The pragmatics of plural predication: Homogeneity and Non-Maximality within the Rational Speech Act Model*

Benjamin Spector

Institut Jean Nicod, Département d'études cognitives, ENS, EHESS, PSL Research University, CNRS
Paris, France
benjamin.spector@ens.fr

Abstract

This paper offers an account of two puzzling properties of the intepretation of plural definites, known as *Homogeneity* and *Non-Maximality*, within the framework of the Rational Speech Act model.

1 Introduction

Plural definites (PDs) display two well-known properties: *Non-maximality* and *Homogeneity*.

Non-maximality refers to the fact the the quantificational force of plural definites is variable across contexts. While they tend to have universal quantificational force, they easily 'allow for exceptions'. For instance, if there are many windows in a building and if I was asked to make sure that some fresh air enters the building, (1a) below can be judged true if I opened all of the windows but two or three. If the quantifier *all* is added, as in (1b), this is no longer the case.

(1) a. I opened the windows.
  b. I opened all the windows.

Homogeneity refers to the fact that even in contexts where a sentence such as (1a) has universal or near-universal force, its negation does not amount to the negation of a universally quantified statement. That is, (2a) does not merely mean that I didn't open all of the windows, but rather suggests that I didn't open *any* of the windows (or at most very few). Again, this effect is removed when *all* is added, as in (2b).

(2) a. I didn't open the windows.
  b. I didn't open all the windows.

Križ and Spector offer an account ([8], henceforth KS) of both properties based on two components: a *semantic* component according to which plural predication generates *multiple interpretations*, and a *pragmatic component* that regulates how the resulting underspecified meanings are used and interpreted in different contexts.

The goal of this paper is to reconstruct the pragmatic component within a general framework for pragmatic reasoning, namely the *Rational Speech Act model* ([3]). In so doing, I hope to strengthen the conceptual motivations for our pragmatic component, by showing them to be derivable from general principles of rational conversation, and to make additional predictions regarding how context influences the interpretation of sentences that contain plural definites.

*Thanks to Leon Bergen for useful discussions in relation with this work. The research reported in this work received support from the Agence Nationale de la Recherche (Grants ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL, and ANR-14-CE30-0010- 01 TriLogMean).

2 KS's account

KS's account relies on a specific semantic rule for predicates, and two interpretative principles.

2.1 Semantic component

KS offers a complete compositional semantics whereby, whenever a predicate is applied to a plural object, the resulting denotation consists of a *set of propositions*. Simplifying somewhat, the set of propositions associated with (3a) is given informally in (3b)

(3) Assume that there are exactly three books a, b and c.

a. Mary read the books

b. {Mary read $a \vee b \vee c$, Mary read $a \vee b$, Mary read $a \vee c$, Mary read $b \vee c$, Mary read $a \vee (b \wedge c)$, $\ldots$, Mary read a, Mary read b, Mary read c, Mary read $a \wedge b$, Mary read $a \wedge c$, Mary read $b \wedge c$, Mary read $a \wedge b \wedge c$ }

This set is obtained by 'plugging' in the position of *the books* all the generalized quantifiers that can be obtained by applying disjunction in all possible ways to all the sums that can be obtained from $\{a, b, c\}$. When negation is applied, as in *Mary didn't read the books*, it applies pointwise to all members of the denotation of the negated sentence or predicate, giving rise to the set $\{\neg$(Mary read $a \vee b \vee c$), $\ldots$, $\neg$(Mary read $a \wedge b \wedge c$)$\}$.

The members of the denotation of a given sentence are called its *candidate meanings*.

2.2 Interpretative principles

KS's account relies on two interpretative principles

- Ban on overinformative meanings

The first principle states that, among the candidate meanings for S (i.e. all the propositions in the denotation of S), those that are *overinformative* relative to a given Question Under Discussion (QUD) are ruled out. This is what governs the following definition of *relevant candidate meanings*.

1. **Overinformativity relative to QUD**
Given an underlying QUD Q, modeled as an equivalence relation over possible worlds notated $\sim_Q$, a proposition ϕ is *overinformative* if it makes distinctions between some possible worlds that are equivalent given the QUD. In other words, ϕ is overinformative relative to Q if:
$\exists w_1, w_2 (w_1 \sim_q w_2 \wedge \phi(w_1) \neq \phi(w_2))$ (where $\phi(w)$ denotes the truth-value of ϕ in w).

2. **Relevant candidate meanings given a QUD**
A candidate meaning for S is a *relevant candidate meaning* for S relative to a QUD Q if and only if it is not overinformative relative to Q.

Importantly, the QUD with respect to which an utterance is interpreted need not correspond to an actual interrogative utterance in the prior discourse. Rather, a QUD is simply an equivalence class over worlds that represents what speakers care about at a certain point of a conversation. In fact, there might be uncertainty about what the QUD is, and KS's proposal is based on an idealization. The RSA model that we will develop will not need this idealization.

- Truth on all intepretations.

The second principle states that a sentence S is judged true in the context of a QUD Q only if all its relevant candidate meanings relative to Q are true. This principle entails as a special case the *Stronger Meaning Hypothesis* ([2]). It is possible for both a sentence S and its negation not to be judged true, and so I have implicitly defined a trivalent semantics. In fact, the principle of 'Truth on all interpretations' is similar in spirit to the guiding intuition of supervaluationism.

2.3 Applying of the theory

2.3.1 Capturing homogeneity

Consider the following pair:

(4) a. Mary read the books on her reading list.
b. Mary didn't read the books on her reading list.

Consider first a context where the underlying QUD is something like *Which books did Mary read?*, which corresponds to the equivalence class such that two worlds are equivalent just in case Mary read exactly the same books in both. In this case, none of the candidate meanings for (4a) is overinformative relative to the QUD, because the various candidate meanings differ from each other only with respect to which books Mary needs to have read for them to be true. 'Truth on all readings' then entails that (4a) is judged true if all its candidate meanings are true. One candidate meaning, namely the proposition that Mary read all the books on her reading list, entails all the others, and so (4a) predicted to be judged true if and only if Mary read all the books on her reading list. As to (4b), again no candidate meaning is overinformative relative to the QUD. (4b) is judged true if all candidate meanings are judged true. There is again a candidate meaning that entails all the others, namely the proposition that Mary didn't read any book on the reading list (this proposition is obtained by 'plugging' in the position of *the books on her reading list* the disjunction of all the individual books on the reading list). (4b) is therefore judged true just in case Mary didn't read any books on her reading list.

2.3.2 Non-monotonic contexts

As discussed in KS, 'Truth on all interpretations' captures the intuitive truth-conditions of sentences in which a plural definite is under the scope of a non-monotonic quantifier, as in:

(5) Only one student read the books on the reading list.

In a context where we care about which books each student read, (5) is predicted to be judged true if, for every disjunction of D of pluralities of books on the reading list, only one student read D. These truth-conditions are equivalent to 'there is a student who read all of the books on the reading list and all other students read no books on the reading list' - the conjunction of the two most 'extreme' candidate meanings, namely 'Only one student read all the books' and 'Only one student read at least one of the books'. This appears to be a good result, given, in particular, the experimental results reported in [7].

2.3.3 Capturing non-maximality

Suppose now that Mary has the following reading obligation: 'read at least at least half of the twenty books on her reading list'. Assume further that the underlying question is whether she in fact did that. In such a context, if we learn that Mary read, say, 15 of the 20 books, it seems that we might truthfully utter something like 'Mary read the books on her reading list, so she will probably pass' (in contrast with 'Mary read all the books on her reading list, so she will probably pass'). This connection between non-maximal readings and the underlying QUD, which is substantiated and discussed at length in [10], is captured as follows. Consider (4a), in the context of the QUD 'Did Mary read at least half of the books on her reading list'. First, note that no candidate meaning entails a negative answer to the question (even the weakest candidate meanings are *compatible* with Mary having read all the books). There is one candidate meaning that is equivalent to the positive answer, namely the proposition that Mary read at least half of the books (which is obtained by taking the disjunction of all pluralities of books that contain at least half of the books). This candidate meaning is not overinformative: it is true in all the worlds in which the answer is positive, false in all others. However, all other candidate meanings are either equivalent to this one, or are overinformative, in the sense that they draw distinctions between worlds in which the answer to the underlying polar question is the same. For instance, the proposition that Mary read, say, at least a specific list of 12 books entails that the answer to the question is positive, but is false in some worlds where the answer is positive. As a result, there is only one relevant candidate meaning, namely the proposition that Mary read at least half of the books, and so (4a) is predicted to convey, in this context, that Mary read at least half of the books on her reading list. In practice this might be too precise a prediction, in that what proposition is exactly conveyed might not be completely clear. In our RSA account, this will be explained by the fact that there might be uncertainty about what the QUD is (cf. section 3.5).

The negative case, (4b), works in the same way. Because the property of being overinformative relative to a QUD is closed under negation, there is again only one relevant candidate meaning for (4b) relative to the QUD 'Did Mary read at least half of the books on her reading list', namely the proposition that Mary didn't read half of the books on her reading list, and so (4b) is predicted, in this context, to convey precisely this proposition.

2.3.4 Motivation

KS argues that their interpretative principles are natural principle of language use. Ruling out overinformative candidate meanings is rational because it is assumed that speakers respect Grice's maxim of relevance, and so do not provide *irrelevant information*. A proposition that is overinformative relative to a QUD Q does *more* than just addressing Q, and is in this sense not relevant. As to 'Truth on all interpretations', it can be viewed as a rational solution to a coordination problem. Faced with a sentence that can have multiple interpretations that are all relevant and equally plausible in a given context, one has no way to decide which reading is intended. The speaker who uses such a sentence to convey one of its possible meanings to the exclusion of others cannot be sure that the intended reading will be the one that the listener will converge on, and in fact incurs a serious risk that the listener will end up believing something that she herself does not believe, in violation of Grice's maxim of quality. Only if the speaker believes that all the possible meanings are true is this risk eliminated, and so the sentence will typically be used in such cases. The listener can herself reason that the speaker will use the sentence only in such situations, and so the requirement that the sentence be true on all its relevant interpretations for it to be used becomes a convention between speakers and listeners.

One of my goals here is to provide a formally explicit foundation for this reasoning.

3 The Rational Speech Act account

In this section, I attempt to reconstruct the account I have just presented by building on a game-theoretic model of pragmatic inference and message choice, namely the Rational Speech Act model ([3]). For simplicity, I will now assume that sentences with a plural definite are underspecified between just two candidate meanings, i.e. the one where the plural definite has universal force, and the one where it has existential force.

3.1 The basic RSA model.

In the basic RSA model, we start from a *literal listener* L_0 who has a prior probability distribution over worlds and knows the literal meanings of sentences. When hearing an utterance u, L_0 updates her prior distribution by conditionalizing it with the proposition expressed by the literal meaning of u. Then we define a speaker S_1 who wants to communicate her beliefs to L_0 and knows how L_0 interprets sentences (i.e. knows L_0's prior probability distribution and knows that L_0 interprets messages by conditionalization). S_1 is characterized by a utility function U_1 such that the utility of a message u if S_1 believes w is *increasing* with the probability that L_0 assigns to w after updating her distribution with u, and *decreasing* with the *cost* of u. Importantly, if the literal meaning of u is incompatible with w, the utility of u is infinitely negative. S_1's probability of choosing a message u when she wants to communicate w is defined by a function that is increasing with the utility of u relative to w. A parameter λ determines the extent to which S_1 maximizes her utility. Next, we define a more sophisticated listener, L_1, who, when receiving a message u, uses Bayes's rule to update her prior distribution on worlds, under the assumption that the author of u is S_1. A speaker S_2 is then defined exactly like S_1, except that now S_2 assumes that she talks to L_1, not L_0. And so on:

1. The interpretation function $\mathcal{L}$, when applied to a message u and a world w, returns 1 if u is true in w, 0 otherwise.
2. $L_0(w|u) \propto P(w)\mathcal{L}(u, w)$.
3. $U_{n+1}(u|w) = log(L_n(w|u)) - c(u)$
4. $S_{n+1}(u|w) \propto e^{\lambda U_{n+1}(u|w)}$
5. $L_{n+1}(w|u) \propto P(w)S_{n+1}(u|w)$.

3.2 Truth on all readings.

Assume now that some sentences are ambiguous, i.e. we start with a set of multiple interpretation functions $\mathcal{I}$. There are as many different literal listeners as there are interpretation functions. We now define S_1 as believing that she talks to a literal listener L_0 but as being uncertain about which interpretation function L_0 is using. This uncertainty is represented by a probability distribution over $\mathcal{I}$. The rational utility function for such an S_1 is such that the utility of a message u is the *expected utility of* u across interpretation functions. The next listener, L_1, assumes that she is receiving a message from S_1 and does not need to care about which interpretation S_1 is using, since there is no uncertainty about S_1's strategy. Then for the higher levels nothing changes. This leads to:

1. The interpretation function $\mathcal{L}$, when applied to a message u and a world w, returns 1 if u is true in w, 0 otherwise.

2. Update rule for a literal listener who uses an interpretation function $\mathcal{L}$: $L_0(w|u, \mathcal{L}) \propto P(w)\mathcal{L}(u, w)$.
3. $\mathbf{U_1(u|w)} = \sum_{\mathcal{L}\in\mathcal{I}} \mathbf{P}(\mathcal{L}).\mathbf{U_1(u|w}, \mathcal{L}) = \sum_{\mathcal{L}\in\mathcal{I}} \mathbf{P}(\mathcal{L}_\mathbf{i})[\mathbf{log}(\mathbf{L_0}(\mathbf{w}|\mathbf{u}, \mathcal{L})) - \mathbf{c}(\mathbf{u})]$
4. $S_{n+1}(u|w) \propto e^{\lambda U_{n+1}(u|w)}$
5. $L_{n+1}(w|u) \propto P(w)S_{n+1}(u|w)$.

Suppose now that, for some interpretation function, the literal meaning of a sentence u is false in w. Then at least one term in the sum in Equation 3 is $-\infty$, and as a result the whole sum is $-\infty$. The utility of u for an S_1 who believes w is thus $-\infty$, and so the probability of choosing u to convey w is 0. 'Truth on all interpretations' is thus derived, and, thereby, homogeneity as well as the behavior of plural definites in non-monotonic contexts.

An important remark is in order. This model significantly departs from the use of *lexical uncertainty* in the RSA literature ([9, 1]), where there are as many S_1's as there are interpretation functions, and it's only at L_1 that reasoning about lexical uncertainty takes place. In these models, the first pragmatic speaker can be viewed as picking an interpretation at random, and is thus less sophisticated than in the model we propose here. The first pragmatic listener then performs Bayesian inference jointly about *both* the speaker's beliefs about the world and the particular reading that the speaker picked. For instance, if one of the reading is very implausible due to the prior probability distribution, the first-level pragmatic listener will assign a low probability to the possibility that the speaker picked that reading. This type of model probably captures important facts about how prior probabilities play a role in disambiguation. The current proposal, on the contrary, cannot capture the role played by prior probabilities in disambiguation. In the next section, I introduce a model that includes questions under discussions, and where the listener, at each level, performs Bayesian inference about the identity of the QUD. This might in principle allow for an *indirect* influence of prior probabilities over worlds on disambiguation, in the following way. If the conveyed meaning of a sentence given a certain QUD is very unlikely (given the priors), the listener will infer that the underlying QUD is a different one, which will in turn have an effect on the interpretation of the sentence. I will not investigate, here, however, the extent to which this indirect mechanism can provide a satisfying theory of disambiguation.

3.3 Non-maximality, QUDs and Overinformativity.

In various works in the RSA framework, speaker utility is relativized to a QUD, which helps account for certain cases of non-literal readings ([6, 5, 9]). E.g., a speaker who is answering 'Are some windows open?' and believes that some but not all are, will receive a utility that is not infinitely negative from using the message *All windows are open*, because this message, though false, entails the correct answer to the QUD. Specifically, the utility of a a message u for a speaker who believes that the actual world is w is no longer determined by the probability that the listener will assign to w after interpreting u, but rather by the probability that the listener will assign to $Q(w)$ after interpreting u, where $Q(w)$ is the true answer to Q in w, i.e. $Q(w) = \{v|v \sim_Q w\}$. This move should thus help us account for the context-sensitivity of plural definites. However, it also carries the risk that *all* will too easily receive a non-universal construal, blurring the contrast between plural definites and quantifiers. In the relevant RSA works, there is uncertainty about the QUD (hence a probability distribution over QUDs), on the part of the listener. The speaker takes into account the listener's uncertainty about QUDs, and this restricts, to a certain degree, the extent of non-literal readings. The basic RSA-model with QUDs is as follows. It assumes that the prior probability distributions over worlds and QUDs are independent.[1]

1. The interpretation function $\mathcal{L}$, when applied to a message u and a world w, returns 1 if u is true in w, 0 otherwise.
2. Update rule for a literal listener who uses an interpretation function $\mathcal{L}$: $L_0(w|u, \mathcal{L}) \propto P(w)\mathcal{L}(u, w)$.
3. $U_{n+1}(u|w, Q) = log(L_n(Q(w)|u)) - c(u)$
$= log(\sum_{v \sim_Q w} L_n(w|u)) - c(u)$
4. $S_{n+1}(u|w, Q) \propto e^{\lambda U_{n+1}(u|w,Q)}$
5. $L_{n+1}(w|u) \propto P(w) \sum_Q P(Q).S_{n+1}(u|w, Q)$

Importantly, in this model, the speaker does not care if the listener forms a false belief if this belief is orthogonal to the QUD. Various simulations I have run suggest that this model licenses too easily non-literal readings, in many different cases. For instance, I applied the model to the following case.

1. 3 worlds (no student came, just some of them came, all of them came).
2. Messages: *Some (students came)*, *All*, *Just some*, *No*, *Not all*.
3. Possible QUDs: 'Did some of the students come?', 'Did all of the students come?', 'What's the case?' (= 'Did no, some but not all, or all of the students come?').
4. Message costs: 0 for *Some*, *No*, *All*, 1 for the others.
5. Flat priors on worlds.
6. Priors on QUDs: 0.8 for *Did some of the students come?*, 0.1 for the two others.

With $\lambda = 5$, The first-level listener (i.e. the first pragmatic listener) assigns a probability of about 0.4 to 'some but not all' after having interpreted *all*, and this probability increases with higher recursion-depth, converging to about 0.47. This appears to be a pretty bad result, since it does not seem to be the case that a sentence such as *All the students came* can be interpreted as being compatible with the possibility that some but not all students came, unless maybe the prior probability distribution over worlds makes it extremely unlikely that all the students came (in which case the sentence might receive a kind of hyperbolic interpretation, suggesting that many students came).[2] Increasing the value of λ does not change the general pattern (the probability of 'some but not all' for a listener who hears *All* remains unacceptably high).

Now, where could a pragmatic ban on overinformative sentences come from? Using *All the students came* to answer *Did some of the students come?* might be misleading in suggesting that the QUD one is answering is actually *Did all students come?*. In the above model, the speaker cares about this only instrumentally, i.e. only to the extent that the listener's wrong belief about what the QUD is might prevent her from identifying the intended meaning of the sentence. I propose to modify the QUD-model so that that the speaker cares not only about communicating

[1] The independence assumption is simplistic and in general false, since it seems quite clear that the prior distribution over worlds has an influence on which questions conversational participants might be interested in (cf. [11, 4]).

[2] We also need to take care of the potential confound of domain restrictions, which we can by considering sentences such as 'All the students who take my class came'.

the true answer to the QUD, but also about communicating what QUD she is using. This amounts to replacing Eq. 3. above with $\mathbf{U_{n+1}(u|w,Q) = log(L_n(Q(w),Q|u)) - c(u)}$ (where $L_n(Q(w),Q|u)$ is the joint probability assigned by the level-n listener to the proposition $Q(w)$ and to the QUD being Q, having heard u). The level-$n+1$ speaker is now modeled as wanting to maximize this joint probability. This gives rise to the following revised model:

1. The interpretation function $\mathcal{L}$, when applied to a message u and a world w, returns 1 if u is true in w, 0 otherwise.
2. Update rule for a literal listener who uses an interpretation function $\mathcal{L}$:
$L_0(w,Q|u,\mathcal{L}) \propto P(Q)P(w)\mathcal{L}(u,w)$.
3. $U_{n+1}(u|w,Q) = log(L_n(Q(w),Q|u)) - c(u)$
$= log(\sum_{v \sim_Q w} L_n(w,Q|u)) - c(u)$
4. $S_{n+1}(u|w,Q) \propto e^{\lambda U_{n+1}(u|w,Q)}$
5. $L_{n+1}(w,Q|u) \propto P(w)P(Q).S_{n+1}(u|w,Q)$

Applied to the previous case, with the same parameters, this model yields a reasonable result with enough recursion-depth. The following table shows the probabilities (rounded to 2 decimal placed) assigned by the level-5 listener to worlds (=rows) depending on the message being interpreted (=columns).

	Some	All	Just Some	No	Not All
No	0.00	0.00	0.00	1.00	0.50
Just Some	0.50	0.03	1.00	0.00	0.50
All	0.50	0.97	0.00	0.00	0.00

Even with a strong bias in favor of the question 'Did some of the students come?', the sentence *All the students came* ends up assigning a very high probability to the world where all the students came. With further iterations, or higher values for λ, the probability assigned to the 'some-but-not-all' worlds converges to 0.

This new model explains the ban on overinformative sentence in the following way: overinformative sentences can be misleading, even when they are *literally* true: they can mislead the listener's beliefs about which QUD the speaker is trying to answer.

3.4 Full proposal

Now that we have a mechanism to capture 'Truth on all interpretations' and the ban on overinformative sentences, we can put them together, which gives rise to the following model:

1. $L_0(w,Q|u,\mathcal{L}) \propto P(w)P(Q).\mathcal{L}(u,w)$
2. $U_1(u|w,Q) = \sum_{\mathcal{L} \in \mathcal{I}} P(\mathcal{L})(log(L_0(Q(w),Q|u,\mathcal{L})) - c(u))$
3. For $n \geq 1$, $S_n(u|w,Q) \propto e^{\lambda U_n(u|w,Q)}$
4. For $n \geq 1$, $L_n(w,Q|u) \propto P(w)P(Q)S_n(u|w,Q)$
5. $U_{n+1}(u|w,Q) = log(L_n(Q(w),Q|u)) - c(u)$

To apply the model to the pair in (4), we need to specify alternative messages. These alternative messages play no role for 'Truth on all interpretations', but they are necessary in order to make sure that overinformative meanings are not chosen, since the reason why an informative message is not chosen is that is has lower utility than other messages that would be less misleading regarding the nature of the QUD being answered. We also want to check that while the model allows for a flexible interpretation of plural definites, this is not so for quantifiers such as *some* or *all*. I assume that *the books* is less costly than *some (of the) books* and *all (of the) books* because, on top of being less complex when the quantifiers are partitive (*some of the/all of the*), it also has a less complex semantic type than quantifiers. I applied the above model to the following case:

1. 3 worlds as before
2. Messages: *No*, *Some*, *All*, *Not All*, *The*, *Not The*
3. 2 interpretation functions, depending on whether *The* is existential or universal
4. 3 QUDs: 'Is Some?', 'Is All?', 'What's the case?'
5. Cost: c(*The*) = 0<c(*Not The*)=c(*Some*)=c(*All*)=1<c(*No*)=1.5<c(*Not All*)=2.
6. Flat priors on worlds and interpretation functions.
7. 2 types of priors on QUDs: either 'What's the case' or 'Is Some?' has a probability of 0.8, and the two others 0.1.

With a recursion depth of 5 (but for lower values as well) and $\lambda = 5$, the results appear to closely match intuitions. When 'What's the case?' has a prior probability of 0.8, the level-5 listener interprets messages as follows (values are rounded to the second digit):

	Some	All	No	Not All	The	Not The
No	0.00	0.00	1.00	0.50	0.00	0.98
Just Some	1.00	0.00	0.00	0.50	0.09	0.02
All	0.00	1.00	0.00	0.00	0.91	0.00

When 'Is some?' has probability 0.8, we get:

	Some	All	No	Not All	The	Not The
No	0.00	0.00	1.00	0.50	0.00	1.00
Just Some	1.00	0.00	0.00	0.50	0.50	0.00
All	0.00	1.00	0.00	0.00	0.50	0.00

So we capture the context sensitivity of sentences with plural definites, as well as homogeneity. When the salient question is 'What's the case?', the positive sentence in the pair (4) ('Mary read the books on the reading list') receives a universal interpretation, and its negative counterpart means that Mary read no books on the reading list. When the salient question is 'Did Mary read at least some of the books on the reading list?', the plural definite ends up equivalent to an existential quantifier (both for the positive and the negative sentence).

As to quantified sentences, they are not interpreted in such a flexible way, which was the desired outcome – somewhat surprisingly, *some* ends up meaning *some but not all* even when the salient QUD is 'Is Some?', possibly because, as discussed below, it also raises the posterior probability of the question 'What's the case?'.

3.5 Added Benefits

As an added benefit, this model is able to capture the fact that the interpretation of plural definites might appear somewhat vague. This is the case for instance if the priors over QUDs are flat. Keeping all other parameters constant, the level-5-listener then interprets (4a) as conveying that Mary read some of the books, and probably all of them (with probability 0.7).

This model also captures the intuition that utterances can 'raise new questions'. When the underlying question 'Is Some?' has a high prior probability, the messages *Some*, *All* and *Not All* result in the listener now believing that the speaker is answering another question. This is illustrated by the following table, which represents the posterior probability assigned by the level-5 listener to each QUD depending on which message was received, in the case where the QUD with the highest prior probability (0.8) is 'Is Some?'. Whether these specific predictions are empirically adequate is a matter for further research.

	Some	All	No	Not All	The	Not The
Is Some?	0.00	0.00	0.89	0.00	1.00	0.89
Is All?	0.00	0.50	0.00	1.00	0.00	0.00
What's the case?	1.00	0.50	0.11	0.00	0.00	0.11

References

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

[2] Mary Dalrymple, Makoto Kanazawa, Yookyung Kim, Sam Mchombo, and Stanley Peters. Reciprocal expressions and the concept of reciprocity. *Linguistics and Philosophy*, 21(2):159–210, 1998.

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

[4] Robert Hawkins and Noah Goodman. Questions and answers in dialogue. 2017.

[5] Justine Kao, Leon Bergen, and Noah Goodman. Formalizing the pragmatics of metaphor understanding. In *Proceedings of the Cognitive Science Society*, volume 36, 2014.

[6] Justine T. Kao, Jean Y. Wu, Leon Bergen, and Noah D. Goodman. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences USA*, 111(33):12002–7, 2014.

[7] Manuel Križ and Emmanuel Chemla. Two methods to find truth-value gaps and their application to the projection problem of homogeneity. *Natural Language Semantics*, 23(3):205–248, 2015.

[8] Manuel Križ and Benjamin Spector. Interpreting plural predication: Homogeneity and non-maximality. Submitted, 2017.

[9] Daniel Lassiter and Noah D Goodman. Adjectival vagueness in a bayesian model of interpretation. *Synthese*, pages 1–36, 2015.

[10] Sophia Malamud. The meaning of plural definites: A decision-theoretic approach. *Semantics & Pragmatics*, 5:1–58, 2012.

[11] Robert van Rooij. Questioning to resolve decision problems. *Linguistics and Philosophy*, 26:727–763, 2003.