What guides speakers' [FC: is 'speakers' needed?] utterance choice in argumentative language use?

Anonymous CogSci submission

"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

Theories of language use in the Gricean tradition, and related formalizations such as the Rational Speech Act (RSA) modeling framework, have focused on information exchange between cooperative interlocutors. More recent work has emphasized the importance of other, potentially less-cooperative functions of language: persuasion, argumentation, manipulation. In this work, we bridge these two traditions by developing a model of argumentative utterance choice in the RSA framework. We propose several formalization of the argumentative strength of a signal. We then measure how well they capture behaviour in a simple utterance production task where participants were assigned an explicit argumentative goal. We show that a popular regimentation of argumentative strength fails to capture some important patterns in the data, and discuss some alternatives. Although our models constitute a substantial improvement over a baseline model, we also find patterns in the data that remain unexplained, suggesting that more work is needed.[FC: not the most inspiring abstract but it's a start]

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 (Grice, 1975). Formal models of pragmatic language use consequently often rely on formal notions of informativity —such as supplied by formal logic or information theory— to 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 (e.g., Parikh, 1991; Blutner, 2000; Frank & Goodman, 2012). Yet, some authors argue that the purpose of communication may be much more about persuasion, argumentation, manipulation, and opinion-negotiation than about pure, objective—indeed, robot-like—information-sharing (e.g., Anscombre & Ducrot, 1983; Mercier & Sperber, 2011; Enfield, 2024). To give an example of selective truth-telling, or *paltering* (Rogers et al., 2017), the outcome of an exam could be truthfully described with roughly similar information value with different expressions, framing the exam as easy (1a) or difficult (1b).

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

While formal notions of informativity are ready at hand, what is 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, as in the above example. 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; Winterstein, 2012) suggests a formalization of argumentative strength in terms of *log-likelihood ratios*, based on early work 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 alternative formalization 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 (Rogers et al., 2017). We study cases like (1), which are 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 (Frank & Goodman, 2012; Franke & Jäger, 2016; Degen, 2023). We then 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 $U(u, s)$:

$$P_S(u \mid s) \propto \exp\left(\alpha U(u, s)\right) \ . \tag{1}$$

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; Franke & Jäger, 2016; 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:

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

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:

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

For a binary semantics (as assumed here), the utility function factors into three well-known aspects of pragmatic language generation, namely that the utterance be true, informative and economical (Scontras et al., 2021):

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

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 (Herbstritt & Franke, 2019). Yet, some phenomena seem to require more elaborate utility functions. For example, models have been explored which incorporate additional utility components related to politeness (Yoon et al., 2020) or opinion (Achimova et al., 2025). Here, we take a similar approach to modelling the utility trade-off between truth and informatively, on the one hand, and, on the other, making an argument in favor of a position or hypothesis H_0, as opposed to the competing position or hypothesis H_1. The form of the utility functions we consider here is:

$$U(u, s, H_0, H_1) = \tag{3}$$
$$\underbrace{\beta \, \log P_{L_0}(s \mid [[u]])}_{\text{truth \& informativity}} + \underbrace{(1 - \beta) \, \text{argstr}(u, H_0, H_1)}_{\text{argumentative strength}} - \underbrace{\text{cost}(u)}_{\text{economy}}$$

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).

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.

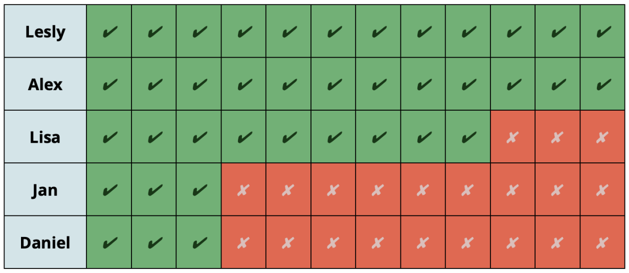
Experiment

To test whether and how speakers choose expressions for purposes of argumentative framing, 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 individual selection of (i) an outer quantifier, (ii) an inner quantifier, and (iii) an adjective, to complete the sentence frame in (2), where OQ and IQ ∈ { None, Some, Most, All }, and ADJ ∈ { right, wrong }.

(2) OQ of the students got IQ of the questions ADJ

Participants. A total of $N = 201$ adult participants with English as their first language were recruited via Prolific and were paid £1.5 (hourly wage approximately £9).

Materials. The results of high-school exams were presented visually in form of matrices, as shown in Figure 1. The rows of



Lesly	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Alex	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Lisa	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗
Jan	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
Daniel	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗

Figure 1: Example of an experimental stimulus showing one set of results for an exam with 12 questions and 6 students.

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 where 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 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).[1]

Procedure. The experiment started with an explanation of the displays and the task. As a within-subjects manipulation (high vs. low framing condition), participants were asked to truthfully describe the exam results as either favorable or unfavorable (students did well/poorly, exam was easy/hard). Each participant saw each of the 20 instances of the same matrix size in completely randomized order. Each participant saw 10 trials in the high and low framing condition each, randomly assigned for each run.

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.[2] 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 2 shows the proportions of sentences which participants generated as descriptions for the exam results. The plot differentiates the two framing conditions (*high* vs. *low*, indicated by color), and the four different shapes of the matrices that represented the exam results (different rows in the figure). The choice distributions seem to differ slightly between results matrices ($\chi^2 \approx 216.8$, df $= 168$, $p < 0.006$). More importantly, the *high* vs. *low* framing induced significantly different distributions over descriptions ($\chi^2 \approx 2300.7$, df $= 103$, $p < 2e^{-16}$). The latter result suggests that our manipulation worked and that participants are indeed able to adapt their description choices to the strategic argumentative framing the context demanded.

Statistical Analysis

We perform a Bayesian analysis of the data produced by the experiment, using the RSA speaker defined in Eq. 1 as a generative model of participants' production behaviour. A full implementation of argumentative strength (Eq. 3) requires the specification of one hypothesis that the speaker argues for (H_0) and one that they argue against (H_1), as well as a functional form that measures how strongly a specific utterance supports H_0 against H_1.

Hypotheses for argstrength calculation

We model each hypothesis H as saying that all students in the class have a certain probability γ_H of getting each answer correct, making the simplifying assumption that there is no variation in skill across students and no variation in probability of getting different answers correct. The probability of observing the full results from an exam is then proportional to the product of the Binomial probability of each student's results: [MF: check: why $\propto$ here? isn't it actually =?][FC: We decided at some point to do everything in terms of the states in the experiment rather than the full combinatorial statespace. I spent time also writing functions to compute argstrengths with the full state space. If we do the latter, the challenge is the sum across utterance-compatible states when calculating e.g., P([[u]]|h). I found a way to do it for all argstrengths except prag-argstrength—it is computationlly infeasible to recalculate the speaker probabilities for each sample with a different α for all those ss in the large matrix conditions.]

$$P(s \mid \gamma_H) \propto \prod_{k \in s} \binom{N}{k} \gamma_H^k (1 - \gamma_H)^{N-k} \qquad (4)$$

where the exam result s is a list of numbers of correct answers for each student, and N is the number of questions.

By experimental design, in the high framing condition the speaker is choosing an utterance to support the hypothesis that students had a high probability of answering correctly rather than a low probability, and the opposite is true in the low framing condition. Consequently, for the models below, the assumed hypotheses are: [MF: elaborate on why exactly these numbers?][FC: They were already there when I joined the project, I spent time playing with them, e.g., treating them as

[1] 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).

[2] The preregistration files are available at https://osf.io/2juzy/overview?view_only=82547823039c4eb9aa2500135ddb2f6f.

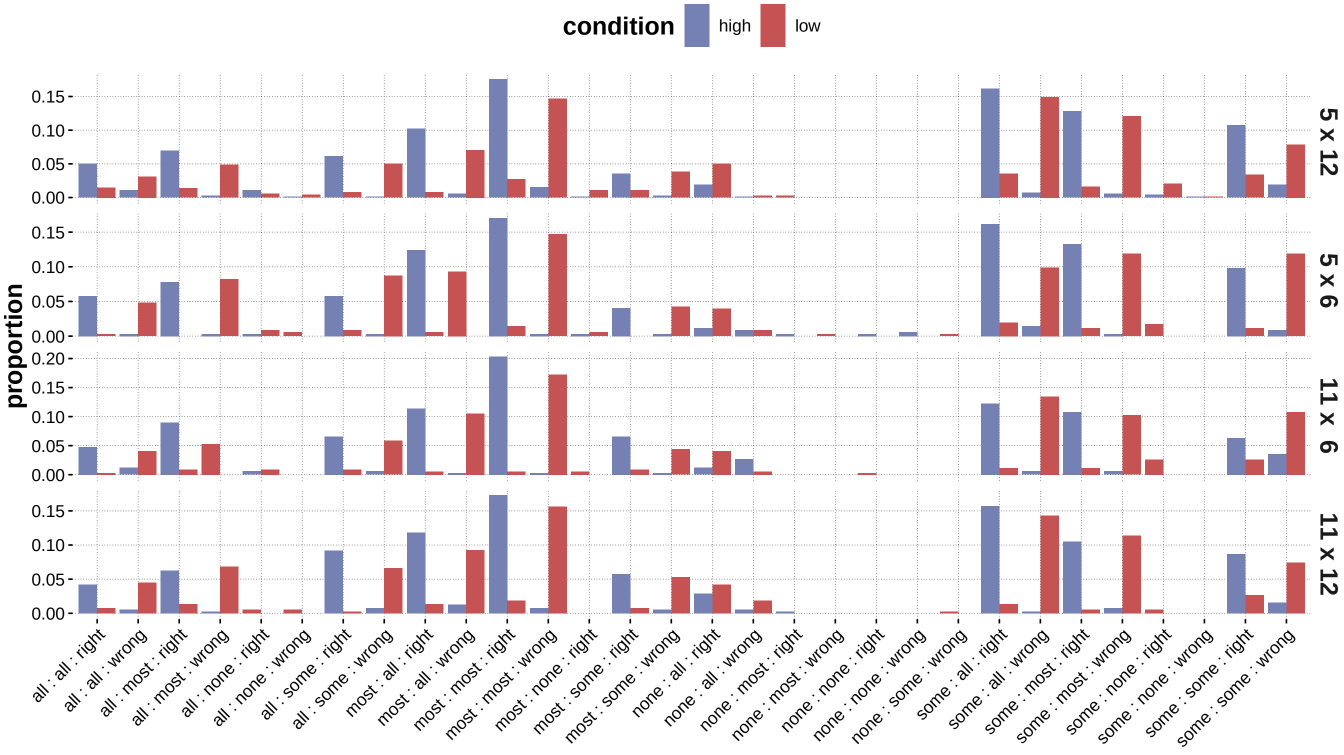


Figure 2: Proportion of sentences generated in Experiment 1. The bars show the observed proportions of sentences generated by the participants using the template in (2). Each row shows the results for a different matrix size.

hidden variables, but the model was underspecified. Is there something to say beyond the previous sentence in the text?]

$$\gamma_{H_0} = 0.85 \quad \gamma_{H_1} = 0.15 \quad \text{for high frame condition.} \tag{5}$$
$$\gamma_{H_0} = 0.15 \quad \gamma_{H_1} = 0.85 \quad \text{for low frame condition.}$$

Functional form of the argstrength component

Having defined the two hypotheses, we next turn to various ways of specifying the functional form of the argumentative strength of utterances.

Log likelihood ratio argstrength. The starting point is the notion of *weight of evidence*, which has been introduced in the previous literature as a formal notion of argumentative strength (e.g., Merin, 1997; van Rooij, 2004; Winterstein, 2012). Concretely, we consider the degree to which the utterance u is more likely to be true (literally) under hypothesis H_0 than under a competing (alternative) hypothesis H_1:

$$\text{lr-argstr}(u, H_0, H_1) = \log \frac{P([\![u]\!] \mid H_0)}{P([\![u]\!] \mid H_1)} \tag{6}$$

where $P([\![u]\!] \mid H)$ is the probability that utterance u is *true* (rather than, e.g., *produced*) given hypothesis H. $P([\![u]\!] \mid H)$ can be calculated based on the literal semantics, with $P(s \mid \gamma_H)$ defined in Eq. (4), as:

$$P([\![u]\!] \mid H) = \sum_{s \in [\![u]\!]} P(s \mid \gamma_H) \tag{7}$$

Pragmatic argstrength. While the previous definition defines argumentative strength in terms of the evidence ratio for u being literally true, it is also conceivable that an utterance's argumentative strength is intuitively assessed not based on its literal meaning but based on its pragmatically enriched meaning. We therefore also consider a notion of pragmatic argumentative strength based on the pragmatic speaker function P_S as defined above in Equations (1) and (2). We first quantify the probability of observing u under said pragmatic speaker behavior if H is true:

$$P_S(u \mid H) = \sum_{s \in S} P_S(u \mid s)\, P(s \mid \gamma_H) \tag{8}$$

We then define pragmatic argumentative strength as:

$$\text{prag-argstr}(u, H_0, H_1) = \log \frac{P_S(u \mid H_0)}{P_S(u \mid H_1)} \tag{9}$$

Maximin argstrength. The third type of argstrength captures the intuition that, rather than calculating argumentative strength by marginalizing across *all* the states compatible with the utterance u, participants only consider the argumentatively weakest state compatible with u:

$$\text{maximin-argstr}(u, H_0, H_1) = \min_{s \in [\![u]\!]} \log \frac{p(s \mid \gamma_{H_0})}{p(s \mid \gamma_{H_1})} \tag{10}$$

A speaker might use maximin-argstr when considering the worst case scenario for a listener who guesses a single one of the states compatible with the utterance. Maximin-argstr can also be motivated on computational grounds, since it only requires the single argumentatively weakest state, which in some cases can be found efficiently.

Model-free argstrength. Instead of assessing argumentative strength in terms of a precise probabilistic relation between hypotheses and utterance-compatible observations, participants might use a heuristic higher-level feature of observations (a suitable *test statistic*). One such feature for our experimental setup is the mean number of right answers across utterance-compatible observations:

$$\mathbb{E}_{\text{right}}(u) = |[[u]]|^{-1} \sum_{s \in [[u]]} \sum_{i \in s} i \tag{11}$$

The argumentative speaker might choose an utterance to maximise this (centered) quantity in the high frame condition, and to minimize it in the low frame condition:

$$\text{mf-argstrength}(u) = \begin{cases} \mathbb{E}_{\text{right}}(u) - \mu_{\mathbb{E}_{\text{right}}} & \text{in high frame} \\ \mu_{\mathbb{E}_{\text{right}}} - \mathbb{E}_{\text{right}}(u) & \text{in low frame} \end{cases} \tag{12}$$

where $\mu_{\mathbb{E}_{\text{right}}}$ is the average value of $\mathbb{E}_{\text{right}}(u)$ across the possible utterances u. We do not model the process of finding a heuristic given an argumentative goal, and therefore this argstrength does not formally depend on the hypotheses defined in Equation (5). [MF: It's not clear from the text (to me) what the $\mu_{\mathbb{E}}$ means exactly. I can guess but I cannot be sure. More importantly, I wonder if this summand doesn't actually cancel out? Is it strictly necessary (apart from stabilizing the inference)?] [FC: Added the definition. The summand is just to give it two neat properties at the same time (1) mf-argstrength(u) in the high condition = - mf-argstrength(u) in the low condition, which is true of the other argstrengths, and (2) mf-argstrength in the low condition can also be written as $\mathbb{E}_{\text{wrong}}(u) - \mu_{\mathbb{E}_{\text{wrong}}}$.] [MF: Even more importantly, I always thought of this approach as a shortcut to a literal listener's posterior of the true rate parameter (the iid probability of getting a single question right). The MLE we use here approximates the Bayesian mean well enough for larger samples sizes (which we might already have), so this is not totally "model free" necessarily; it define argumentative strength in terms of a "valuation function" on beliefs of the literal listener about a latent world-variable after hearing u.][FC: Yeah, I also share this intuition, but note that $\mathbb{E}_{\text{wrong}}(u)$ can just get bigger with bigger matrix sizes, unlike the MLE, so assuming the same β across conditions for a given participant, it is doing something different from just a prob estimate. I am not sure how much detail we should get into here.]

Statistical models

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 and β.

The parameter encoding the cost for 'none' is always pooled. For the pragmatic argstrength model, we use the same value of α for the calculation of the argumentative strength and for the participant's utility computation. Parameter recovery simulations on synthesis data for the pooled models and array size 5×12 show that the α, β, and cost parameters can be recovered accurately from <100 samples for all five models.

Results

Figure 3 shows the results of loo-based model comparison of all models trained on the dataset. We can roughly distinguish three levels. First, both the pooled and hierarchical vanilla RSA model have a comparatively poor predictive accuracy. Second, the semantic and pragmatic lr-argstrength models improve over the vanilla RSA model, and perform similarly to each other both for the pooled and the hierarchical cases. Third, the maximin and non-parametric models offer the best predictive accuracy. By expected log-likelihood under leave-one-out cross-validation, the best model is the hierarchical non-parametric model. The posterior predictive p-value of this model with loglikelihood as the test statistic is 0.407, meaning that the data does not exclude this as a possible model. However, the second best model, the hierarchical maximin model, is not significantly worse under a simple z-test ($p \approx 0.97$), as suggested by (Lambert, 2018). The most important high-level conclusion from these results is that including *some* quantitative notion of argumentative strength seems required; the vanilla RSA model is not enough—and of the set of candidate notions investigated here *all* yield very substantial improvements to predictive accuracy.

Fitting only on the data of each matrix-size condition independently and comparing the resulting models also reveals interesting differences. [MF: Reader may wonder whether (i) models were trained on all conditions and evaluated on pairwise comparisons, or (ii) trained on only data from the pair to be compared.][FC: Clarified now. Here I am speaking to people (like me) who are worried about keeping the same β across conditions for each participant, becuase it's not clear the argstrengths scale well for different matrix sizes.] For instance, the maximin model is better than the model-free one for the 5×12 condition, while the opposite is true in the 11×6 and 5×6 conditions (the models are not credibly different in the 11× 12 condition). Since each model encodes an account of argumentative strength that we do not expect to interact with the matrix size condition, these differences in performance might reflect a failure of some linking modeling assumption, e.g., the Binomial likelihood for student results.

The posterior predictive checks for each model also show systematic patterns in how they correctly fit or depart from the data, which suggests ways in which the corresponding accounts of argumentative strength succeed or fail in this experimental setup. We briefly discuss these qualitative patterns of success and failure for each model. [MF: It would be great to be able to say whether the best fitting model is in some absolute sense good enough. We could compute a posterior

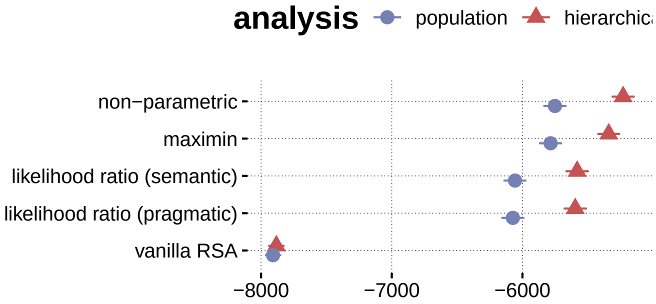


Figure 3: Results of model comparison based on the full data set. For each model, shapes indicate the expected log-probability mass from leave-one-out cross validation, with error bars showing the standard error of these estimates. The y-axis lists the different types of models, ordered by ascending goodness-of-fit. The shapes and colors indicate the method of model fitting: with or without hierarchical structure.[FC: hierarchic in legend feels unusual to me. Also note 'non-parametric', which is the name in the codebase but we decided to call 'model-free' instead]

predictive p-value, for example.][FC: added this in first paragraph]

Vanilla RSA Since this model does not take into account the argumentative goal of participants, it predicts a higher-than-observed frequency of informative but argumentatively weak signals, such as 'all : all : wrong' in the high framing condition. It also predicts a lower-than-observed frequency for signals that can be exploited for paltering, such as 'some : some : right' in the high condition when most students got all answers wrong.

lr-argstrength The simple lr-argstrength and its pragmatic variation make very similar predictions. They capture many patterns of utterance choice in the data relating to argumentativity, e.g., participants' preference for 'some : most : right' over 'most : all : wrong' in the high frame condition when both utterances are true. However, in the two 12-questions exam conditions, this argstrength predicts a much stronger preference of 'some : all : right' over 'most : most : right' than we find in the data (in the high frame condition, when both signals are true). This is significant for its overall predictive performance, because these were the most frequently chosen utterances for those conditions.

Maximin argstrength Intuitively, what makes 'most : most : right' a better argument than 'some : all : right' in the high frame condition is that the former excludes some particularly bad cases, i.e., where one participant answered correctly and all the other ones got all of them wrong. This is captured by the maximin argstrength, which can fit this pattern in the data better than lr-argstrength.

In each condition, maximin argstrength also predicts a long tail of argumentatively weak signals with equal argstrength, e.g., 'some : some : wrong', 'all : all : wrong', and several others (in the high condition). This is a surprising prediction, but it makes sense of a pattern in the data: when participants have to describe poor exam results in the high condition, they produce a larger variety of signals than for other observations.

Model free argstrength In many cases, mf-argstrength makes similar predictions as maximin-argstrength. One systematic way the two depart is that the mf-argstrength model systematically postdicts higher frequency than the maximin-argstrength model for 'all|some|right' in the high condition and 'all|some|wrong' in the low condition. Another is that mf-argstrength systematically postdicts lower frequency than maximin-argstrength for 'some : most : wrong' in the low condition and 'some : most : right' in the high condition. Which of the two models is more accurate depends on the matrix array size, explaining some of the difference in performance between models trained separately on the matrix array conditions.

Overall All models predict a lower-than-observed frequency for '(some,most,all) : (some,most) : wrong' in the low condition and '(some,most,all) : (some,most) : right' in the high condition (e.g., for state $[1, 1, 1, 0, 0]$). In these cases, models distribute their predictions across a larger set of utterance which uses adjective 'wrong' in the high framing condition and 'false' in the low framing condition. It is possible that participants might avoid adjectives that emphasize a polarity opposite to the one for which they are arguing. However, all measures of argumentative strengths we defined are truth-functional, and therefore cannot capture the connotation of conveying the same content in terms of 'right' or 'false'.

One particularly striking observation is that participants often chose 'most : most' when 'all : most' would have been both more informative and argumentatively stronger according to all models. All models failed to capture this pattern, and indeed it is difficult to explain it as a rational strategic choice. Future work could look into other possible explanations, such as processing ease.

Finally, participants chose 'none : all : right' more often than predicted by any model, for both frame condition. For example, in the high frame condition, it is chosen for exam results where all students got all answers wrong.[FC: Not sure what more to say here, can also leave it out]

Discussion

In this paper, we proposed several different formalizations of the notion of argumentative strength and evaluated them with data from a simple utterance choice experimental setup. We found that the traditional notion of *log-likelihood ratio* argumentative strength captures the data better than a model that completely ignored argumentative goals, but worse than other notions. Strikingly, the models with the closest fit to the data are simpler than the log-likelihood ratio model, in that they consider fewer utterance-compatible states (maximin argstrength) or use a more context-specific approximation (model-free).

We explored a natural selection of models, but many other options could be considered. For instance, full lr-argstrength and maximin argstrength differ in how many states they consider: all utterance-compatible ones and just a single one respectively; 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, guided e.g., by prototypicality.

Our data suggests that a full account of utterance choice would need to go beyond the purely intensional definitions of argumentative strength we have considered. Even in this simple design, the data contained several patterns that could not be captured by any of the models. These might be due to the connotation of different terms ('right', 'wrong') or processing complexity ('all : most'). We leave a more systematic exploration of these options to future work.

References

Achimova, A., Franke, M., & Butz, M. V. (2025). The alignment model of indirect communication (G. J. Baxter, Ed.). *PLOS One*, *20*(5), e0323839. https://doi.org/10.1371/journal.pone.0323839

Anscombre, 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

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, *35*(1), 3–44. https://doi.org/10.1515/zfs-2016-0002

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

Lambert, B. (2018). *A student's guide to bayesian statistics*. Sage Publications.

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

Winterstein, G. (2012). What *but*-sentences argue for: An argumentative analysis of *but*. *Lingua*, *122*(5), 1864–1885.

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