 3: Experiment 1 model predictions XXX Robert: when we've got the models right and have actually fit the rationality parameter to the data, we should put a scatter plot here, or show the data and predictions side by side. This is hard to compare with the results on the previous page.

would suffice. These theories also cannot account for other contextual variation in what counts as a useful answer, such as questions like “where are you?”

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?

General discussion

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 as well. Behind every question lies some 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 an-

swers, even if the explicit question is expressed using the same words.

In this paper we have presented computational-level evidence that answerer behavior is best described by a pragmatic model that reasons about questioner intentions, using the question utterance as a signal. This analysis provides a novel perspective on the role of questions in dialogues: it is well-accepted under the Gricean view that answerers strive to be relevant, but we find that there is also a burden placed on the questioner to provide sufficient information about should be considered relevant in the first place. By artificially restricting the question space in our experiments, we have shown that when necessary, questioners can reliably choose an optimally informative signal.

XXX

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, 998.
- 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. (2013). The principles and practice of probabilistic programming. In *ACM SIGPLAN Notices* (Vol. 48, pp. 399–402).
- 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.
- Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 243–259.
- 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.
- 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.