Modeling sentence polarity asymmetries: Fuzzy interpretations in a possibly wonky world

Muxuan He Elsi Kaiser Khalil Iskarous
University of Southern California
{muxuanhe, emkaiser, kiskarou}@usc.edu

Abstract

Negation is an important aspect of human language and reasoning. Prior work has proposed that positive- and negative-polarity sentences exhibit a number of asymmetries. This paper focuses on two of them: (i) Regarding cost, marked forms like negation are known to elicit more production cost than the unmarked positive polarity, and (ii) regarding pragmatic inference, the negative polarity is said to presuppose the prominence of its positive-polarity counterpart, but not the other way around. We present novel empirical evidence regarding these two asymmetries and offer one of the first formalizations of these asymmetries within the Rational Speech Act (RSA) framework. We show that existing extensions of the standard RSA model, e.g., soft semantics and common ground update, while not originally proposed to address sentence polarity asymmetries, can nonetheless be applicable to these phenomena.

1 Introduction

As one of the most influential cognitive models of pragmatics, the Rational Speech Act model (RSA; Frank and Goodman, 2012) formalizes the recursive reasoning involved in language use and communication. See formulas (1) - (4) for a formal definition of the standard RSA model:

$$P_{L0}(s|u) \propto [\![u]\!](s) \cdot P(s) \quad (1)$$

$$[\![u]\!](s) \in \{0, 1\} \quad (2)$$

$$P_{S1}(u|s) \propto exp(\alpha\ (\ln P_{L0}(s|u) - Cost(u)) \quad (3)$$

$$P_{L1}(s|u) \propto P_{S1}(u|s)\ \cdot P(s) \quad (4)$$

This model centers on a pragmatic listener, $P_{L1}(s|u)$, who infers the intended state s from an utterance u by reasoning about a pragmatic speaker, $P_{S1}(u|s)$, who selects utterances based on their utility U. This speaker derives informativeness (how much an utterance reduces uncertainty about the intended meaning or referent) from a literal listener, $P_{L0}(s|u)$, who interprets u deterministically as true or false ($[\![u]\!](s) \in \{0, 1\}$) and factors in the cost of u, $Cost(u)$. The speaker is modeled as a SoftMax-optimal agent choosing utterances to best convey s. Both listeners apply Bayesian inference to update beliefs over states from the prior, *P(s)*, which serves as the shared common ground (Stalnaker, 1978, 2002).

The RSA model and its close extensions successfully cover a wide range of pragmatic phenomena (see Degen, 2023; Scontras et al., 2021 for a review), including those involving negation, such as indirect politeness and negative strengthening (e.g., *not bad* vs. *not amazing* in Yoon et al., 2020), projective content that survives negation (Qing et al., 2016), and presupposition triggering (Warstadt, 2022). However, the use of RSA to specifically address the pragmatic consequences of sentence polarity asymmetries has received less attention. Theoretical work on negation (e.g., Jakobson, 1963; Givón, 1978; Horn, 1989) suggests that positive and negative polarities show (at least) two asymmetries, which we refer to as Asymmetry Hypotheses 1 and 2:

- Asymmetry Hypothesis 1: Marked forms like negation are typically realized using more complex structures and longer linguistic forms, which are known to elicit higher production cost than their unmarked counterparts; and
- Asymmetry Hypothesis 2: Negation presupposes that its positive-polarity counterpart is relevant or prominent in the common ground, not the other way around.

In this paper, we aim to (i) empirically test the pragmatic consequences of the two asymmetry hypotheses and to (ii) characterize the empirical patterns associated with two types of asymmetry within the RSA framework.

The first asymmetry is closely linked to the trade-off between informativeness and cost that the

Proceedings of the Society for Computation in Linguistics (SCiL) 2025, pages 74-83.
Eugene, Oregon, July 18-20, 2025

pragmatic speaker in RSA must consider. Given that a pragmatic speaker aims to maximize informativeness and minimize cost, the standard RSA model predicts that a negative utterance is less likely to be produced than a similarly informative positive-polarity utterance, i.e., when the states they refer to have similar prior probabilities. Consider part-whole relations as a concrete example. Assuming that situations like *The house doesn't have a bathroom* and *The house has a* ***ball****room* have similar prior probabilities (see below for details on a norming study of state priors), utterances describing these situations should be similarly informative. However, when the standard RSA model (in particular, the pragmatic speaker) penalizes higher-cost utterances, the negative utterance yields a lower utility and is therefore less likely to be produced.

The second asymmetry regarding presupposition accommodation is closely related to common ground update. Assuming that negation presupposes the probability of its positive-polarity counterpart, a negative utterance requires that this positive counterpart be either part of the common ground or can be accommodated. If it is not already common ground knowledge, listeners must accommodate the presupposition before the negative utterance can successfully update the common ground with the negated information. Thus, if a speaker says *The house doesn't have a ballroom*, then in principle the negative utterance presupposes the possibility of *The house has a ballroom*. However, since *ballroom* is not a typical part of *house*, the listener must accommodate this atypical part-whole relation before the negative utterance can be deemed pragmatically motivated and smoothly integrated into common ground.

Utterance choices can be easily probed by asking naïve participants how likely they are to mention certain things. In contrast, directly asking whether a negative utterance presupposes the possibility of its positive-polarity counterpart is less likely to yield interpretable results. To probe this second asymmetry, we instead asked participants to rate the typicality of the whole entity under discussion (e.g., *house*, see more details in Experiment 2 in Section 3.2).

As we show in Section 3, (i) the empirical data patterns are more complex than those predicted by either hypothesis, (ii) while the standard RSA model aligns with the predictions of Hypothesis 1, it fails to account for our findings, and (iii) the standard RSA model lacks a mechanism for common ground updating such that it can't capture Hypothesis 2, let alone explain the observed data. In light of this, we extend the standard model to better capture our empirical findings.

2 Related Work

The standard RSA model (formalized in (1)-(4)) tends to idealize the key components–such as common ground and the literal listener–that are, in practice, subject to uncertainty in real-world communication. Before delving into the empirical findings and our extended RSA models, we review relevant work on common ground update and soft semantics (as opposed to deterministic semantics).

2.1 Common ground update in RSA

Degen et al. (2015) observed that the single prior mechanism in the standard RSA model predicts no scalar implicature in a *some*-utterance that introduces a high-prior event, e.g., *Some marbles sank into water*, while both theoretical observations (Geurts, 2010) and empirical data (Degen et al., 2015) suggest that the scalar implicature is, in fact, strong. To solve this issue, Degen et al. proposed a complex prior, $P(s|w)$ in (5), which determines the world (wonky vs. normal) based on the variable *wonkiness*, *w*. In their ***w*RSA model** (see (5) – (8)), the pragmatic listener, $P_{L1}(s, w|u)$, jointly infers the actual state and the world wonkiness.

$$P(s|w) \propto \begin{cases} 1 & if\ wonky\ world \\ P(s) & if\ normal\ world \end{cases} \quad (5)$$

$$P_{Lo}(s|u,w) \propto [\![u]\!](s) \cdot P(s|w) \quad (6)$$

$$P_{S1}(u|s,w) \propto exp(\alpha\ (ln\ P_{Lo}(s|u,w) - Cost(u)) \quad (7)$$

$$P_{L1}(s,w|u) \propto P_{S1}(u|s,w)\ \cdot P(s|w) \cdot P(w) \quad (8)$$

This model predicts that, when observing a *some*-utterance that introduces a high-prior event, the pragmatic listener backs off to the wonky world where the event has a lower prior probability. This adjustment makes the *some*-utterance a more reasonable utterance choice for the speaker. Degen et al.'s study shows that this extended model fits the empirical data much better than the basic model, in terms of updating both state and world priors.

Kravtchenko and Demberg (2022b), using the core ideas from the *w*RSA model to predict *atypicality inferences* in redundant descriptions of habitual events, found that low-utility utterances led listeners to infer that the habituality of an agent's actions was lower than typically expected.

However, as Cremers et al. (2023) point out, Degen et al. (2015)'s implementation of the *w*RSA model deviates from strict Bayesian reasoning. Instead of directly using the empirically obtained prior distribution over world states in the pragmatic listener's belief of common ground, the model assigns a weighted combination of two worlds: one uniform ('wonky world') and one empirical (representing 'normal world'), which contaminates the so-called 'observation'. Therefore, Cremers et al. (2023) replaced *P(s|w)* with *P(s|normal world)* for the literal listener. See (9) for the modification that we adapted from Cremers et al. (2023):

$$P_{L1}(s, w|u) \propto P_{S1}(u|s, w) \cdot P(s|normal) \cdot P(w) \quad (9)$$

Degen et al.'s proposal of a complex prior inspired more work on the joint inference of common ground and state (Qing et al., 2016; Warstadt, 2022) that involve another approach, namely, Question under Discussion (QUD; Roberts, 1996/2012). By inferring a pragmatic speaker's question under discussion, the pragmatic listener finds a way to rationalize utterances.

For the present study, we want to start with the approach of complex prior, for which our empirical data provide a meaningful test ground. However, this does not exclude QUD as a future direction.

2.2 Soft semantics in RSA

The literal listener's model in the vanilla RSA model and most of its variants interprets an utterance with a deterministic Boolean semantics. Using the examples from Degen et al. (2020), the utterance "small" assigns a probability of 0 to the referent 'big red ball' (false) and the referent 'big blue ball' (false) and assigns a probability of 1 to the referent 'small blue ball' (true), in a finite set consisting only of these three objects.

"Small ball" is the optimal utterance for a listener to most efficiently identify the 'small blue ball', but in natural production, speakers are often redundant, producing "small blue ball" instead. To address this and other empirical-modeling discrepancies with referential expressions, Degen et al. (2020) introduced soft semantics—a continuous semantics—into the RSA model.

Continuing with the examples from Degen et al. (2020), the soft semantics of the utterance "small" can assign a probability of .48 to the 'small blue ball' and a probability of .26 to both the 'big blue ball' and the 'big red ball', reflecting flexibility in literal meaning. Such fuzzy (i.e., vague in the sense of fuzzy logic, Zadeh, 1978) interpretations can be simply represented as follows:

$$[\![u]\!](s) \in [0, 1] \subset \mathbb{R} \quad (10)$$

The literal interpretation is no longer restricted to a binary 'true' vs. 'false' but instead ranges from 0 to 1 in a continuous manner. Regarding the implementation of this continuous semantics, probabilities of literal meanings are decided during model fitting, e.g., using optimization techniques such as Maximum Likelihood Estimation (Degen et al., 2020), or by plugging in pre-normed data when applicable (Yoon et al., 2020). In addition to Degen et al. (2020), the model of the literal listener can also be modified by introducing lexical uncertainty to the lexicon (Bergen et al., 2012).

Degen et al. (2020)'s approach can be interpreted as introducing noise to literal meaning. Relatedly, Bergen and Goodman (2015)'s noisy-channel RSA introduces noise to the transmission of utterance itself that affects literal meaning as well: The received utterance may differ from the intended utterance at the string level. Kravtchenko and Demberg (2022b) adapted the noisy-channel RSA to model the effects of framing on atypicality inferences, showing that emphasis (e.g., via exclamation punctuation) strengthens these inferences. They argue that with emphasis redundant utterances are less prone to misremembering or being ignored, and thus more likely to trigger pragmatic inferences.

In the case of negation, soft semantics might be able to capture both types of noises, namely fuzzy interpretations of negative utterances and their potentially noisy transmission. This is suggested by various prior observations regarding negation: (i) Theoretically, negation is said to presuppose the existence of the negated (Horn, 1989), (ii) empirically, negative sentences trigger the activation of both the negated representation (e.g., *door-not open*) and the negative representation (e.g., *door-open*) (Kaup et al., 2006), and (iii) negation impacts memory in that negative situations can be misremembered as their positive counterparts (Maciuszek & Polczyk, 2017; Cornish & Wason, 1970).

3 Sentence Polarity Asymmetries

We collected utterance choice preferences in Experiment 1 to test Hypothesis 1 and the standard RSA model. We collected typicality ratings in Experiment 2 to test Hypothesis 2. As previewed

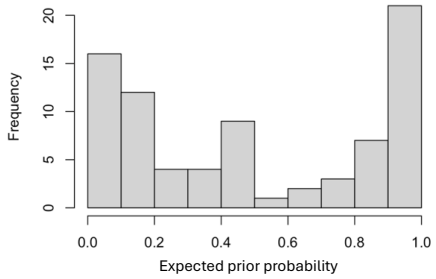

Figure 1: **Norming study.** Histogram of expected values of each smoothed prior distribution

earlier, the results of the experiments reveal more nuanced sentence polarity asymmetries than can be fully captured by either of the two asymmetry hypotheses or the standard RSA model.

3.1 Norming of the state prior

In both experiments, our stimuli were sentences describing real-world part-whole relations such as *house-garage* and their negative forms *house-no garage*. To test how prior probabilities of these part-whole relations influence utterance likelihood and sentence interpretation in the standard RSA model and human data, we first conducted a norming study. This norming study (n=57) measured prior probabilities of 81 part-whole pairs.

The pairs consisted of 27 whole entities and three part entities for each whole entity. On each trial, participants saw two words: the whole entity in capitals (e.g., *CLASSROOM*) and the part entity in lower case (e.g., *stove*). Participants gave their ratings on a slider scale (0-100%) to answer questions about *state prior probability*, e.g., how likely they think a stove is part of or seen in a classroom. The percentage rating distributions for each pair were smoothed using a nonparametric density estimation method suited for ordinal categorical variables with the np package in R (Hayfield & Racine, 2008), following Degen et al. (2015). This non-parametric smoothing method is used in all experiments reported here to handle outliers in our relatively small samples, while preserving the ordinal nature of the rating data.

As Figure 1 shows, the data have a wide range of coverage while somewhat oversampling the high and low ends of probability. This is ideal for generalizing findings across levels of state priors.

3.2 Informativeness-cost trade-off

Experiment 1 (n=52) measured utterance likelihoods of individual part-whole relations being explicitly mentioned. On each trial, participants read a two-sentence sequence followed by a question. The first sentence is a lead-in that introduces the 'whole' entity, e.g., *Emma visited a friend's house yesterday*. The second sentence states a fact about what the place has (i.e., the 'part' entity), in either positive or negative polarity (*The house has a bathroom* or *The house doesn't have a bathroom*). Each participant saw an equal amount of positive and negative-polarity items. For each item, participants rated utterance likelihood, e.g.

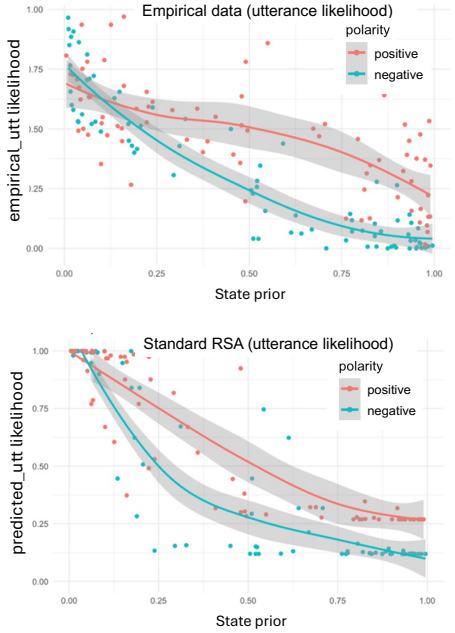

Figure 2: **a.** Empirically collected utterance likelihood (top) **b.** Model (standard RSA) predictions of utterance likelihood (bottom)

How likely do you think it is that Emma would mention that? Participants gave their ratings on a slider scale (0- 100%).

Figure 2a shows the utterance likelihoods from the human participants for both positive-polarity and negative statements. Visual inspection indicate that (i) for both sentence polarities, utterance likelihoods decrease as the state priors increase, (ii) for negative polarity, the decrease of utterance likelihoods as state priors increase is steeper than positive polarity. These patterns suggest a main effect of state prior and an interaction between state prior and sentence polarity.

Beta regression analysis confirms that there is a main effect of state prior (β = -4.36, SE = 0.28, z = -15.74, $p < .001$), and an interaction effect between state prior and sentence polarity (β = 2.43, SE =

0.35, z = 6.96, $p < .001$). From the positive sign of the interaction effect, we can confirm that the negative polarity yields a steeper decrease in utterance likelihood as the state prior increases. In addition, we found no main effect of sentence polarity (β = -0.20, SE = 0.20, z = -1.05, $p = .296$).

These results reveal patterns that Asymmetry Hypothesis 1 does not predict. On one hand, overall, positive utterances are not always perceived as having higher utterance likelihood. On the other hand, speakers are more likely to communicate low-informativeness information using positive polarity and more likely to communicate high-informativeness information using negative polarity.

Model predictions (standard RSA): Now let us see whether the standard RSA model can capture these observations. The model (as in (1)-(4)) is run in R using the rwebppl package[1].

The model considers two states: $U_{state} = \{s_{pos}, s_{neg}\}$ and three possible utterances: $U_{utterance} = \{u_{pos}, u_{neg}, u_{null}\}$. These utterances are mapped to truth values of different states. When the null utterance, u_{null} (say nothing), is made, people simply rely on their prior expectations (state prior) to interpret the situation. The positive utterance, u_{pos} "A has B", maps to the truth of only the positive state, s_{pos}. The negative utterance, u_{neg} "A doesn't have B", maps to the truth of only the negative state, s_{neg}.

The utterance utility term consists of an informativeness component, a cost component, and a speaker rationality parameter. α is set to 1 and utterance cost is specific to each of the three utterances (Cost(u_{null})=0; Cost(u_{pos})=1; Cost(u_{neg})= 2). *P(s)* is the normed state priors data that we plugged in the model as input.

Figure 2b shows the model-predicted utterance likelihoods for both sentence polarities. Visual inspection indicates that (i) similar to the empirical data, for both sentence polarities, utterance likelihoods decreased as the state priors increase, and (ii) for positive polarity, the predicted utterance likelihood is always higher than the negative. These patterns suggest a main effect of sentence polarity and a main effect of state prior.

Beta regression analysis reveals a main effect of state prior (β = -4.43, SE = 0.36, z = -12.47, $p < .001$). However, unlike human data, we found in the model predictions a main effect of sentence polarity (β = 0.78, SE = 0.28, z = 2.83, $p < .01$), indicating that the positive polarity always yields higher utterance likelihood than the negative polarity. Moreover, we did not find a significant interaction between state prior and sentence polarity (β = 0.13, SE = 0.44, z = 0.28, $p = .78$).

The results suggest that the standard RSA model follows predictions of the Hypothesis 1 and fails to fully capture the empirically observed patterns.

Comparing empirical data and model predictions: The discrepancy centers on the lower bound of the state prior that approaches a probability of 0: Based on human data, negative-polarity situations that have low priors (e.g., *The classroom doesn't have a board.*) are more likely to be communicated than positive-polarity situations that have similarly low priors (e.g., *The classroom has a stove.*). However, given that our human data were not collected in a spontaneous production study, it is possible that the Experiment 1 participants did not consider the role of utterance cost. We want to be cautious about committing to this pattern of sentence polarity asymmetry, so we ran another model simulation with the utterance cost constant as 1 for both sentence polarities.

Beta regression analysis now shows a main effect of state prior (β = -4.45, SE = 0.32, z = -13.75, $p < .001$), no effect of sentence polarity (β = 0.14, SE = 0.27, z = 0.53, $p = .59$), and no interaction between state prior and sentence polarity (β = 0.01, SE = 0.41, z = 0.02, $p = .98$). This shows that the model-predicted utterance likelihood of negative and positive sentences patterns alike, which is not surprising given how the model parameters do not differentiate them.

The results above suggest that even when cost is controlled, the standard RSA model fails to capture the sentence-polarity asymmetry observed in our empirical utterance likelihood data.

In the other model implementations in this paper, we thus assume higher cost for negative utterances than positive ones (also in line with cognitive psychology and linguistics research).

3.3 Common ground update

Experiment 2 (n=52) collected typicality ratings of the whole entity (e.g., *house*) using the same stimuli as in Experiment 1, except that the fact statement of a positive/negative part-whole relation was embedded in direct speech in Experiment 2, e.g., *"The house has a bathroom," Emma told her*

[1] https://github.com/mhtess/rwebppl

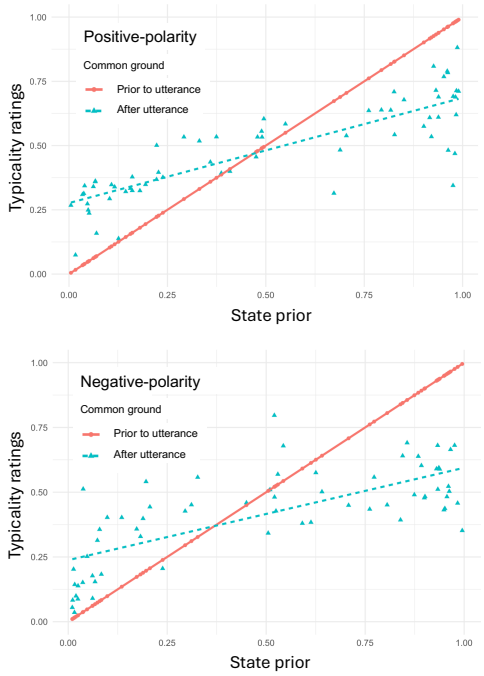

Figure 3: **a.** Ratings pre- vs. post- positive polarity utterances (top) **b.** Ratings pre- vs. post- negative polarity utterances (bottom)

partner. Participants were instructed to rate the typicality of *house* based on what the protagonist said about it (e.g., *How likely do you think it is that the house is a typical house?*).

Following standard RSA practice (Frank and Goodman, 2012; Degen et al., 2015), we compare ratings before and after utterances are presented to participants. Pre-utterance ratings (the norming data in Section 3.1) reflect (the listener's belief of) common ground prior to communication, while post-utterance typicality ratings (this section, Experiment 2) reflect updated common ground triggered by the utterance, in line with the discussions about Asymmetry Hypothesis 2.

Figure 3a shows these two types of ratings for positive polarity (solid line: state prior; dashed line: updated common ground). Figure 3b shows the same results for negative polarity.

The ratings were analyzed with Pearson correlation and beta regression. *First*, we assessed the correlation between state prior (norming) and typicality ratings (Experiment 2). To test this, we conducted a Pearson correlation: the typicality ratings are more strongly correlated with the state prior in positive polarity ($r_{utt}(76)= 0.84$, $p<0.01$) than in negative polarity ($r_{utt}(76)= 0.75$, $p<0.01$).

Second, to compare sentence polarities directly, we analyzed the interaction between polarity and state prior on typicality ratings using beta regression. We found a main effect of polarity (β= -0.41, SE = 0.08, z = -5.16, $p<0.01$) where the negative polarity yielded lower typicality ratings than the positive polarity, a main effect of state prior (β= 1.66, SE = 0.11, z = 15.01, $p<0.01$) where typicality ratings increased with state priors, but no interaction (β= 0.05, SE=0.22, z=0.21, p=0.83).

These results suggest that negation triggers stronger common ground update/inferences. However, importantly, our results suggest that (i) the positive-polarity is not free of inferences, and (ii) both sentence polarities can trigger atypicality inferences (Kravtchenko and Demberg, 2022ab) and what we call *typicality inferences* (i.e., low prior states are inferred to be more typical post- vs. pre-utterances).

Model predictions (standard RSA): The standard RSA model uses Boolean semantics, so the model updates the state posterior to 1 based on the only state that a non-null utterance makes true, but makes no inferences about common ground.

Comparing empirical data and model predictions: The comparison is fairly straightforward: The standard RSA model cannot handle common ground update.

Motivated by the discrepancies between empirical observations and model predictions (of the standard RSA), in the following Sections 4 to 6, we extend the standard model to better capture our empirical findings.

4 *fuzzy*RSA

The goal of Section 4 is to pinpoint the sentence polarity asymmetry related to the informativeness-cost tradeoff (i.e., a pragmatic speaker aims to maximize informativeness and minimize cost). Building on prior work, we introduce soft semantics into the standard RSA model to capture the asymmetry observed in utterance likelihood. We call this extended model the ***fuzzy*RSA model**.

4.1 Model

The *fuzzy*RSA model is extended from the standard RSA model by configuring different interpretation functions across sentence polarities. For a negative utterance, the fuzzy interpretation is defined as a constant probability distribution of a negative state and a positive one (see (11), where $n \in [0,1]$), with its optimal value determined during model fitting. For instance, when n is

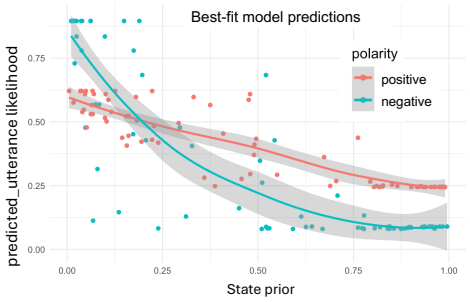

Figure 4: ***fuzzy*RSA.** Utterance likelihood values generated from the best-fit model.

assigned a value of .7, "A doesn't have B" assigns a probability of .7 to "*A-no B*" and .3 to "*A-B*".

$$[\![u_{neg}]\!](s_{neg}) = n,\ [\![u_{neg}]\!](s_{pos}) = 1 - n \quad (11)$$

This constant formulation reflects the 'inherent' pragmatic feature of negation as a presupposition trigger, which applies to all negative utterances.

For a positive utterance, the fuzzy interpretation is defined as a parametrized sigmoid function of the priors of positive states (see (12-13)), also fit during model optimization.

$$\begin{cases} [\![u_{pos}]\!](s_{pos}) = Sigmoid(P(s_{pos});\ \theta) \\ [\![u_{pos}]\!](s_{neg}) = 1 - [\![u_{pos}]\!](s_{pos}) \end{cases} \quad (12)$$

$$S_{\theta=\{L,k,x_0,c\}}\left(P(s_{pos})\right) = \frac{L}{1+e^{-k(P(s_{pos})-x_0)}} + c \quad (13)$$

The sigmoid function in (13) increases rapidly for state priors that are relatively low and gradually approaches the maximum value (i.e., approaching 1) towards relatively high state priors. The sigmoid function captures a systematic relationship between the state prior and the probability of interpreting a positive utterance as intended. Compared to the negative polarity, the interpretation function associated with positive polarity disincentivizes the communication of low-prior positive states.

4.2 Model fitting

We optimized model parameters by minimizing the joint loss across negative and positive polarities. This joint loss was computed as the sum of squared differences between model predictions and empirical data. A grid search over pre-specified parameter ranges—informed by exploratory model simulations—was used to identify the best fitting-values: n=.8, α=1, θ={L=0.7, k=6, x_0=.35, c=0.3}. The best-fit model has a mean square error (MSE) of 0.04 (compared to a MSE of 0.06 for standard RSA model).

4.3 Model predictions

Figure 4 shows that the *fuzzy*RSA model predicts patterns that resemble the empirical data. The results suggest that the *fuzzy*RSA model provides a better approximation of the empirical data and potentially of the cognitive processes involved in inferring utterance likelihood.

5 *wonky*RSA

In another extended model, we introduce a complex prior to capture the asymmetry in typicality ratings and provide a mechanism for common ground update. We call it the ***wonky*RSA model**.

5.1 Model

As discussed earlier, we integrate Cremers et al. (2023)'s modification into Degen et al's (2015) 'wonky world' model, resulting in the following:

$$P_{L0}(s|u,w) \propto [\![u]\!](s) \cdot P(s|w) \quad (14)$$

$$P_{S1}(u|s,w) \propto_{exp} (\alpha\ (ln\, P_{L0}(s|u,w) - Cost(u)) \quad (15)$$

$$P_{L1}(s,w|u) \propto P_{S1}(u|s,w)\ \cdot P(s|normal) \cdot P(w) \quad (16)$$

In the *wonky*RSA model, presupposition accommodation is reflected in an updated wonkiness, i.e., the wonky world has a higher or lower probability based on how much accommodation is needed.

Before the accommodation, the common ground is *P(s|w = normal)*. After the accommodation, the common ground is a complex probability distribution: *P(s|w = normal)* with a probability of (1-*P(w)*) and *P(s|w= wonky)* with a probability of *P(w)*. In other words, the updated common ground can be represented by the marginalized probability of a state across both worlds. We assume that the post-utterance ratings collected (typicality ratings; Experiment 2) reflect this updated common ground, which we refer to as *expected typicality*, formalized as following:

$$\mathbb{E}(typicality) = \sum_{world} P(world) * P(s|world) \quad (17)$$

5.2 Model fitting

We optimized model parameters by minimizing the joint loss across negative and positive polarities. This joint loss was computed as the sum of squared differences between expected typicality and typicality ratings. A grid search over pre-specified parameter ranges—informed by exploratory model

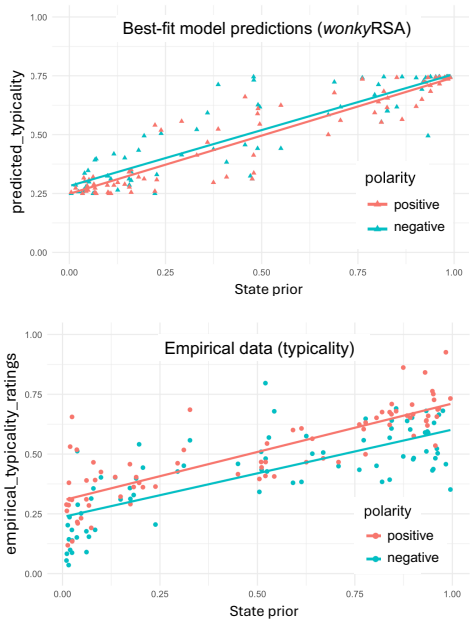

Figure 5: **a.** Model predictions of post-utterance expected typicality in positive vs. negative polarities (top) **b.** Post-utterance typicality ratings in positive vs. negative polarities (bottom)

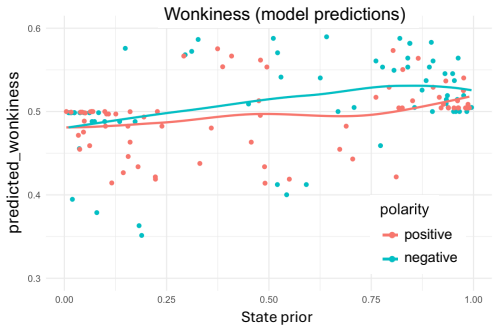

Figure 6: *wonky*RSA's predictions of **wonkiness**

simulations—was used to identify the best fitting-values: w=.5, α=2. The best-fit model has a MSE of 0.02 (while the standard RSA model is unable to make predictions regarding presupposition accommodation).

5.3 Model predictions

The best-fit model is able to capture two aspects of our empirical patterns. (i) Figure 5a shows that the *wonky*RSA model predicts both typicality and atypicality inferences in both sentence polarities. (ii) Figure 6 shows that the model-predicted wonkiness values more or less align with the inference patterns: lower-than-prior wonkiness is predicted where typicality inferences are observed, and higher-than-prior wonkiness is predicted where atypicality inferences are observed.

However, the *wonky*RSA model is not yet able to reflect the stronger inferences associated with the negative polarity: Instead of predicting lower typicality ratings for negative polarity (Figure 5b), the model predicts similar typicality values for both sentence polarities (Figure 5a).

This is not surprising given that the *wonky*RSA model does not differentiate two sentence polarities. Therefore, it is necessary to further extend the *wonky*RSA model, which we will discuss in Section 6.

6 *funky*RSA

In a third extended model, we bring together two approaches, soft semantics and the complex prior, from the preceding two models in Sections 4 and 5. This is an attempt to introduce polarity asymmetry into the *wonky*RSA model. We call this combinatory model the ***funky*RSA model**.

6.1 Model

The *funky*RSA model integrates components from *fuzzy*RSA and *wonky*RSA, formalized as shown:

$$P_{L0}(s|u,w) \propto [\![u]\!](s) \cdot P(s|w) \quad (18)$$

$$\begin{cases} [\![u_{pos}]\!](s_{pos}) = Sigmoid(P(s_{pos});\ \theta) \\ [\![u_{pos}]\!](s_{neg}) = 1 - [\![u_{pos}]\!](s_{pos}) \end{cases} \quad (19)$$

$$S_{\theta=\{L,k,x_0,c\}}\left(P(s_{pos})\right) = \frac{L}{1+e^{-k(P(s_{pos})-x_0)}} + c \quad (20)$$

$$P_{S1}(u|s,w) \propto exp(\alpha\ (ln\, P_{L0}(s|u,w) - Cost(u)) \quad (21)$$

$$P_{L1}(s,w|u) \propto P_{S1}(u|s,w)\ \cdot P(s|normal) \cdot P(w) \quad (22)$$

6.2 Model predictions

Instead of fitting the model from scratch, we plugged in the values of parameters that contributed to the best-fit *fuzzy*RSA and *wonky*RSA models. Note that these two models differ in their values of the speaker rationality parameter α. We thus ran the *funky*RSA model with both values which yielded similar results for typicality. Figure 7 shows the model predictions of typicality in both polarities.

The model does predict a difference between sentence polarities; however, the predicted difference does not align well with the empirical findings: The negative polarity does not yield lower typicality values than the positive polarity.

This suggests that while optimal parameter values from *fuzzy*RSA and *wonky*RSA models

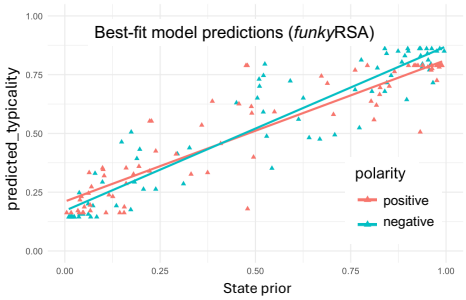

Figure 7: *funky*RSA model's predictions of typicality($\alpha = 1$).

provided a starting point, they do not yield satisfactory predictions when applied directly to the *funky*RSA model. Due to the increased complexity and computational cost of jointly optimizing all parameters in the *funky*RSA model, we leave full optimization for future work.

For utterance likelihood, we assume that the empirical ratings reflected participants' choices in a normal world. the *funky*RSA model makes the same predictions as the *fuzzy*RSA model regarding utterance likelihood.

7 Discussion

In this paper, we (i) empirically tested two hypotheses about sentence polarity asymmetries and (ii) introduced three extended RSA models that demonstrated the potential to better capture our empirical data than the standard RSA model.

The empirical data from Experiments 1 and 2 reveal patterns that are not predicted by the standard RSA model. Results of utterance likelihood ratings (Experiment 1) show that, although negation is theoretically deemed as a less optimal utterance choice than the positive polarity regarding the informativeness-cost tradeoff, negative utterances are not always less likely than positive utterances. Results of typicality ratings (Experiment 2) show that both state priors and sentence polarity play a role in triggering pragmatic inferences. Although negative utterances were associated with stronger inferences, positive utterances also yielded pragmatic accommodation.

To capture these novel empirical findings within the RSA framework, we targeted two components of an RSA model, namely the interpretation function that gives rise to literal meaning, and the configuration of common ground that allows presupposition accommodation. Inspired from prior work on soft semantics in RSA, our *fuzzy*RSA model uses different soft-semantics interpretation functions for different sentence polarities. Adapted from prior work on wonky world RSA models, our *wonky*RSA model provides a complex prior for common ground update. Combining *fuzzy*RSA and *wonky*RSA models, we then propose the *funky*RSA model which aims to introduce interpretation-level sentence polarity asymmetry into the *wonky*RSA model. The three extended RSA models yield somewhat better predictions than the standard RSA model and somewhat satisfying results that align better with the results of Experiments 1-2.

However, some questions remain open. *First*, regarding the different configurations in how different sentence polarities are literally interpreted, we formalized a sentence polarity asymmetry at a semantic level (i.e., through fuzzy interpretations). This worked for the predictions of utterance likelihood (*fuzzy*RSA model) but not for the predictions of typicality (*funky*RSA model), which might suggest that sentence polarity asymmetry is not limited to the difference in literal interpretations. Thus, future work should explore approaches to formalizing the sentence polarity asymmetry more closely related to common ground update. *Second*, regarding the complex prior used in the *wonky*RSA model, we explored one version of the wonky world—a uniform prior. This, however, is a potential source of sentence polarity asymmetry. For example, the wonky world assumed for negative utterances may differ from that for positive ones We plan to explore other configurations of the wonky world in future work.

8 Conclusion

This paper presents novel empirical findings on sentence polarity asymmetries and offers one of the first formalizations of these asymmetries within the RSA framework. The contributions are two-fold. Theoretically, this study highlights the important role of prior knowledge in pragmatic reasoning and offers new insights into both production and comprehension of negation. Empirically, we show that existing extensions of the RSA model, e.g., soft semantics and common ground update, while not originally proposed to address sentence polarity asymmetries, can nonetheless be applicable to these phenomena. This supports the generalizability of these approaches, as well as strengthens the broader applicability of the RSA framework.

Acknowledgments

We thank the anonymous reviewers for thoughtful feedback on this project.

References

Bergen, Leon, Noah Goodman, and Roger Levy. 2012. That's what she (could have) said: How alternative utterances affect language use. In Proceedings of the Annual Meeting of the Cognitive Science Society, 34(34).

Bergen, L., & Goodman, N. D. 2015. The strategic use of noise in pragmatic reasoning. *Topics in Cognitive Science*, 7(2), 336–350.

Cornish, E. R., & Wason, P. C. 1970. The recall of affirmative and negative sentences in an incidental learning task. *Quarterly Journal of Experimental Psychology*, 22(2), 109–114.

Cremers, A., Wilcox, E. G., & Spector, B. 2023. Exhaustivity and Anti-Exhaustivity in the RSA Framework: Testing the Effect of Prior Beliefs. *Cognitive Science*, 47(5), e13286.

Degen, Judith, Michael Henry Tessler, and Noah D. Goodman. 2015. Wonky worlds: Listeners revise world knowledge when utterances are odd. In Proceedings of the 37th Annual Meeting of the Cognitive Science Society, 548–553. Austin, TX: Cognitive Science Society.

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. 2020. When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review*, 127(4), 591–621.

Degen, J. 2023. The rational speech act framework. *Annual Review of Linguistics*, 9(1), 519–540.

Frank, Michael C., and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. Science 336(6084):998. https://doi.org/10.1126/science.1218633

Geurts, Bart. 2010. Quantity Implicatures. Cambridge: Cambridge University Press.

Givón, T. 1978. Negation in language: Pragmatics, function, ontology. In *Pragmatics*, 69–112. Brill.

Hayfield, Tristen, and Jeffrey S. Racine. 2008. Nonparametric econometrics: The np package. Journal of Statistical Software 27(5).

Horn, Laurence R. 1989. A Natural History of Negation. Chicago, IL: University of Chicago Press.

Jakobson, R. 1963. Implications of language universals for linguistics. In *Roman Jakobson: Selected Writings II*, 580–592. The Hague: Mouton.

Kaup, B., Lüdtke, J., & Zwaan, R. A. 2006. Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, 38(7), 1033–1050.

Kravtchenko, Ekaterina, and Vera Demberg. 2022a. Informationally redundant utterances elicit pragmatic inferences. Cognition 225:105159. https://doi.org/10.1016/j.cognition.2022.105159

Kravtchenko, E., & Demberg, V. 2022b. Modeling atypicality inferences in pragmatic reasoning. In Proceedings of the Annual Meeting of the Cognitive Science Society, 44(44).

Maciuszek, J., & Polczyk, R. 2017. There was not, they did not: May negation cause the negated ideas to be remembered as existing? *PLoS One*, 12(4), e0176452.

Qing, Ciyang, Noah D. Goodman, and Daniel Lassiter. 2016. A rational speech-act model of projective content. In Proceedings of CogSci.

Roberts, C. 1996. Information structure: Towards an integrated formal theory of pragmatics. In Je Hak Yoon and A. Kathol (eds.), *Ohio State University Working Papers in Linguistics (OSUWPL)*, 49: Papers in Semantics. Columbus, OH: The Ohio State University Department of Linguistics.

Roberts, C. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5, 1–69. [Reprint of Roberts (1996).]

Scontras, Gregory, Michael Henry Tessler, and Michael Franke. 2021. A practical introduction to the Rational Speech Act modeling framework. arXiv preprint arXiv:2105.09867.

Stalnaker, Robert. 1978. Assertion. In Formal Semantics: The Essential Readings, 147-161.

Stalnaker, Robert. 2002. Common ground. Linguistics and Philosophy 25(5/6):701-721. https://doi.org/10.1023/A:1020867916902

Warstadt, Alex. 2022. Presupposition triggering reflects pragmatic reasoning about utterance utility. In Proceedings of the 2022 Amsterdam Colloquium.

Yoon, Eunice J., Michael Henry Tessler, Noah D. Goodman, and Michael C. Frank. 2020. Polite speech emerges from competing social goals. Open Mind 4:71-87. https://doi.org/10.1162/opmi_a_00035

Zadeh, L. A. 1978. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1(1), 3–28.