

“Today in Timiryazevsky Park [two] world leaders participated in a 100 meter foot race. Our Soviet Premier, Nikita Khrushchev, finished a very respectable second place. The poor American President, John F. Kennedy finished a miserable next-to-last.”
— ascribed to Russian newspaper *Pravda*

Abstract

Concrete.

Keywords: pragmatic language use; argumentative framing; argument strength; probabilistic modeling;

Introduction

Overwhelming evidence from decades of research demonstrates that human language use is often strongly oriented towards providing a high degree of useful information for addressees [MF: REFs]. This picture aligns well with the classic Gricean framework (Grice, 1975), which posits a cooperative principle alongside Maxims of Quality (truth), Quantity (informativity), and Relevance—where the latter is typically interpreted as relevance to the listener.

Formal models of pragmatic language use consequently often rely on some kind of formal, relative, or quantitative notion of informativity, versions of which are supplied by formal logic (e.g., asymmetric logical entailment) or information theory. Based on such formalized quantitative notions of informativity, we can describe speakers’ pragmatic choice of expressions as a form of utility maximization, defined—at least in part—by the drive to optimize relevant information exchange (Parikh, 1991; Blutner, 2000).

However, human talk does not always revolve around the straightforward exchange of relevant information alone. We also express an interpretative stance towards the facts we describe, communicating sentiment and opinion alongside descriptions of the world. Indeed, some work suggests that human talk is much more about persuasion, argumentation, manipulation, and opinion-negotiation than about pure, objective—indeed, robot-like—information-sharing (Anscombe & Ducrot, 1983; Mercier & Sperber, 2011; Enfield, 2024). For example, as a case of selective truth-telling, or *paltering* (Rogers et al., 2017), the same outcome of an exam could be truthfully described in a way that makes the students appear successful or less successful overall:

- (1) a. Most of the students got most questions right.
b. Some of the students got all questions wrong.

Yet, there is much less formal work on this kind of language use. What seems to be blatantly lacking are good formal descriptions of what makes one utterance more suitable than another for a speaker with an intent to argumentatively frame their contribution. In this work, we therefore ask which kind of utility function might underlie a speaker’s choice to prefer one utterance over another in a persuasive or argumentative context: how can we capture formally what makes one argument better than another? We address this question through the notion of *argumentative strength*.

Some prior work (e.g., Merin, 1997; van Rooij, 2004; Cummins & Franke, 2021) suggests a formalization of argumentative strength in terms of *log-likelihood ratios*, based on early work from the philosophy of science on observational evidence (Good, 1950); see the formalization below. However, so far there has been no stringent empirical testing of this notion, nor a systematic comparison against closely related competitor notions in terms of their predictive accuracy for experimental data. Providing such a comparison is the main contribution of this work.

We focus on a setting of selective truth-telling, or paltering **Roberts**. We study cases similar to the real-life examples examined by Cummins and Franke (2021), but consider a constrained experimental setting that allows statistical model comparison of bespoke probabilistic models which differ only in the quantitative notion of argumentative strength they assume underlies the speakers’ choice of expression. In the following, we formulate these models in the tradition of the RSA framework **RS**A. Next, we describe the experimental set-up, and report on the statistical model comparisons, as well as our results. Our main findings are that [MF: fill me ...].

Probabilistic models of argumentative language

Probabilistic models of pragmatic reasoning usually define a speaker and listener policy. Here, we will focus exclusively on the speaker’s policy. In line with the usual assumption of Bayesian decision-makers, the speaker’s policy of choosing an utterance u , when trying to communicate a state s is defined in terms of a soft-max operation (Franke & Degen, 2023), with parameter α on the utility function $U(u, s)$:

$$P_S(u \mid s) \propto \exp(\alpha U(u, s)) . \quad (1)$$

The utility function $U(u, s)$ captures how good it is for the speaker to choose u when the true state, according to the speaker, is s . We here follow the Rational Speech Act (RSA) modeling framework (e.g., Frank & Goodman, 2012; Degen, 2023) which adopts the usual Gricean assumptions that the speaker wants to speak truly, maximize the amount of information conveyed about the state s , and to minimize their own speaking effort (Grice, 1975). To implement these assumptions, utilities are defined as a sum of the information-theoretic surprisal of s conditional on u being true and the (negative) cost of u , which is a stand-in for production effort or ease of accessibility:

$$U(u, s) = \log P(s | [[u]]) - \text{cost}(u), \quad (2)$$

where $[[u]] \subseteq S$ is the semantic denotation of u , formally represented as the set of world states in which u is true. Under the wide-spread assumption of a flat prior over s , the conditional probability of s given that u is true can be written as:

$$P(s | [[u]]) = \begin{cases} |[[u]]|^{-1} & \text{if } s \in [[u]] \\ 0 & \text{otherwise.} \end{cases}$$

With this, if the semantics for utterances is binary, the utility function from above can be factored into three well-known aspects of pragmatic language generation, namely the requirements that the speaker's utterance be true, informative and economical (Scontras et al., 2021):

$$U(u, s) = \underbrace{\log [s \in [[u]]]}_{\text{truth}} + \underbrace{\log |[[u]]|^{-1}}_{\text{informativity}} - \underbrace{\text{cost}(u)}_{\text{economy}}.$$

The speaker's policy defined above in Equation (1), when used with the standard utility function in Equation (2) has been productively used to explain choices of utterances for different linguistic constructions of phenomena, e.g., for referential expressions (Frank & Goodman, 2012), generics (Tessler & Goodman, 2019), conditionals (Grusdt et al., 2022), quantifiers and implicature (Goodman & Stuhlmüller, 2013; van Tiel et al., 2021), gradable adjectives (Lassiter & Goodman, 2017), or probability expression (Herbsttritt & Franke, 2019). [MF: insert some more references (without MF as co-author!)] Yet, some phenomena seem to require more elaborate utility functions. For example, in the realm of social meaning, extensions of the vanilla RSA model sketched above have been explored which incorporate additional utility components related to politeness (Yoon et al., 2020). Here, we take a similar approach to modelling the utility trade-off between describing the world informatively and making an argument in favor of a position or hypothesis H_0 , as opposed to the competing position or hypothesis H_1 . The general form of the extended speaker utility function we consider in this paper will be:

$$U(u, s, H_0, H_1) = \underbrace{\beta \log P_{L_0}(s | [[u]])}_{\text{truth \& informativity}} + \underbrace{(1 - \beta) \text{argstr}(u, H_0, H_1)}_{\text{argumentative strength}} - \underbrace{\text{cost}(u)}_{\text{economy}}$$

Following the previous literature, the parameter β models the degree to which a speaker values optimizing informativity

of an utterance or making a strong argument for position H_0 (relative to H_1). For the special case of $\beta = 1$, this formulation reduces to the previous utility function which did not have argumentative strength as an additional speaker objective for utterance selection.

In the following we will explore different models of the speaker's utterance choice:

1. The **vanilla RSA model** provides the conservative baseline. It contains no speaker objective for argumentative speech; alternatively we can think of it as a model with $\beta = 1$.
2. The **likelihood-ratio model** assumes that argumentative strength can be operationalized in analogy to a common measure of observational evidence, the log-likelihood ratio (based on literal interpretation of the utterance).
3. The **pragmatic likelihood-ratio model** is similar to the previous model but computes argumentative strength via log-likelihood ratios based on a pragmatic enrichment of the utterance.
4. The **maximin model** provides a computationally simpler definition of argumentative strength in terms of a form of worst-case reasoning.
5. The **model-free model** uses a situation-specific notion of argumentative strength in terms of the posterior expectation of true answers; this approach is “model-free” in the sense that it does not commit to a strong theoretic position on what argument strength is supposed to be.

- will explore a bunch of notions
- the starting point is log likelihood ratio
- used by a lot of previous work
- but no real quantitative model fits so far

Experiment

To test whether and how speakers choose expressions to frame a complex situation with argumentative information culling, we used an experimental design which presents a perspicuous but complex state of affairs (the results of a high-school exam) and allows participants to choose flexibly from a larger, but still constrained set of alternative expressions. The design used here is essentially the same as that of Experiment 1 reported in (Macuch Silva et al., 2024), except that we here used a larger set of visual scenes (different array sizes, see below). While the work reported by Macuch Silva et al. (2024) also elicited and analyzed free production data, we here focus on a more constrained free-choice task in order to harness the complexity of the data for subsequent modeling. Participants could essentially choose one of 32 sentences, but they did so by selecting, using a drop-selection menu, (i) an outer quantifier, (ii) an inner quantifier, and (iii) an adjective, to complete the sentence frame in (2).

(2) OQ of the students got IQ of the questions ADJ
with OQ and IQ $\in \{ \text{None}, \text{Some}, \text{Most}, \text{All} \}$, and ADJ $\in \{ \text{right}, \text{wrong} \}$.

Participants. A total of $N = 201$ participants were recruited via Prolific (self-identified gender: 88 female, 111 male, 1 other and 1 non-disclosed; mean age (of those who revealed it) 30.3 (standard deviation 8.07), min 18 and max 60). Participants had to be at least 18 years old in order to participate. They were paid £1.5. Based on a mean completion time of just below 10 minutes (median just below 9 minutes), this amounted to an average hourly payment of £1.5. [MF: check: no other Prolific internal selection criteria; English etc.?][FC: We only excluded participants with more than 4 false responses]

Materials. The results of high-school exams were presented visually in form of matrices, as shown in Figure ?? (top left).[FC: Figure is missing] The rows of matrices corresponded to students (indicated by names), the columns indicated questions. A checkmark on green background in a cell represented that the student got the question right. A cross on a red background represented a false answer. The results were always arranged to show students ordered in terms of performance (students with more correct answers on in higher rows). The names of students were sampled at random for each trial from a list of common English first names.

Four sizes of matrices were used, differing in the number of students (5 or 11) and the number of questions in the exam (6 and 12). For example, the matrix in Figure ?? is an instance of a 5×12 matrix. For each matrix size, there were 20 instances, each one corresponding to one of the 20 situations which can be logically distinguished based on sentences of the form in (2). More concretely, the 20 situations are all the “possible world states” that can be differentiated with a language that contains only the sentences in (2) under their standard logical meaning, assuming that *some* means *at least one* and *most* means *more than half* (see Macuch Silva et al., 2024, for details).

Procedure. The experiment started with an explanation of the displays and the task. Participants were instructed to describe the results of high-school exams as either favorable or unfavorable (high vs. low framing condition). Each participant consistently saw one size of the four results matrices, and they saw each one of the 20 instances of that matrix type exactly once in completely randomized order. Trials were randomly assigned to a high or low framing condition, so that each participant saw 10 trials in the high and 10 trials in the low framing condition.

Results. Following preregistered protocol, we excluded all the data from a participant if the participant selected the same response in all trials or if the participant gave more than four responses which are literally false as a description of the results shown in the corresponding trial. This reduced the original number of $N = 201$ participants to $N = 186$. We also excluded any remaining responses that are literally false. This

resulted in another 113 individual responses being removed from the data set.

Figure 1 shows the proportions of sentences which participants generated as descriptions for the exam results. The plot differentiates the two different argumentative framing conditions (*high* vs. *low*, indicated by color in Figure 1), and the four different shapes of the results matrices to be described (different rows in the figure). Visually, the distributions for different matrix sizes seem to be rather similar, but there appear to be striking contrasts in the choices of descriptions between the two different argumentative framing conditions. This impression is corroborated by a simple χ^2 -test. The choice distributions seem to differ between results matrices ($\chi^2 \approx 216.8$, $df = 168$, $p < 0.006$). The *high* vs. *low* framing induced significantly different distributions over descriptions ($\chi^2 \approx 2300.7$, $df = 103$, $p < 2e^{-16}$). The latter result, indicating a difference between *high* and *low* argumentative framing conditions, suggests that our manipulation worked and that participants are indeed able to adapt their description choices to the strategic argumentative framing the context demanded.

Models

We consider different model variants, each using a different notion of argumentative strength.

The calculation of argumentative strength for all models requires the specification of two hypotheses, one that is being argued for (H_0) and one that is being argued against (H_1). We model each hypothesis as saying that all students in the class have a certain, equal probability γ of getting each answer correct; therefore, we make the simplifying assumption that there is no variation in skill across students and no variation in probability of getting different answers correct. [FC: I actually spent a long time in the early phases of the project exploring variations of this - not sure it makes sense to pick up this thread though.] Under this assumption, the probability of observing the results for a class is proportional to the product of the Binomial probability of each student’s results:

$$P(s | \gamma) \propto \prod_{k \in S} \binom{12}{k} \gamma^k (1 - \gamma)^{12-k} \quad (3)$$

where S is the set of exam arrays participants can observe in the experiment, each exam is encoded as a list of numbers of correct answers, and $P(s | \gamma)$ is normalized across S . In the high framing condition, $\gamma = 0.85$ for H_0 and $\gamma = 0.15$ for H_1 , and the reverse for the low framing condition.

Log likelihood ratio argstrength RSA

The starting point is the notion of *weight of evidence*, which has been introduced in the previous literature as a formal notion of argument strength, following [MF: refs]. This notion requires fixing two competing hypotheses H_0 and H_1 and formalizes the argumentative strength of an utterance u as evidence in favor of H_0 as opposed to the (alternative, competing) hypothesis H_1 . Concretely, we consider the degree to which

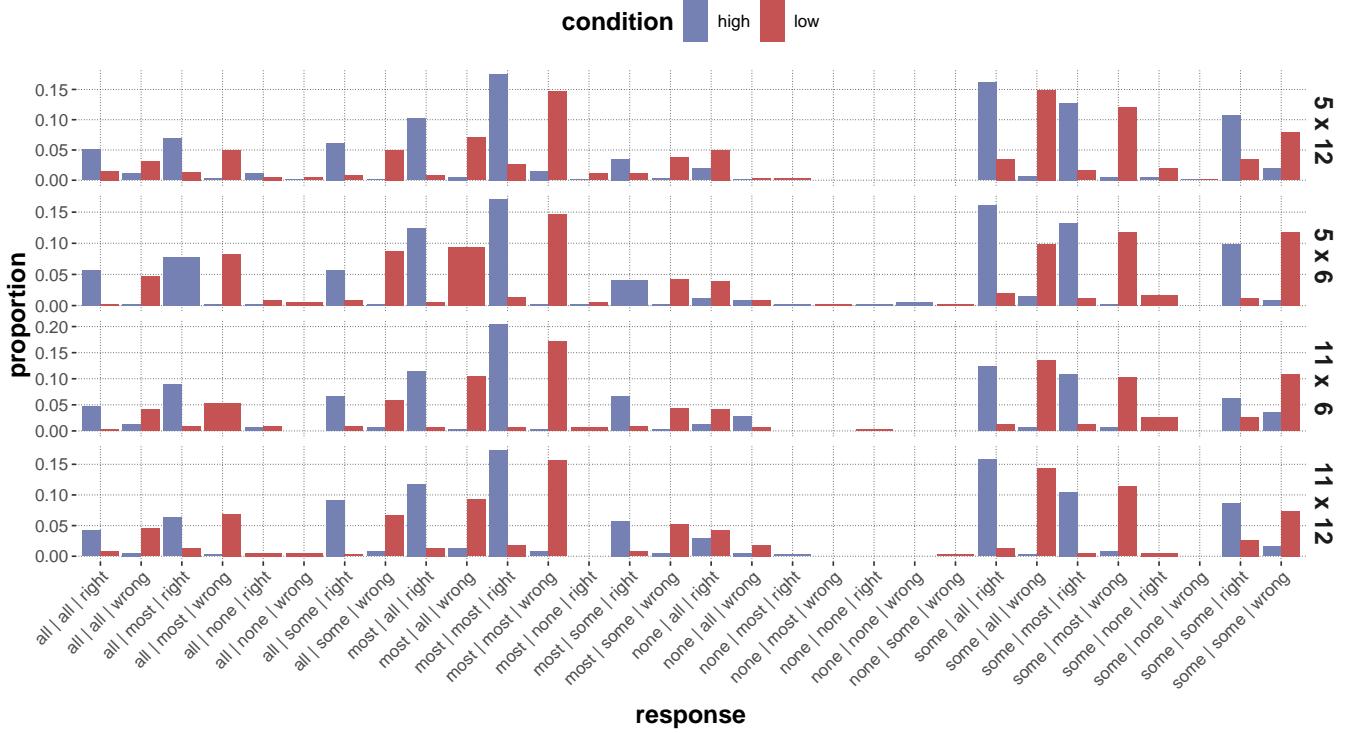


Figure 1: Proportion of sentences generated in Experiment 1. The bars show the observed proportions of sentences generated by the participants using the template in (2). Each row shows the results for a different matrix size.

the utterance u is more likely to be true (literally) under hypothesis H_0 than under a competing (alternative) hypothesis H_1 :

$$\text{argstr}(u, H_0, H_1) = \log \frac{P(\llbracket u \rrbracket | H_0)}{P(\llbracket u \rrbracket | H_1)} \quad (4)$$

where $P(\llbracket u \rrbracket | H)$ is the probability that utterance u is true given hypothesis H . For and H asserting that each student answers each question correctly with probability γ_H :

$$P(\llbracket u \rrbracket | H) = \sum_{s \in S} P(\llbracket u \rrbracket | s)P(s | \gamma_H) = \sum_{s \in S} \llbracket u \rrbracket^s P(s | \gamma_H) \quad (5)$$

Pragmatic argstrength RSA

The second model is the same as the log likelihood ratio argstrength RSA model described above, except for the utility function which uses a different measure of argumentative strength: [MF: explain notation w (world states)]

$$\text{argstr}(u) = \log \frac{P_S(u | H_0)}{P_S(u | H_1)} = \log \frac{\sum_{w \in W} P_S(u | w)P(w | H_0)}{\sum_{w \in W} P_S(u | w)P(w | H_1)}$$

where P_S is defined above in Equation (1).

This pragmatic speaker is different from the real pragmatic speaker, because they do not include argumentative strength in their calculation, corresponding to the idea that the real speaker is considering a listener who doesn't take into account the speaker's argumentative goal. Conceptually, a pragmatic

speaker that uses this type of argumentative strength exploits the listener's assumption of cooperativeness.

Pragmatic speakers intuitively differ for signals that have a stronger alternatives. In the experiment, these are 'some' (alternative: 'most') and 'most' (alternative 'all').

Maximin argstrength RSA

The third model we fit is meant to capture the intuition that, rather than calculating argumentative strength by marginalizing across *all* the states compatible with the utterance, participants only consider the argumentatively weakest one. More formally, for each utterance u participants might consider the following argumentative strength:

$$\text{maximin-argstr}(u) = \min_{s \in S} \log \frac{p(s | \llbracket u \rrbracket, \gamma = 0.85)}{p(s | \llbracket u \rrbracket, \gamma = 0.15)} \quad (6)$$

$$p(s | \llbracket u \rrbracket, \gamma) \propto \prod_{k \in s} \binom{12}{k} \gamma^k (1 - \gamma)^{12-k} \quad (7)$$

where $\gamma = 0.85$ encodes H_0 and $\gamma = 0.15$ encodes H_1 . Other than the calculation of the argumentative strength, the model is identical to the model presented in Section .

Maximin- and lr-argstrength differ not only in terms of the absolute argstrength of signals, in some cases also in their ranks. For instance, in the high condition 'most|none|wrong' is better than 'all|most|right' for lr, but the other way around for maximin.

Weak signals for minimax argstrength can be bad arguments across the board. The most striking is 'some|some|right', which is always quite a bad signal argumentatively (in absolute terms). However, given a specific situation a bad signal can still be the best option. For instance, 'most|most|right' in the high condition is actually a slightly bad signal (with sigma=0.85), but often the best available.

Instead of calculating the maximin for a specific value of γ , we could also calculate it assuming that the agent is arguing *for* the value of γ being greater than a particular threshold and *against* the value of γ being lower than that threshold. This could be formalized as:

$$\text{maximin-argstrength}(u) = \min_{s \in S} \log \frac{p(s \mid [[u]] = 1, \phi \geq \theta)}{p(s \mid [[u]] = 1, \phi < \theta)}$$

There is an intuition: what makes 'most|most|right' a better argument than 'some|all|right' is that the former excludes some particularly bad cases (namely, the case where one participant answered correctly and all the other ones got all of them wrong).

An intuitive explanation for this intuition is that the speaker is thinking 'conservatively' about the listener. The imagined listener does not consider all possible observations, but rather guesses the situation (among the ones compatible with the utterance) that lends the least support to the argued-for state (e.g. γ parameter or betabinomial parameters).

NOTE: I am saying 'conservatively imagined listener' rather than 'conservative listener'. It's not that the listener is conservative - the listener might not even know what the speaker is arguing for. Rather, the speaker is thinking: 'the listener will have to guess *at least this*, and so this utterance will be *at least this strong*'.

In other words: the speaker wants to maximise the minimal observation-wise argstrength of the signal. In other words: for each signal, the speaker considers all observations compatible with the signal, each of which individually has a certain 'strength' wrt to the argument. Then, they calculate argstrength as the strength of the *least convincing* observation.

Consider the utterance 'all|some|right':

- If we are in the high condition, the conservatively imagined listener will guess '3|3|3|3|3',
 - If we are in the low condition, the conservatively imagined listener will guess '12|12|12|12'

From a computational point of view, in one sense it still requires the agent to calculate for all possible states, but it seems likely that there are ways of pruning the space of considered states in any given situation (based on a partial order of maximin-argstrength).

In practice, for the high condition:

$$\text{maximin-argstrength}(u) = \min_{s \in S} \log \frac{p(s \mid [[u]] = 1, \phi_+)}{p(s \mid [[u]] = 1, 1 - \phi_+)}$$

and for the low condition same but with ϕ_- .

Conceptually: it's the minimum state-wise Bayes factor (the evidence for the worst-case scenario).

In slogan form: an argumentation chain is only as strong as its weakest link.

Model-free argstrength

The agent might also try to maximize (in the high condition) or minimize (in the low condition) the expected total number of correct answers across all students given the utterance (recall that s is a list of the number of correct answers for each student):

$$\text{totalcorrect}(u) = |[[u]]|^{-1} \sum_{s \in [[u]]} \sum_{i \in s} i$$

In words, the argumentative strength encodes the expected total number of right answers, which is to be (soft)maximised in the high condition and (soft)minimized in the low condition.

Results

For each of the measures of argumentative strength above, we implement and fit two models:

1. A population model with completely pooled α and β .
 2. A hierarchical model with by-participant α , β (For the pragmatic argstrength model, we pool the value of α for the calculation of the argumentative strength and of the utility).

The parameter encoding the cost for ‘none’ is always pooled.

Figure 2 shows the results of loo-based model comparison. By expected log-likelihood under leave-one-out cross-validation, the best model is the hierarchical non-parametric model. However, the second best model, the hierarchical maximin model, is not significantly worse under a simple z -test [MF: add reference Lambert].

For several of the models, I calculated the posterior predictive p-values for all the models (see Bayesian p-value section above). In all cases they were quite close to 0.5, indicating good compatibility of the data with the fitted posterior. However, note that posterior predictive p-values are generally not uniformly distributed in [0,1] and hierarchical models pose a challenge (see e.g. this paper ([clickable](#))).

It is also instructive to have a look at the *pointwise* posterior predictive p-values. - Roughly, they quantify the compatibility of the model with each individual datapoint. - They answer the question "What is the probability that the posterior prediction for this specific datapoint has an equal or higher discrepancy than the actually observed one?" - In this case, the measure of discrepancy is just the likelihood. - The odd thing is that many datapoints have pointwise posterior predictive p-value of 1. - This means that the real datapoint has a probability greater than or equal to the one sampled from the posterior *for all trace samples*. - Note however that in the case of a categorical distribution, the 'equal to' can do a lot of work. - In particular, if the posterior samples the same factor as the original data, they will always have the same likelihood. -

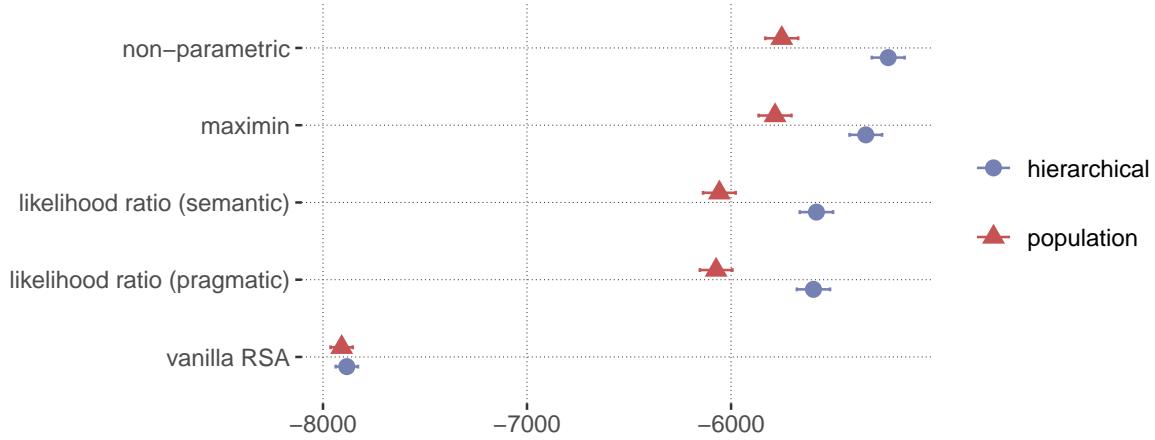


Figure 2: Results of model comparison based on the full data set. For each model, shapes indicate the expected log-probability mass from leave-one-out cross validation, with error bars showing the standard error of these estimates. The y-axis lists the different types of models, ordered by ascending goodness-of-fit. The shapes and colors indicate the method of model fitting: with or without hierarchical structure.

This will happen if the model across the trace gives a high probability to the specific datapoint that was observed. - So it's not so worrying. - Pointwise predictive values close to 1 however indicate low compatibility of the model and the data.

lr-argstrength Looking at the datapoints that the different models fail to predict reveals some illuminating patterns. For the lr-argstrength model, we look at the signals that were produced most often by participants for a certain observation but which are not on the Pareto frontier of informativity/lr-argstrength, which means that no value of β could account of them. These are:

- Observation 4 ([12, 12, 9, 0, 0]), high condition:
 - Optimal signal: 'some|all|right' (5 times)
 - Most common signal: 'most|most|right'
- Observation 10 ([12, 12, 3, 0, 0]), low condition:
 - Optimal signal: 'some|all|wrong' (7 times)
 - Most common signal: 'most|most|wrong' (15 times)
- Observation 15 ([12, 12, 9, 3, 3]), high condition:
 - Optimal signal: 'some|all|right' (14 times)
 - Most common signal: 'most|most|right' (20 times)
- Observation 16 ([12, 12, 9, 9, 9]), high condition:
 - Optimal signal: 'some|all|right' (8 times)
 - Most common signal: 'most|most|right' (11 times)
- Observation 19 ([9, 9, 3, 0, 0]), low condition:
 - Optimal signal: 'some|all|wrong' (7 times)
 - Most common signal: 'most|most|wrong' (11 times)

(Note: there are other ways in which the predictions differ from observations, e.g. often signals on the Pareto frontier

are not produced as often as one would expect.) They all involve 'most|most|' being used when 'some|all|' is predicted to be a better signal by the model!

maximin argstrength Maximin argstrength makes better predictions than lr-argstrength in several instances:

- '1.png' and '3.png', low condition: The most produced signal ('most|all|wrong') is the best for maximin but not for BF (although still close to best).
- '2.png', low condition and '13.png' high condition: As I mentioned above, maximin makes sense of why so many different signals were produced.
- '4.png' and '15.png', high condition: The most produced signal ('most|most|right') is best for maximin but not for BF. Neither argstrength explains why 'some|none|right' was not produced more, but the cost for 'none' might.
- '8.png', high condition: The most produced signal ('all|most|right') is best for maximin but not for BF.
- '9.png' and '14.png', low condition: The most produced signal ('all|most|wrong') is best for maximin but not for BF.
- '10.png' and '19.png', low condition: The most produced signal ('most|most|wrong') is best for maximin but not for BF.

Maximin argstrength also predicts that a lot of the 'bad' signals (given the condition) are equally bad, e.g., 'some|some|wrong' and 'all|all|wrong' (in the high condition). On the other hand, with lr-argstrength 'all|all|wrong' is a bad signal and but 'some|some|wrong' is neutral (neither good nor bad). This might seem like a weird prediction of maximin at first, but interestingly it makes sense of a part of the data that

was puzzling before. When participants are forced to describe a 'bad' situation in the high condition, they don't just produce the 'winner' according to BF-argstrength, but rather produce a larger variety of signals than for other observations.

On the other hand, maximin argstrength underpredicts the frequency of 'some|some|wrong'. The truth might lie in the middle. Participants might consider some but not all of the observations for the signal. This behaviour can be rationalized: If the speaker is perfectly rational but reasons about an imperfect listener, they might only focus on the few states that the listener is likely to consider given a signal.

In some cases, maximin-argstrength makes worse predictions than lr-argstrength. - '5.png', high condition: 'most|all|right' is the most produced signal, and it is the best under lr but not maximin (which instead predicts 'all|most|right' to be argumentatively better). '14.png', high condition: a signal that is predicted to be good by maximin ('none|all|wrong') is not produced much. This can be explained by the cost for 'none'.

Discussion

- There are a couple of implementation decisions for this model (and all the following ones):
- Whether to consider all *possible* observations (every way that 5 students can answer 12 questions) or just the 20 observations that were presented in the model. I call the former the 'Michael method' in the code and the latter is the one I am implementing for simplicity, because it allows me to simply manipulate arrays and keep everything vectorized.
- Whether to give a small fuzzy truth value to utterances that are literally false or treat them as just having truth value 0. The former is the method I use here (though see the calculation below in the model with Solt's 'most'), the latter is the one used in the original Greta implementation.

An interesting empirical observation is that participants prefer 'most|most' over 'some|all'. A first explanation could be that 'most' has a stronger interpretation than assumed in the lr-argstrength model. Then, 'some' becomes weaker than 'most' and 'some students got all answers wrong' could be argumentatively weaker than 'most students got most of the answers wrong'. We played with several variations of this idea that increases the argumentative strength of utterances involving 'most':

- Calculate with pragmatic argumentative strength using S1.
- Give a stronger literal meaning (based on Solt's account of 'most')
- Use a 'manually' pragmatically enriched sense of 'most' (based on my paper with Jakub).

None of these variations were better than the lr-argstrength.

Another option is that high enough values of γ make 'most|most|right' argumentatively stronger than 'some|all|right'. We tested a model where (although with a prior that pushed in that direction) γ is estimated to be very

close to 1 (indeed, so close that most|most| is argumentatively stronger than some|all).

A third option lifts the assumption in the originl model of a single binomial parameter for all students. If different students have different binomial p parameters (e.g. structured hierarchically), knowing that one student performed very well does not tell you as much as knowing that many students performed reasonably well, as the former might be a fluke, making 'some|all' weaker. If each student is still described by a Binomial parameter, but these parameters are distributed as a Beta distribution, the resulting model of exam result probabilities follows a Beta-Binomial distribution. We leave a more detailed analysis of this option to future work.

Some observations remain puzzling. - '17.png': low condition: Neither model can really make sense of why 'most|all|wrong' is by far the most produced signal. lr-argstrength predicts it should be 'most|none|right' and maximin that it should be 'all|most|wrong'. - '16.png', high condition: Really strange, because it seems like there's no way of cashing out argumentative strength where 'most|most|right' is better than 'all|most|right' in the high condition.

References

- Anscombe, J.-C., & Ducrot, O. (1983). *L'argumentation dans la langue*. Mardaga.
- Blutner, R. (2000). Some aspects of optimality in natural language interpretation. *Journal of Semantics*, 17, 189–216. <https://doi.org/10.1093/jos/17.3.189>
- Cummins, C., & Franke, M. (2021). Rational interpretation of numerical quantity in argumentative contexts. *Frontiers in Communication*, 6, 89. <https://doi.org/10.3389/fcomm.2021.662027>
- Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, 9(1), 519–540. <https://doi.org/10.1146/annurev-linguistics-031220-010811>
- Enfield, N. J. (2024). *Language vs. Reality: Why Language is Good for Lawyers and Bad for Scientists*. MIT Press.
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), 998. <https://doi.org/10.1126/science.1218633>
- Franke, M., & Degen, J. (2023). The softmax function: Properties, motivation, and interpretation. <https://doi.org/10.31234/osf.io/vsw47>
- Good, I. J. (1950). *Probability and the weighing of evidence*. Griffin.
- Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling lanuage understanding as social cognition. *Topics in Cognitive Science*, 5, 173–184. <https://doi.org/10.1111/tops.12007>
- Grice, P. H. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, vol. 3, speech acts* (pp. 41–58). Academic Press.
- Grusdt, B., Lassiter, D., & Franke, M. (2022). Probabilistic modeling of rational communication with conditionals. *Semantics & Pragmatics*, 15. <https://doi.org/10.3765/sp.15.13>

- Herbstritt, M., & Franke, M. (2019). Complex probability expressions & high-order uncertainty: Compositional semantics, probabilistic pragmatics & experimental data. *Cognition*, 186, 50–71. <https://doi.org/10.1016/j.cognition.2018.11.013>
- Lassiter, D., & Goodman, N. D. (2017). Adjectival vagueness in a bayesian model of interpretation. *Synthese*, 194(10), 3801–3836. <https://doi.org/10.1007/s11229-015-0786-1>
- Macuch Silva, V., Lorson, A., Franke, M., Cummins, C., & Winter, B. (2024). Strategic use of english quantifiers in the reporting of quantitative information. *Discourse Processes*, 61(10), 498–523. <https://doi.org/10.1080/0163853X.2024.2413311>
- Mercier, H., & Sperber, D. (2011). Why do humans reason? arguments from an argumentative theory. *Behavioral and Brain Sciences*, 34, 57–111. <https://doi.org/doi:10.1017/S0140525X10000968>
- Merin, A. (1997). *If all our arguments had to be conclusive, there would be few of them* [Arbeitspapiere des SFB 340, Bericht Nr. 101]. <http://www.semanticsarchive.net/Archive/jVkZDI3M/101.pdf>
- Parikh, P. (1991). Communication and strategic inference. *Linguistics and Philosophy*, 473–514(14), 3.
- Rogers, T., Zeckhauser, R., Gino, F., Norton, M. I., & Schweitzer, M. E. (2017). Artful paltering: The risks and rewards of using truthful statements to mislead others. *Journal of Personality and Social Psychology*, 112(3), 456–473.
- Scontras, G., Tessler, M. H., & Franke, M. (2021). A practical introduction to the rational speech act modeling framework. <https://doi.org/10.48550/ARXIV.2105.09867>
- van Rooij, R. (2004). Cooperative versus argumentative communication. *Philosophia Scientiae*, 8(2), 195–209.
- van Tiel, B., Franke, M., & Sauerland, U. (2021). Probabilistic pragmatics explains gradience and focality in natural language quantification. *Proceedings of the National Academy of Sciences*, 118. <https://doi.org/10.1073/pnas.2005453118>
- Tessler, M. H., & Goodman, N. D. (2019). The language of generalization. *Psychological Science*, 126(3), 395–436. <https://doi.org/10.1037/rev0000142>
- Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2020). Polite speech emerges from competing social goals. *Open Mind: Discoveries in Cognitive Science*, 4, 71–87. https://doi.org/10.1162/opmi_a_00035