The Role of Prior Beliefs in The Rational Speech Act Model of Pragmatics: Exhaustivity as a Case Study

Ethan Wilcox, Benjamin Spector

The Role of Prior Beliefs in The Rational Speech Act Model of Pragmatics: Exhaustivity as a Case Study

Ethan Wilcox[1] and **Benjamin Spector**[2]

[1]Department of Linguistics, Harvard University, wilcoxeg@g.harvard.edu
[2]Institut Jean Nicod, Département d'études cognitives, ENS, EHESS, PSL University, CNRS , benjamin.spector@ens.fr

Abstract

This paper examines the interaction between prior beliefs and pragmatic inferences, focusing on exhaustivity effects. We present three experiments that tests how prior beliefs influence both interpretation and production of language, and compare the results with the predictions of the Rational Speech Act model, a Bayesian model of linguistic interpretation. We find that prior beliefs about conditional probabilities have no affect on language production, but do affect interpretation, producing *anti-exhaustivity* effects. We find that the RSA model achieves a relatively good fit both for the human production and interpretation data, but only for highly-implausible utterance costs.

Keywords: Pragmatics, Rational Speech Act model, Exhaustivity.

Introduction

The interpretation of linguistic utterances in context depends on the prior beliefs of speakers and hearers. For instance, if someone says "Mary visited a cardiologist today", one will infer that Mary is more likely than a random person to have a heart-related medical condition. It is easy to account for such inferences as a probabilistic inference: the hearer starts with a prior probability distribution over possible world states, and then conditionalizes this distribution with the new information that Mary visited a cardiologist today. Typically, though, pragmatic inferences go well beyond what can be predicted with such a simple model of linguistic interpretation. They also involve, for instance, reasoning about *other sentences* that the speaker could have uttered given some assumptions about their communicative goals (Grice, 1975). For instance, if I'm asked a question such as *Among Peter, Mary and Sue, who attended the show today?*, an answer such as *Mary did* tends to trigger the inference that the others did not, even if there is no expectation that what any of them does depends on what the others do. Such *exhaustivity effects* are typically accounted for in terms of Grice's maxim of quantity: if in fact both Peter and Mary had attended the show, a knowledgeable speaker would say that rather than just talking about Mary.

Now, in some situations these two types of effects (effect of prior beliefs, exhaustivity effects) are pitted against each other. For instance, we might know that Peter and Mary are a couple, who usually go out together, so that, upon learning that Mary attended the show, one would assign a high probability to the possibility that Peter did too, which would go against the exhaustivity effect just mentioned.

The Rational Speech Act Model (RSA) is a model of pragmatic reasoning which integrates both the role of prior beliefs and that of pragmatic reasoning about alternative utterances (Frank & Goodman, 2012). It can in principle make very precise predictions about their interactions. The RSA model views the speaker as being engaged in a trade-off between two goals: maximizing informational content and minimizing the cognitive cost of an utterance. As we will see shortly, in the baseline RSA model, this trade-off is affected in a drastic way by the prior beliefs shared by listeners and speakers, to the extent that, in some situations, an *anti-exhaustivity effect* is predicted: in some cases, the utterance *Mary did*, in the above context, is expected to be the best message to use to convey that both Mary and Peter attended the show, and thus to be interpreted in this way. However, because an RSA model has several free parameters, it is difficult to assess a) whether it is compatible with a given set of data, and b) whether it provides an *explanatory* account of the data. The goal of this paper is to gather data about the effect of priors on exhaustivity effects, both for interpretation and production, and to assess how well the baseline RSA model can account for these data in a principled way.

Degen et al. have already tested the effect of priors on pragmatic interpretation within the RSA framework, focusing on a similar but different type of inference, namely the inference from *some* to *not all* (Degen, Tessler, & Goodman, 2015). We will discuss the relationship between our study and Degen et al. (2015) in the next section.

The Rational Speech Act Framework and exhaustivity effects

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. 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. A *rationality parameter* α 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].

[1]See Bergen, Levy, and Goodman (2016) for the mathematical description of the model.

Now consider a case where world states are individuated by the truth-values of two propositions A and B (for instance *Mary attended* and *Peter attended*), and where the available utterances are A, B, A *and not* B, B *and not* B and A *and* B. Consider a situation where the speaker wants to communicate the world state $\{A\}$ (where A is true and B is false) . She can choose between the two messages A and A *and not* B. While A is less informative than A *and not* B, it has nevertheless a significant probability of use, because it is less costly. Upon hearing A, the first-level pragmatic listener L_1 will reason as follows, if the priors are sufficiently uniform across world states. The message is only compatible with two world states, namely $\{A\}$ And $\{A,B\}$. But the the speaker is more likely to mean $\{A\}$ than to mean $\{A,B\}$. If she wanted to communicate $\{A,B\}$, there were two other possible messages, namely B and A *and* B. B is furthermore no more costly than A, and A *and* B is costly but also more informative. In contrast with this, if she wanted to communicate $\{A\}$, there was only one other possible message, namely A *and not* B, and furthermore this message, while more informative, is very costly (more than A *and* B). As a result, it is likely that the intended meaning was in fact $\{A\}$, and the exhaustivity effect is derived.

However, things can change drastically with non-uniform priors. Imagine now a speaker who wants to communicate $\{A,B\}$. She has a choice between using the messages A, B, A *and* B. While the latter message is the most informative, it is also more costly than the two others. Suppose further that the prior conditional probability of B given A is very high. The literal listener L_0, upon hearing A, will assign a high probability to the world state $\{A,B\}$. In this case, A may turn out to have a higher utility than A *and* B for S_1: it is quite good at communicating the world state $\{A,B\}$ (given the priors), and it is less costly than A *and* B. Furthermore, with such non-uniform priors, a speaker who would want to communicate $\{A\}$ might be very unlikely to use the message A: despite the fact that A is less costly than A *and not* B, it is so poor at conveying the intended world state (due to the priors), that the speaker now has an extra incentive to use the costly sentence A *and not* B. Now, upon hearing A, the pragmatic listener L_1 will reason as follows. The intended world state is either $\{A\}$ or $\{A,B\}$. If the latter, S_1 was in fact quite likely to use A. If the former, the speaker was more likely to use A *and not* B. So the intended world state is probably $\{A,B\}$. This time an *anti-exhaustivity effect* is derived (Roni Katzir, p.c.). However, this prediction is highly sensitive to the values of the free parameters of the model (rationality, costs).

Degen et al. (2015) discuss a related case. The RSA model, under a broad range of values for the free parameters, predicts that when the conditional probability of an *all*-statement given the truth of the corresponding *some*-statement is very high, *some* is going to be used to convey *all* and to be so understood. Degen et al. consider a discourse such as: *Max threw fifteen marbles in the water. Some of the marbles sank*. Because we expect all marbles to sink, this is a case where the prior probability of $\forall$ (the world where all marble sank) is very high, and where the basic RSA model predicts that the sentence will in fact convey that all marbles sank. But the experimental results show that actual listeners typically derive a *some but not all*-reading. In Degen et al.'s model, unlike in the basic RSA model, the pragmatic listener is uncertain about the speaker's beliefs about the listener's priors. Even if $\forall$ has a very high prior probability for the listener, the pragmatic listener L_1 assigns a substantial probability to the possibility that the speaker believes that the literal listener L_0 is in fact entertaining uniform priors over world states. So the pragmatic listener L_1 has a higher-order prior probability distribution over the set of first-order prior distributions (over world-states) that the speaker might attribute to the literal listener L_0. When processing a sentence, this listener updates both her probability distribution over worlds and her higher-order probability distribution over the set of priors that the speaker is considering. The proposed model is such that when hearing *some*, the listener concludes that the speaker probably believes that the listener is using uniform priors, and as a result *some* ends up conveying $\exists\neg\forall$. Simulations show that in order to obtain this result, the pragmatic listener L_1 must view the speaker (S_1) as believing that there is a high probability that the literal listener's prior distribution over world states is uniform. For the range of values that are typically used in RSA models for α (somewhere between 1 and 10), this probability must be substantial (Degen et al. report that it has to be equal to .5 to achieve the best fit with experimental data). Given this, a conceptual limitation of this account is that it models the listener as believing that the speaker views the listener as likely to be unaware that marbles typically sink when thrown into water (despite the fact that the priors over world states that Degen et al. collected show that people do in fact expect that when marbles are thrown into water, they will all sink). But no empirical evidence is provided to support these assumptions, and so it is not clear that much is gained compared to a model that would simply ignore the actual priors and take as input relatively uniform priors.

Now, in the case of exhaustivity effects, the situation is even more extreme. In the *some-all* case, the *all* sentence is no more costly than the *some*-sentence. Because of this, even with extremely biased priors, a fully rational speaker would always choose *all* to convey *all*, since it is still more informative than *some*, and would never use *some* (*some but not all* would be used to convey $\exists\neg\forall$). For this reason, with very high values for α (corresponding to a very rational speaker), a correct result is derived in Degen et al.'s model, even if the probability that the speaker assigns to the possibility that the listener does not expect all marbles to sink is very low (but still positive). In the exhaustivity case, avoiding the anti-exhaustivity effect is harder, because the message A *and* B is more costly than the message A, and so will not necessarily be the message used by a fully rational speaker who believes A *and* B, if the prior conditional probability of B given A is very high (the gain in informativity provided by A *and* B compared to A might be too small to justify the extra cost). Even with

a fully rational speaker, for a broad range of reasonable cost values, there exist contexts where the speaker is predicted to use *A* to mean *A and B*. In this paper, we will compare the predictions of the baseline RSA model with experimental data pertaining to exhaustivity and anti-exhaustivity effects.

Independently of this theoretical goal, our contribution is to provide experimental data pertaining to cases where priors are biased against the exhaustive reading of a sentence *A* in the context of *Which of A and B is true?*.

Human Judgement Experiments

To test the effect of priors on human linguistic judgements, we conducted three online experiments. Each experiment involved a simple scenario in which a character was moving furniture from her apartment onto the street, and questions were asked about what the character was able to move or how she was likely to report the progress of her moving to a friend. Experiments were hosted on IbexFarm. Participants were recruited on Amazon Mechanical Turk.[2]

Experiment 1: Priors

As this work aims to test the effect of priors on human linguistic judgements, our first experiment gathered prior probabilities for two scenarios, which were used in later experiments. In the priors experiment subjects were shown a scenario in which a character is moving her apartment and tests the weight of two furniture items. The character picks up one item, at which point respondents were asked whether they thought she could pick up the second item as well. Input format were forced-choice, yes/no radio buttons. The experiment was divided into two conditions: In the first, High Conditional Probability condition, the character was shown picking up a chair and asked whether she could also pick up a footstool, which was visually about half the size. In the second, Low Conditional Probability condition participants saw the character picking up the footstool and asked if they thought she could also pick up the chair. Participants were asked two simple comprehension questions at the end of the experiment, and only responses from participants who answered both correctly were used. We collected 60 responses, of which 57 (95%) were usable. The proportion of respondents who selected yes in each condition was taken as the population-level prior on conditional probability in each case.

The results can be seen in Fig. 1, on the left-hand panel. Error bars represent binomial 95% confidence intervals using the binconf function in R on default settings (Wilson method). A Fishers Exact Test indicates that participants were significantly less likely to endorse the yes response in the Low Conditional Probability condition (p=0.02225).

Experiment 2: Elicitation

We conducted a second experiment to test the effect of priors on the elicitation of simple conjunctives. If humans subjects incorporate priors in their utterance and endorsement of simple conjunctives, then we expect the relative rate of the conjoined utterance ("A and B") to be lower in high-conditional probability contexts, where P(B‖A) is very high (because in this case the utterance A is quite good at communicating the $A \wedge B$ world state (which we will denote by $\{A,B\}$ henceforth). Furthermore, we also expect that, if they want to communicate the world $A \wedge \neg B$ (which we will now notate$\{A\}$), there will be less likely to use the message A in the high-probability condition, and more likely to use *A and not B*.

In this setup, participants were shown the same 'moving' scenario from the previous experiment, involving a chair, a footstool and a character who tells a friend that she would move 'everything I can' down to the curb. In the subsequent panel participants were shown the character with the furniture she was able to move depending on the condition to which the participant was assigned, which are enumerated in Table 1, along with the condition name and a *tag*, with which we refer to the condition in charts and figures. Participants are asked to endorse an utterance that they think the character would use to describe the situation to a friend, who has prior familiarity with the items, over the telephone. Input were force-choice radio buttons with six possible utterances: 'I moved the chair', 'I moved the footstool', 'I moved the chair but not the footstool', 'I moved the footstool but not the chair', 'I moved the chair and the footstool' and ' I moved the footstool and the chair.'

Experimental Stimuli	Tag	Condition Name
Chair + Footstool	{A,B}	[BOTH, HIGH PROB]
Chair	{A}	[SINGLE, HIGH PROB]
Footstool + Chair	{A,B}	[BOTH, LOW PROB]
Footstool	{A}	[SINGLE, LOW PROB]

Table 1: Elicitation Experimental Conditions

Following the critical question, we asked two simple comprehension questions and whether the participant was a native speaker of English. Only data from those respondents who answered both correctly and identified as a native English speaker were used. The experiment was given to 174 subjects, of which 126 (72.4%) answered the follow-up questions satisfactorily. A further 33 subjects were filtered as repeat subjects from one of our other experiments, bringing the total number of responses to 93.

The results from this experiment can be seen in Figure 1, in the middle panel, with world state on the x-axis and the proportion of "A and B" responses on the y-axis. Red dots represent proportion of "A and B" responses in the high probability condition, blue dots the low probability condition; error bars represent 95% confidence intervals. Endorsements of the "A and B" utterance were near floor in the $\{A\}$ world (m=0.02, m=0.11 in the High Probability and Low Probability conditions, respectively). However, the endorsements were not at ceiling in the $\{A,B\}$ world state (m=0.68, m=0.84 in the High Prob and Low Prob conditions, respectively).

[2]Experiments were pre-registered online at http://aspredicted.org/blind.php?x=7qm9pz

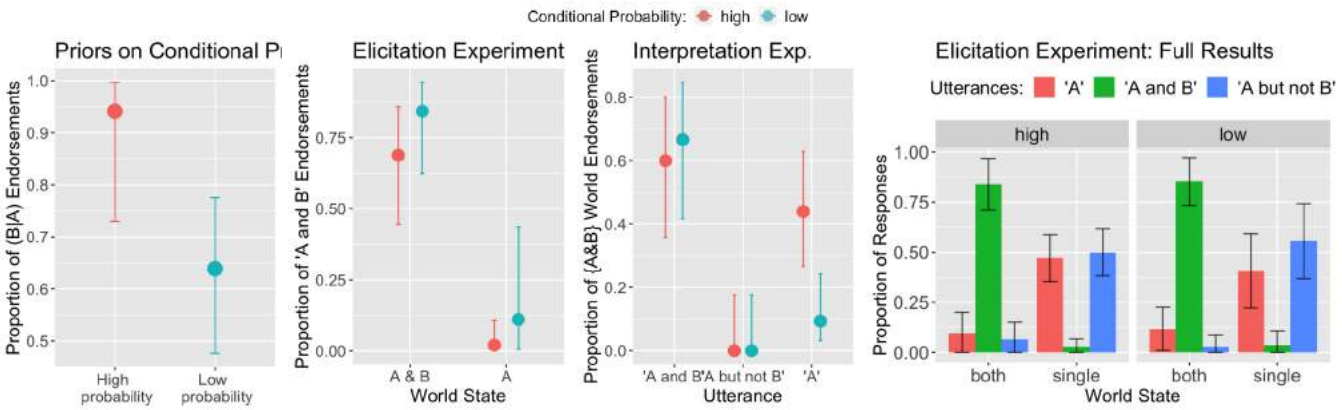

Figure 1: Human Judgements from the Online Study

To test whether priors on conditional probability had an effect on utterance endorsement, we fit a linear model to the data using the proportion of 'A and B' responses as our dependent measure, and experimental conditions as predictors, which were coded using 1/-1. We found a main effect of WORLD STATE, whereby participants were less likely to endorse "A and B" in the $\{A\}$ world ($p<0.001$), as expected. However, we found no interaction between world state and prior conditional probability ($p=0.507$), which is visually evident from the fact that both prior probability conditions fall within each other's confidence intervals. In fact, the relative rate of "A and B" endorsement in the high and low conditional probability conditions, ran counter to our expectations, with respondents marginally less likely to endorse the relevant utterance in the high conditional probability condition. Note that while the subjects in this study were willing to use 'A' to endorse the {A,B} world, their rates of endorsement in both conditions (between 15-32%) were well below their expectation of P({A,B}|{A}) (between 65-95%).

In addition to our pre-registered analyses, we conducted a follow-up analysis to assess whether the priors on conditional probability affected the rate of endorsements for the exhaustified utterances, 'A but not B.', in order to communicate the world $\{A\}$ The results for all utterance endorsements can be seen in Fig. 1, on the far right panel. As the conditional probability increases, we might expect the rate of endorsements for the exhaustified utterance (the blue bar) to increase in the SINGLE condition, given that the bare utterance, 'A' might be quite bad at communicating $A \wedge \neg B$. Our pre-registered analysis, which examines only the rate of endorsement for the 'A and B' utterance, would not capture these dynamics.

In order to assess the impact of conditional probability priors on the rate of the exhaustified utterance we fit a linear regression model using the proportion of exhaustified ('A but not B') utterance endorsements as our dependent variable and utterance types as our predictors. We found a main effect of world state ($p < 0.001$) whereby exhaustified utterances were less likely to be endorsed in the BOTH condition (as fully expected), but no significant interaction between world state and conditional probability ($p = 0.515$).

Experiment 3: Interpretation

Experimental Stimuli	Tag	Condition Name
"The chair and the footstool"	'A and B'	[BOTH, HIGH PROB]
"The footstool and the chair"	'A and B'	[BOTH, LOW PROB]
'The chair but not the footstool"	'A but not B'	[ONLY, HIGH PROB]
"The footstool but not the chair"	'A but not B'	[ONLY, LOW PROB]
"The chair"	'A	[SINGLE, HIGH PROB]
"The footstool	'A'	[SINGLE, LOW PROB]

Table 2: Interpretation Experimental Items

The third experiment aimed to test the effects of prior conditional probability on utterance interpretation. The RSA model predicts that human subjects will be more likely to interpret the utterance 'A' as referring to an {A,B} world in cases where P(B|A) is higher.

For this experiment, participants were shown the same 'moving' scene as in the others. A character commits to moving 'what I can lift' down to the curb, and tells a friend what she is capable of lifting depending on the condition to which the subject was assigned. There were six conditions, corresponding to six possible utterances: 'I can lift the chair and the footstool','I can lift the footstool and the chair', 'I can lift the chair but not the footstool', 'I can lift the footstool but not the chair', 'I can lift the chair', 'I can lift the footstool'(cf. Table 2).

In the subsequent slide, participants see the character by the curb, with a grayed-out area where the furniture would be, are told that the character 'has moved all the items she can lift down the curb', and are asked to select which items they believe have been moved down. They are provided with a visual reference of the furniture items, scaled to size, at the bottom of the screen. The input form was a check box, and

in the instructions to the experiment participants were told that they could check as many or as few of the boxes as they wished.

Following the critical question, participants were asked two comprehension questions and whether or not they were a native speaker of English. The survey was given to 475 participants of which 338 (71%) answered the follow-up questions satisfactorily. Another 77 were filtered out, as they were repeat responders from the previous experiment, leaving the total number of responses analyzed to 261.

The results from this experiment can be seen in 1, on the right-hand panel. The utterance types are on the x-axis, with the proportion of respondents who checked both boxes (thereby endorsing the {A,B} world) on the y-axis. Red dots indicate responses for the high conditional probability condition, blue for the low conditional probability condition. Error bars indicate 95% confidence intervals. The proportion of {A,B} world endorsements is at floor when respondents heard the "A but not B" utterance, as predicted. However, when respondents heard the "A and B" utterance, endorsements of the {A,B} were relatively low (m=0.6, m=0.66 in the high and low probability conditions, respectively). This means that when respondents read "I will move what I can lift down to the curb" followed "I can lift the chair and the footstool down", and are then told that the character moved all the furniture he could lift, they are willing to endorse a world where only one had been moved (the footstool in 72% of the cases). We believe this behavior is partly due to the experimental setup: subjects may expect the character to do as little work as possible without the help of her friend, who they were told would assist in the moving process later on. We had initially thought that the commitment to 'move what I can lift' would ensure that modalized sentence of the form 'I can do X' would be interpreted as implying that the character did X, but this result suggests that this was not always the case.

To test whether the conditional probability had an effect on the rate of {A,B} world endorsements, we fit a linear regression model using experimental conditions as predictors. We found a significant main effect of ONLY utterances and SINGLE utterances ($p<0.001$ for both), whereby subjects were less likely to endorse the {A,B} world for these two conditions. In addition, we found an interaction between the prior probability and the SINGLE utterance types ($p=0.0144$), whereby participants were more likely to endorse the {A,B} world in the high conditional probability after hearing the non exhaustified utterance. Overall these results indicate that gradient prior probabilities gradiently affect utterance interpretation, raising the question why we did not observe a similar gradience in the elicitation experiment.

Model Fit

We fit the vanilla Recursive Speech act Model presented in (Frank & Goodman, 2012) to the human data we collected, with one level of recursion depth (that is, we fit S_1 and L_1).

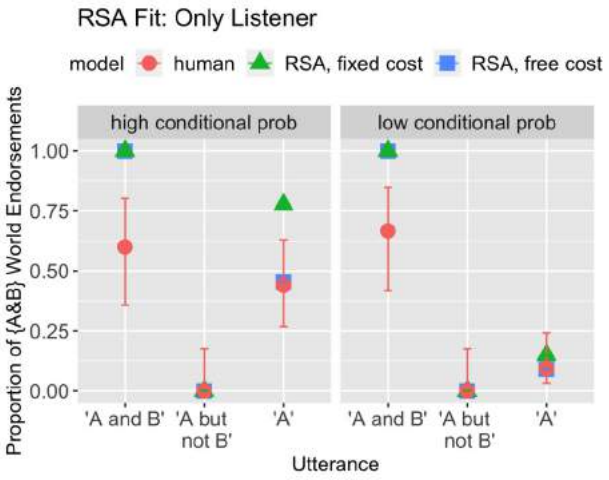

Figure 2: RSA Model fit with fixed cost (green triangles) and free cost ratios (blue squares) to human judgements (red circles).

The model has three possible world states: {A}, {B} and {A,B}. World {A} had a prior of 0.32, world {B} had a prior of 0.14 and world {A,B} had a prior of 0.54, rendering P({A,B} | {A}) = 0.8, close to the human *high conditional probability* prior, and P({A,B} | {B}) = 0.63, close to the human *low conditional probability* prior. The model includes seven messages: 'A', 'B', 'A but not B', 'B but not A', 'A and B' and 'null', which is defined as true in every situation, and was assigned a fixed cost of 100. 'A' and 'B' were assigned a cost of 0.[3] The costs of 'A and B' (c_1) and 'A but not B' (c_2) were free parameters, as was the rationality parameter, α.

In order to assess how well the RSA model captured the human judgements, we conducted four fits, which are summarized in Table 3. Each fit was made by iterating through a wide range of alphas and cost parameters (0-20 for each). This technique guarantees that we found a locally optimal fit within the range of cost and optimally parameters typically seen in the rest of the Recursive Speech Act literature (Scontras, Tessler, & Franke, 2017). In the fixed cost ratio fits, the cost for "A but not B" must be greater than but could not be more than 2 times that of "A and B". This constraint makes sense if we view cost as reflecting, for instance, the number of logical operators in a sentence, or the number of words used. Thus, we wanted to see if an fit existed with cognitively plausible relative costs between these two types of utterances. But we also relaxed this constraint in the 'free cost' fit, where the only constraint that the the cost of "A but not B" is higher than that of "A and B".

The results for the listener-only fit can be seen in Figure 2. Here, the x-axis is the possible utterances, and the x-axis is the proportion of respondents who endorsed the {A,B} world (in the human case) or the posterior distribution on the {A,B} world (in the model case). The left panel represents the high

[3] In the RSA model, it is the *difference* between relative costs that matters: adding a fixed constant to each cost value has no effect.

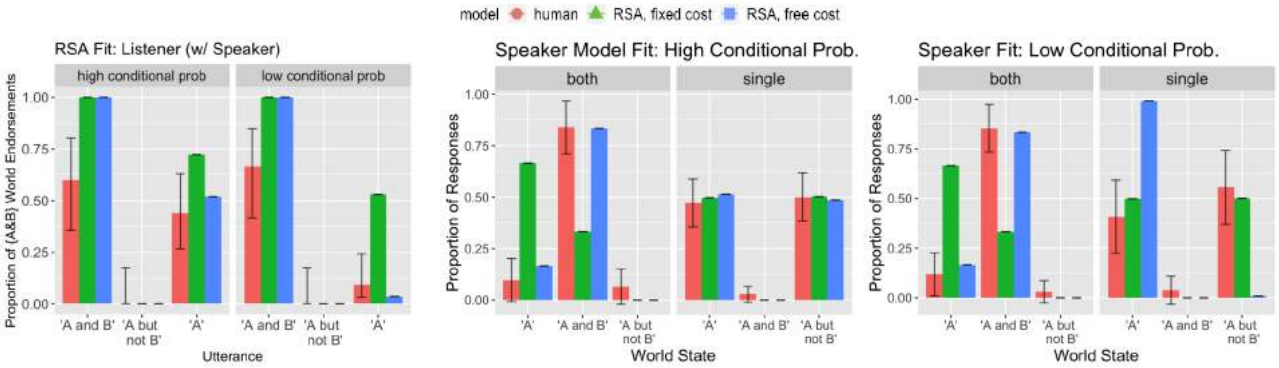

Figure 3: RSA Model fit with fixed-cost ratios (green) and free cost ratios (blue) to human judgements (red).

Name	Layers	Restrictions	α	c_1	c_2	MSE
Fit 1	L_1	$c_2 < 2 * c_1$	9.19	1.57	3.17	0.097
Fit 2	L_1	$c_1 < c_2$	5.52	0.010	8.42	0.068
Fit 3	L_1, S_1	$c_2 < 2 * c_1$	0.01	0.27	0.54	0.092
Fit 4	L_1, S_1	$c_1 < c_2$	7.97	0.01	1.59	0.059

Table 3: Summary of The Four Fits and Optimal Parameters

conditional probability condition and the right panel the low conditional probability condition. The vanilla RSA's meaning function constrains the listener's posterior for utterances 'A and B' and 'A but not B' such that all probability is assigned to the {A,B} world and the {A} world, respectively. Therefore, it is entirely the model's posterior on the 'A' utterance that determines the relative goodness of the fit. For the restricted cost ratio fit (green triangles) the best model is able to match human behavior in the low conditional probability condition, but favors the {A,B} much more greatly than do human respondents in the high conditional probability condition (the green triangle is well above the red error bars), resulting in a mean squared error of 0.097. When the restriction on relative costs is relaxed (blue squares) the model is able to achieve a very precise fit, with a mean squared error of 0.0678. The reason why the free-cost fit is able to perform significantly better than the fixed cost fit is that it can assign much higher relative cost to "A but not B" than to "A and B".

For example, in Fit 2 the utterance "A but not B" is 840 times more costly than the utterance "A and B". This results in strong model performance because high relative cost of the exhaustified utterance counterbalances its informativity at communicating the $\{A\}$ world. This renders the 'A' utterance a good choice to communicate the $\{A\}$ world, despite the strong priors on the $\{A, B\}$. Furthermore, the low cost of "A and B" ensures that it will be chosen often in the $\{A, B\}$ world, even in the high-probability condition.

The results for Fits 3 and 4, which fit both the speaker and listener layers, can be seen in Figure 3, with the listener layer graphed at left and the speaker layer graphed in the center and right images. For the listener layer, the x-axis shows utterances, and the y-axis posterior probability endorsements for the {A,B} world. For the speaker layer, the facets represent the different worlds conditions and the x-axis shows the possible utterances, with the relative proportion assigned to each utterance (for the RSA models) or proportion of endorsements (for the human) on the y-axis.

As to performance of the model: in the restricted cost ratio fit (the green bars) the performs only moderately well. For the ten critical conditions where the posterior distributions are not constrained to either 100% of 0%, the best fit falls outside of the human judgements' 95% confidence intervals 6 times, resulting in a mean squared error of 0.092. For the free cost model (blue bars) the model is able to perform slightly better, falling outside of the human judgements' 95% confidence intervals only twice (both in the {A} world, low probability condition). This fit gives an MSE of 0.059. Two remarks are in order. First, in the free cost model, "A but not B" is 158 times more costly than the utterance "A and B". Second, the best model achieves a good fit for the listener and for the speaker in the high-probability condition, but drastically underestimates the rate of endorsement of "A but not B" in the low probability condition as a way to express the $\{A\}$-world.

Discussion

The results of the interpretation experiment establishes that prior probabilities modulate exhaustivity effects, as is expected under the RSA approach. In our data, they do so for interpretation, but not for production. The RSA model can achieve a good fit with our experimental data for the interpretation experiment only with implausible parameters. With the kind of cost values that are typically assumed (cf. fixed cost fit), it overestimates the effect of prior probabilities. When we relax constraints on costs, an excellent fit is achieved, but the cost of "A but not B" has to be 832 times that of "A and B". When we want to fit both interpretation and production, the best model drastically underestimates the use of sentences such as "A but not B" - precisely because it assigns it an extremely prohibitive cost. Note that we are only evaluating the baseline RSA model. More sophisticated models have been proposed within the RSA framework, and we are not evaluating those. What our results suggest is that a key ingredient of the baseline RSA model, namely the tradeoff between infor-

mativity and cost, which predicts a huge influence of priors on interpretation and production, might make it hard to capture both interpretation and production data. On the interpretation side, the model needs to assign a very high cost to *A but not B*, but then on the production side, the model predicts that *A but not B* is not usable.

That being said, this conclusion is provisional, as caution is in order when interpreting the results we present here. We only tested two different conditions, in one type of scenario, and the data are somewhat noisy (cf. the high rate of rejection of $\{A, B\}$ after hearing "I can do A and B"). The fact that we used modal sentences when we collected priors and in the interpretation task is a limitation of this study.[4] Future work is needed to a) gather additional and less noisy data so as to reach more reliable conclusions, b) construct alternative models, including refined versions of the baseline RSA model, which could then be compared to it.

Acknowledgments

B.S. would like to thank Leon Bergen, Danny Fox and Roni Katzir for relevant discussions, and acknowledges funding from a grant from the Agence Nationale de la Recherche (ANR-17-EURE-0017). E.G.W. would like would like to acknowledge support from the Mind Brain Behavior Interfaculty Initiative Graduate Student Grant.

References

Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*.

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Cogsci.*

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

Scontras, G., Tessler, M. H., & Franke, M. (2017). *Probabilistic language understanding: An introduction to the rational speech act framework* (Tech. Rep.). Retrieved 2018-1-9 from https://problang.org.

[4] In another experiment that we do not present here, we collected judgments for non-modal sentences in near-identical scenarios, and we got a much higher endorsement (85%) for the $\{A, B\}$ world for the conjunctive sentence *A and B*.