RSArg

- intro:
  - pragmatics mostly on cooperative dialogue
  - much less on argumentative language use; and what exists is often informal (**AnscombreDucrot1983:Largumentation-**)
  - problem: what's the goal of argumentative language use (what's the payoff function?)
- probabilistic pragmatics / RSA:
  - vanilla version
  - extensions with multiple utility components
  - sketch idea of adding some notion of argumentative strength to utils
  - mention log-odds ratio as an obvious candidate but postpone full model and definition of $argStr(u)$ until after the experimental part
- Exp 1 & 2 (in one swoop):
  - describe experimental design
  - mention (e.g., in footnote and expand in appendix) what was preregistered when
  - report on results w/ visuals and some descriptive stats showing the "argumentativity matters"
- Models:
  - describe different RSArg models, expand on argStr(u)
  - maybe: motivate models with reference to some aspect of the observed data
- Model fits & comparison:
  - describe Bayesian models (hierarchical models, priors etc.)
  - discuss results of model fits and comparison
- Discussion

[MF: comment by Michael] [HW: comment by Hening] [FC: comment by Fausto]

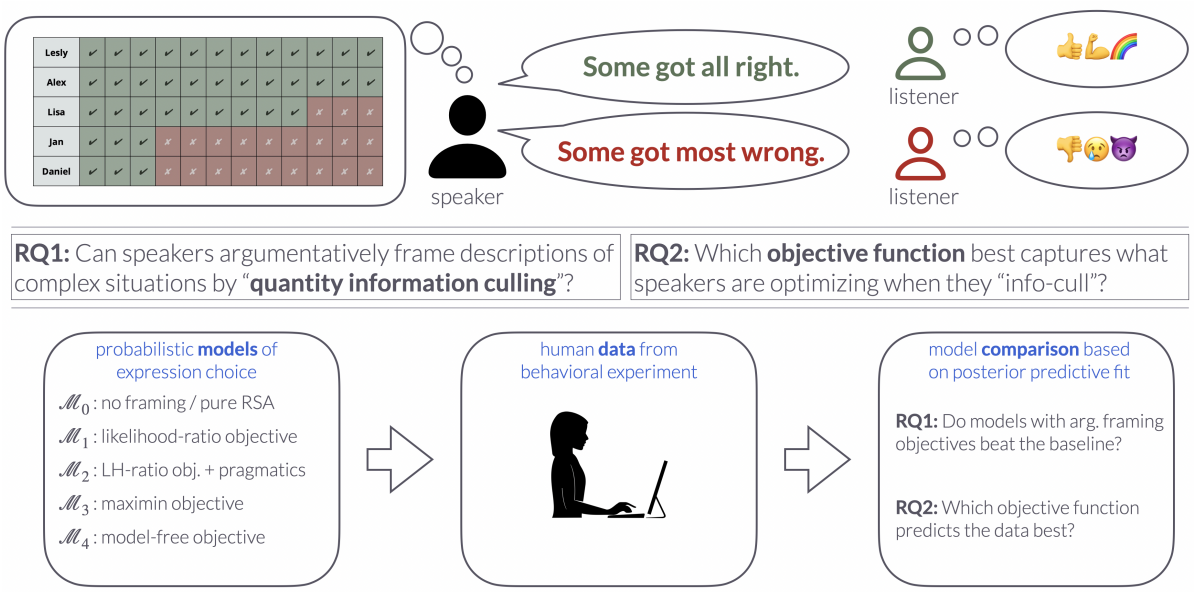

Figure 1: We investigate speakers' flexibly to strategically choose expressions to present a complex situation, like the results of an exam, in a way that makes it appear more positive or more negative, e.g., implying more of a high or a low success rate of the exam. Our main research questions are: (1) whether speaker are able to systematically engage in strategic "information culling" to achieve argumentative framing; and (2) which objective function best characterizes speakers' aggregate behavior, i.e., what is it that speaker *do* when try to frame a situation in one way or another. To address these questions, we compare a number of probabilistic models based on their ability to explain the experimental data. Models differ in the way in which they operationalize the argumentative strength of an expression. [MF: improve visualization]

1 Probabilistic pragmatics & and argumentative discourse

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 (**FrankeDegen2023:The-softmax-fun**), 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 (**FrankGoodman2012:Predicting-Prag**; **Degen2023:The-Rational-Sp**) 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 (**Grice1975:Logic-and-Conve**). 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:

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

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 (**ScontrasTessler2021:A-practical-int**):

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

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 (**FrankGoodman2012:Predicting-Prag**), generics (**Tessler2019:The-Language-of**), conditionals (**GrusdtLassiter2021:Probabilistic-m**), quantifiers and implicature (**GoodmanStuhlmuller2013:Knowledge-and-I**; **Tielvan-TielFranke2021:Probabilistic-p**), gradable adjectives (**LassiterGoodman2015:Adjectival-vagu**), or probability expression (**HerbstrittFranke2019:Complex-probabi**). [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 (**YoonTessler2020:Polite-Speech-E**). 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:

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

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.

2 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 (**Vinicius-Macuch-SilvaWinter2024:Strategic-use-o**), except that we here used a larger set of visual scenes (different array sizes, see below). While the work reported by **Vinicius-Macuch-SilvaWinter2024:Strategic-use-o** 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 in which participants could choose one of the 32 sentences of the scheme in (1).

(1) Expression choice required selecting an outer and inner quantifier and an adjective:

$$\left.\begin{matrix}\text{None}\\ \text{Some}\\ \text{Most}\\ \text{All}\end{matrix}\right\} \text{ of the students got } \left\{\begin{matrix}\text{none}\\ \text{some}\\ \text{most}\\ \text{all}\end{matrix}\right\} \text{ of the questions } \left\{\begin{matrix}\text{right}\\ \text{wrong}\end{matrix}\right\}.$$

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.?]

Materials. The results of high-school exams were presented visually in form of matrices, as shown in Figure 1 (top left). 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 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 example, the matrix in Figure 1 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 (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 (1) under their standard logical meaning, assuming that *some* means *at least one* and *most* means *more than half* (**Vinicius-Macuch-SilvaWinter2024:Strategic-use-o**).

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 disfavorable (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.

3 Models

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

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

$$\text{argstr}(u, H_0, H_1) = \log \frac{P([\![u]\!] \mid H_0)}{P([\![u]\!] \mid H_1)} \tag{4}$$

where $P([\![u]\!] \mid H)$ is the probability that utterance u is true given hypothesis H.

To apply this definition to the case of our experiment, we have to specify what the competing hypotheses are, and how they condition the probability of u being literally true. There are certainly degrees of freedom in this operationalization. The preregistered approach we report on here is as follows. Participants are instructed to argue that the students in a certain classroom have either a high probability of getting the questions right (*high condition*) or a low probability of getting the questions right (*low condition*). We make the simplifying assumption that the participants assume that all students in a class have the same probability γ of answering each question correctly. Under this assumption, each observed exam result consists of n_s samples (one per student) from a Binomial distribution with success parameter γ and n_q (i.e., the number of questions). Result for a participant therefore only conveys how many questions the participants answered correctly. On the other hand, students are identified individually with names in each trial. H_0 is then the hypothesis that the binomial probability parameter p equals γ. Therefore:

$$P([\![u]\!] \mid \gamma) = \sum_{s \in S} (s \in [\![u]\!]\, K_s) \tag{5}$$

$$K_s \propto \prod_{k \in s} \binom{12}{k} \gamma^k (1-\gamma)^{12-k} \tag{6}$$

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 the binomial probabilities are normalized across the arrays in the experiment.

Compared to the vanilla RSA model, this model has two additional parameters: β and γ. These parameters are hard to infer jointly, but we have some prior knowledge that γ is high in the high condition and low in the low condition. Therefore, we set $\gamma = 0.85$ in the high condition and $\gamma = 0.15$ in the low condition.

We fit two versions of this model:

1. A version with completely pooled α and β.
2. A version with by-participant α, β.

A very basic way of checking how the model differs from predictions is to check for which cases the most produced signals for an observation are *not* on the Pareto frontier of informativity/argstrength as calculated, and which signals are produced instead. This gives conceptually the most obvious ways in which the basic model does not capture the data (there's also the cost of the signal but as it turns out it doesn't matter in this case).

It turns out that this sometimes happen for the lr argstrength. 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!

The initial thinking was that this could be because 'most' is in fact considered by the participant argumentatively stronger than the base model predicts. This way, 'some' becomes much weaker than 'most' and 'some students got all answers wrong' could argumentatively weaker than 'most students got most of the answers wrong'. Models 3, 4, 6, 7, and 11 all go this route, and they involve some change to the calculation of the argumentative strength of utterances involving 'most', which makes them argumentatively stronger:

- 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).

However, none of these variations were better than the literal default argstrength for utterances involving 'most'. This means either that these breaks from the model prediction are not so important for the overall likelihood of the dataset, or that these explanations do not make the right kind of predictions overall.

What are other possible explanations? I can see two.

1. For high enough values of γ, most|most|right becomes argumentatively stronger than some|all|right.

Option 1 is what comes out of model 12, 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 more argumentatively strong than some|all).

2. The basic model assumes a single binomial parameter for all students. Therefore, 'some|all|right' is quite argumentatively strong. However, 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.

Consider the second option. - In this case, most|most|right might be argumentatively stronger than some|all|right. - This approach is also going to change the argstrength of the other utterances. Question is how. - Let's say that each student is still described by a Binomial parameter, but these parameters are distributed as a Beta distribution, so the resulting model is hierarchical, i.e. a Beta-Binomial distribution. - To calculate argstrength in this case, we need to calculate p(utterance being true | argued-for state) = $\sum_{s \in \text{observations}} I(s \in \text{utterance}) p(s \mid \text{argued-for state})$. - The observations for simplicity can be just the 20 observations of the model. - The argued-for state in this case can either be a specific combination of parameters of the Beta-Binomial or some range of the parameters. Suppose the former for simplicity. - $p(s \mid \text{argued-for state})$: This is, similarly to the previous model, the probability of an observation given the argued-for-state.

3.2 Pragmatic argstrength RSA

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

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

where P_S is defined above in Equation (1).

We fit two versions of this model:

1. A version with completely pooled α and β.

2. A version with by-participant α, β. This version assumes that each participant uses the same (estimated) value of α for the calculation of the argumentative strength and of the utility.

DOUBLE CHECK THIS

This is the same as the basic model, but instead of calculating the argumentative strength with the truth value, we calculate it with the pragmatic production probabilities. - To see the conceptual connection, note that the original argstrength can be rewritten in terms of the probability of making an observation and the probability of a signal being true given the observation:

$$\begin{aligned}
\text{argstrength}(u) &= \log \frac{p(u \mid \gamma = 0.85)}{p(u \mid \gamma = 0.15)} && (7)\\
&= \log \frac{\sum_{i=1}^{20} p(u, o_i \mid \gamma = 0.85)}{\sum_{i=1}^{20} p(u, o_i \mid \gamma = 0.15)} && (8)\\
&= \log \frac{\sum_{i=1}^{20} p(u \mid o_i) p(o_i, \gamma = 0.85)}{\sum_{i=1}^{20} p(u \mid o_i) p(o_i, \gamma = 0.15)} && (9)\\
&= \log \frac{\sum_{o \in u} p(o, \gamma = 0.85)}{\sum_{o \in u} p(o, \gamma = 0.15)} && (10)\\
& && (11)
\end{aligned}$$

But now instead of considering the probability that an utterance is true, consider the probability of the literal speaker producing an utterance given an observation.

$$\begin{aligned}
\text{argstrength}(u) &= \log \frac{p(u \mid \gamma = 0.85)}{p(u \mid \gamma = 0.15)} && (12)\\
&= \log \frac{\sum_{i=1}^{20} p(u, o_i \mid \gamma = 0.85)}{\sum_{i=1}^{20} p(u, o_i \mid \gamma = 0.15)} && (13)\\
&= \log \frac{\sum_{i=1}^{20} p(u \mid o_i) p(o_i, \gamma = 0.85)}{\sum_{i=1}^{20} p(u \mid o_i) p(o_i, \gamma = 0.15)} && (14)\\
&= \log \frac{\sum_{i=1}^{20} p(o_i, \gamma = 0.85) \begin{cases} \frac{1}{|u|} & \text{if } o \in u \\ 0 & \text{else} \end{cases}}{\sum_{i=1}^{20} p(o_i, \gamma = 0.15) \begin{cases} \frac{1}{|u|} & \text{if } o \in u \\ 0 & \text{else} \end{cases}} && (15)\\
&= \log \frac{\sum_{o \in u} p(o, \gamma = 0.85) \frac{1}{|u|}}{\sum_{o \in u} p(o, \gamma = 0.15) \frac{1}{|u|}} && (16)\\
&= \log \frac{\frac{1}{|u|} \sum_{o \in u} p(o, \gamma = 0.85)}{\frac{1}{|u|} \sum_{o \in u} p(o, \gamma = 0.15)} && (17)\\
&= \log \frac{\sum_{o \in u} p(o, \gamma = 0.85)}{\sum_{o \in u} p(o, \gamma = 0.15)} && (18)\\
& && (19)
\end{aligned}$$

It's the same!

- So the argumentative strength above was basically the Bayes factor calculated by the listener using the literal speaker as a model. - Instead of the literal speaker, a pragmatic speaker can be used. - This pragmatic speaker is different from the real pragmatic speaker, because they do not include argument strength in their calculation, corresponding to the idea that the real speaker is imagining a listener who doesn't take into account the speaker's

argumentative aims. - Nonetheless, I always assume a literal listener with a uniform prior over states (Is this reasonable?) - Conceptually, a pragmatic speaker that uses this type of argumentative strength relies on (exploits) the listener's assumption of cooperativeness. - In this case, there are three differences: - 'some' implicates 'not most' - 'some' implicates 'not all' - 'most' implicates 'not all' - The other signals do not generate implicatures.

3.3 Maximin argstrength RSA

The third model we fit is meant to capture the intuition that, rather than minimizing full argumentative strength as defined above, participants might try to find the utterance u such that the argumentatively weakest among the states compatible with u is maximal. More formally, for each utterance u participants might consider the following argumentative strength:

$$\text{maximin-argstr}(u) = \min_{s \in S} \log \frac{p(s \mid [\![u]\!], \gamma = 0.85)}{p(s \mid [\![u]\!], \gamma = 0.15)} \tag{20}$$

$$p(s \mid [\![u]\!], \gamma) \propto \prod_{k \in s} \binom{12}{k} \gamma^k (1-\gamma)^{12-k} \tag{21}$$

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 3.1.

We fit two versions of this model:

1. A version with completely pooled α and β.
2. A version with by-participant α, β.

Collapse of badness

Maximin argstrength predicts that a lot of the 'bad' signals (given the condition) are equally bad. For instance, 'some|some|wrong' in the high condition is argumentatively as bad as 'all|all|wrong' with minimax-argstrength. On the other hand, with BF-argstrength 'all|all|wrong' is a bad signal to send, 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, which can be seen by comparing figures '13.png' of 'by_observation_maximin_argdelta' and 'by_observation_argdelta'. When participants are forced to describe a 'bad' situation in the high condition, they don't just produce the 'winner' according to BF-argstrength ('some|some|wrong'), but rather produce a larger variety of signals than for other observations (e.g. 'none|all|right'). On the other hand, maximin argstrength cannot explain why 'some|some|wrong' is produced more often.

It feels like the truth is somewhere in the middle, e.g. we consider a few observations for the signal but not all. Maybe a few of the ones close to the 'threshold', or some of the most 'prototypical' for the signal. I feel like the best story depends on where the approximation 'lies':

- If the speaker is perfectly rational but reasons about an imperfect listener, then the question will be 'What are the few states that the listener is likely to consider upon hearing that signal?'.
- On the other hand, if the speaker themselves is approximately rational, the question is 'What states is the speaker most likely to consider when picking a signal?'

Badness of weak signals

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). I don't have strong intuitions either way about this.

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, as can be seen by the argdeltas in 'by_observation_maximin_argdelta'.

Different orderings

In some cases, BF and maximin predict different orderings of goodness. For instance, in the high condition 'most|none|wrong' is better than 'all|most|right' for BF, but the other way around for maximin (I haven't looked yet at how gamma influences this though). Again, I don't have clear intuitions about this.

Better predictions of maximin argstrength for individual observations

- '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.

Worse predictions of maximin argstrength for individual observations

- '5.png', high condition: 'most|all|right' is the most produced signal, and it is the best under BF 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' though.

Some stuff that neither model explains

- '17.png': low condition: Neither model can really make sense of why 'most|all|wrong' is by far the most produced signal. BF thinks 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.

Adjectival maximin

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 is greater than this threshold" and *against* "The value of γ is lower than this threshold". Formally, we can write it as follows:

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

Let's think about how to calculate the components:

$$p(s \mid [[u]] = 1, \phi \geq \theta) = \frac{p(\phi \geq \theta \mid [[u]] = 1, s) p(s \mid [[u]] = 1)}{p(\phi \geq \theta)} \tag{22}$$

3.4 Model-free argstrength

In this version of the model, we define the measure of argumentative strength so that the arguing agent tries 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:[a]

$$\text{modelfree-argstr}(u) = |[\![u]\!]|^{-1} \sum_{s \in [\![u]\!]} \sum_{i \in s} i \qquad \text{High condition} \tag{23}$$

$$\text{modelfree-argstr}(u) = -|[\![u]\!]|^{-1} \sum_{s \in [\![u]\!]} \sum_{i \in s} i \qquad \text{Low condition} \tag{24}$$

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.

This model has three free parameters to fit for each participant: α, β, and the cost for 'none'. Similarly to the previous models, we fit two versions of this model:

1. A completely pooled version
2. A version with by-participant α and β (and completely pooled 'none' cost)

[a] Here, s is interpreted as a list of numbers, one for each student in the class, encoding the number of correct answers by the student.

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.

4 Results

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].

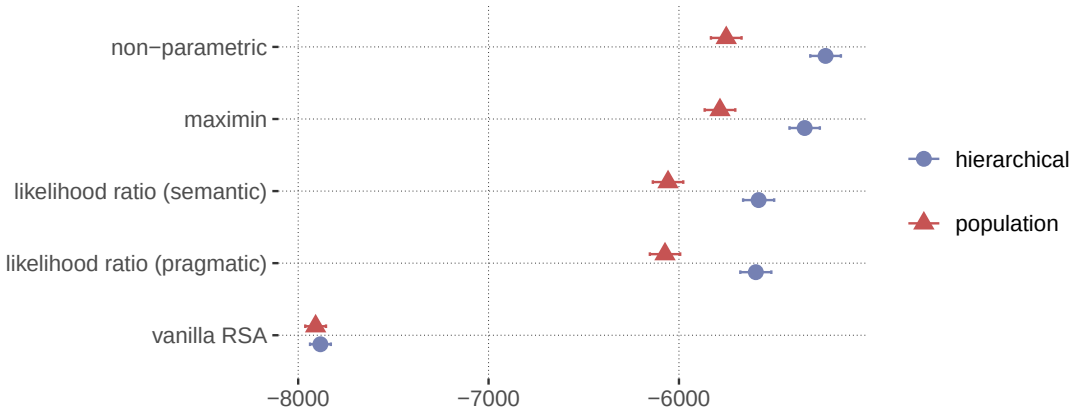

Figure 2: Results of model comparison based on the full data set. For each model, shapes indicate the expected log-probability mass from leave-one-out cross validation, with error bars showing the standard error of these estimates. The *y*-axis lists the different types of models, ordered by ascending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

- 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.

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. - 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.