RSArg

- intro:
  - pragmatics mostly on cooperative dialogue
  - much less on argumentative language use; and what exists is often informal (Anscombre and Ducrot, 1983)
  - 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]

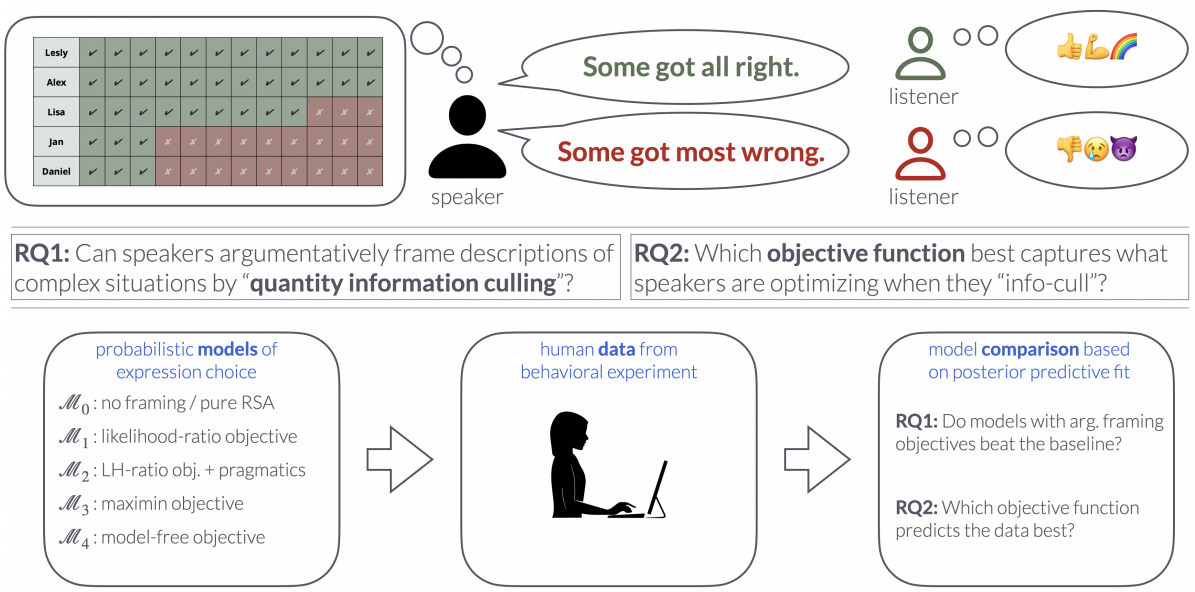

Figure 1: We investigate speakers' flexibly to strategically choose expressions to present a complex situation, like the results of an exam, in a way that makes it appear more positive or more negative, e.g., implying more of a high or a low success rate of the exam. Our main research questions are: (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 (Franke and 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 and Goodman, 2012; Degen, 2023) which adopts the usual Gricean assumptions that the speaker wants 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:

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

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, Tessler, and Franke, 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 and Goodman, 2012), generics (Tessler and Goodman, 2019), conditionals (Grusdt, Lassiter, and Franke, 2022), quantifiers and implicature (Goodman and Stuhlmüller, 2013; van Tiel, Franke, and Sauerland, 2021), gradable adjectives (Lassiter and Goodman, 2017), or probability expression (Herbstritt and Franke, 2019). [MF: insert some more references (without MF as co-author!)] Yet, some phenomena seem to require a 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:

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

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.

- 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

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 (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 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* (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 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 any response that is literally false as a description of the results shown in the corresponding trial. This resulted in [MF: how many trials were excluded?] We also excluded all the data from a participant if the participant selected the same response in all trials or if the participant gave more than 4 false responses. This resulted in [MF: how many people were excluded?] After cleaning, we had data from $N = [MF : fillme]$ participants.

[MF: TODO: implement exclusion criteria in R; get the numbers for this section; replot with proper data exclusion]

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.

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.

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:

$$\text{maximin-argstr}(u) = \min_{s \in S} \log \frac{p(s \mid [\![u]\!], \gamma = 0.85)}{p(s \mid [\![u]\!], \gamma = 0.15)} \tag{7}$$

$$p(s \mid [\![u]\!], \gamma) \propto \prod_{k \in s} \binom{12}{k} \gamma^k (1-\gamma)^{12-k} \tag{8}$$

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.

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]

$$\text{modelfree-argstr}(u) = |[\![u]\!]|^{-1} \sum_{s \in [\![u]\!]} \sum_{i \in s} i \qquad \text{High condition} \tag{9}$$

$$\text{modelfree-argstr}(u) = -|[\![u]\!]|^{-1} \sum_{s \in [\![u]\!]} \sum_{i \in s} i \qquad \text{Low condition} \tag{10}$$

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.

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].

References

Anscombre, Jean-Claude and Oswald Ducrot (1983). *L'argumentation dans la langue*. Brussels: Mardaga.

Degen, Judith (2023). "The Rational Speech Act Framework". In: *Annual Review of Linguistics* 9.1, pp. 519–540.

Frank, Michael C. and Noah D. Goodman (2012). "Predicting Pragmatic Reasoning in Language Games". In: *Science* 336.6084, p. 998.

Franke, Michael and Judith Degen (2023). "The softmax function: Properties, motivation, and interpretation". In.

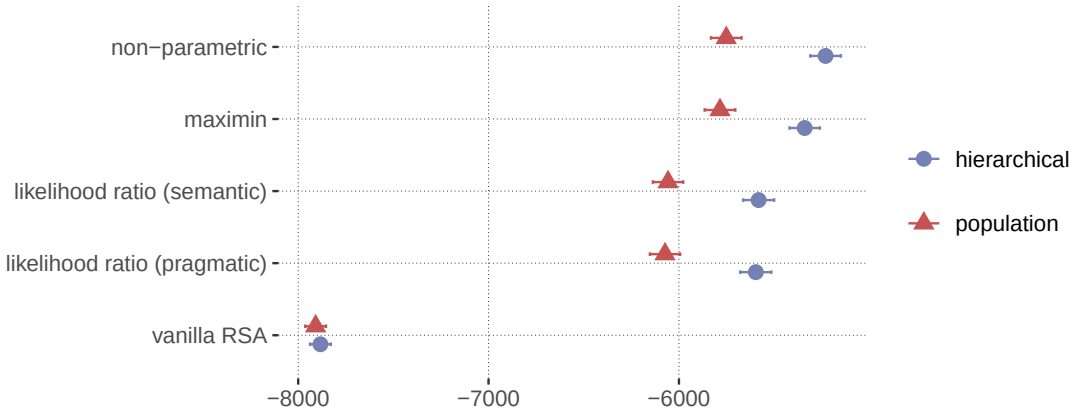

Figure 2: Results of model comparison based on the full data set. For each model, shapes indicate the expected log-probability mass from leave-one-out cross validation, with error bars showing the standard error of these estimates. The *y*-axis lists the different types of models, ordered by ascending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

Goodman, Noah D. and Andreas Stuhlmüller (2013). "Knowledge and Implicature: Modeling Lanuage Understanding as Social Cognition". In: *Topics in Cognitive Science* 5, pp. 173–184.

Grice, Paul Herbert (1975). "Logic and Conversation". In: *Syntax and Semantics, Vol. 3, Speech Acts*. Ed. by Peter Cole and Jerry L. Morgan. New York: Academic Press, pp. 41–58.

Grusdt, Britta, Daniel Lassiter, and Michael Franke (2022). "Probabilistic modeling of rational communication with conditionals". In: *Semantics & Pragmatics* 15.

Herbstritt, Michele and Michael Franke (2019). "Complex probability expressions & high-order uncertainty: compositional semantics, probabilistic pragmatics & experimental data". In: *Cognition* 186, pp. 50–71.

Lassiter, Daniel and Noah D. Goodman (2017). "Adjectival vagueness in a Bayesian model of interpretation". In: *Synthese* 194.10, pp. 3801–3836.

Macuch Silva, Vinicius et al. (2024). "Strategic use of English quantifiers in the reporting of quantitative information". In: *Discourse Processes* 61.10, pp. 498–523.

Scontras, Gregory, Michael Henry Tessler, and Michael Franke (2021). *A practical introduction to the Rational Speech Act modeling framework*.

Tessler, Michael Henry and Noah D. Goodman (2019). "The Language of Generalization". In: *Psychological Science* 126.3, pp. 395–436.

van Tiel, Bob, Michael Franke, and Uli Sauerland (2021). "Probabilistic pragmatics explains gradience and focality in natural language quantification". In: *Proceedings of the National Academy of Sciences* 118.

Yoon, Erica J. et al. (2020). "Polite Speech Emerges From Competing Social Goals". In: *Open Mind: Discoveries in Cognitive Science* 4, pp. 71–87.