Chapter 1

Introduction

Language is built upon a foundation of trust, and there is a case to be made that this trust is justified. Our modern lives depend almost entirely on testimonial knowledge (Mercier and Sperber, 2017; Levy, 2021), and language itself is evolutionarily stable because cooperativity is generally advantageous (McCready, 2015). However, listeners obviously do not always automatically accept assertions on the basis of conversational norms (Sperber et al, 2010), and speakers do not expect their addressees to take implausible testimony for granted (Oey et al, 2023). Even when Grice's (1975) Quality maxim is presumably in effect, interlocutors might have access to different bodies of evidence against which they evaluate facts, or else they may differ subjectively on matters of taste or opinion. This study concerns the use of *reasons* to negotiate the acceptance of controversial beliefs.

In argumentative scenarios, the function of a reason is to reduce the listener's dependence on a speaker's trustworthiness, by making the content itself more *plausible*, or predictable in context (Mercier, 2020). Beliefs justified by many coherent reasons, then, should be more credible– or at least, they should be no less credible– than beliefs justified by few or none. This analysis is intuitive, and it is predicted by a number of formal and probabilistic models of argument strength (Anscombre and Ducrot, 1983; Merin, 1999; Winterstein, 2018). The present study, however, will investigate a competing intuition: that arguments are actually weakened when they are *redundantly* justified– that is, when (far) more justification is provided than what would be expected or required to raise an interlocutor's credence to the threshold for acceptance. Put famously by Queen Gertrude in Act III, Scene II of *Hamlet*: "The lady doth protest too

much, methinks" (see Figure 1.1).

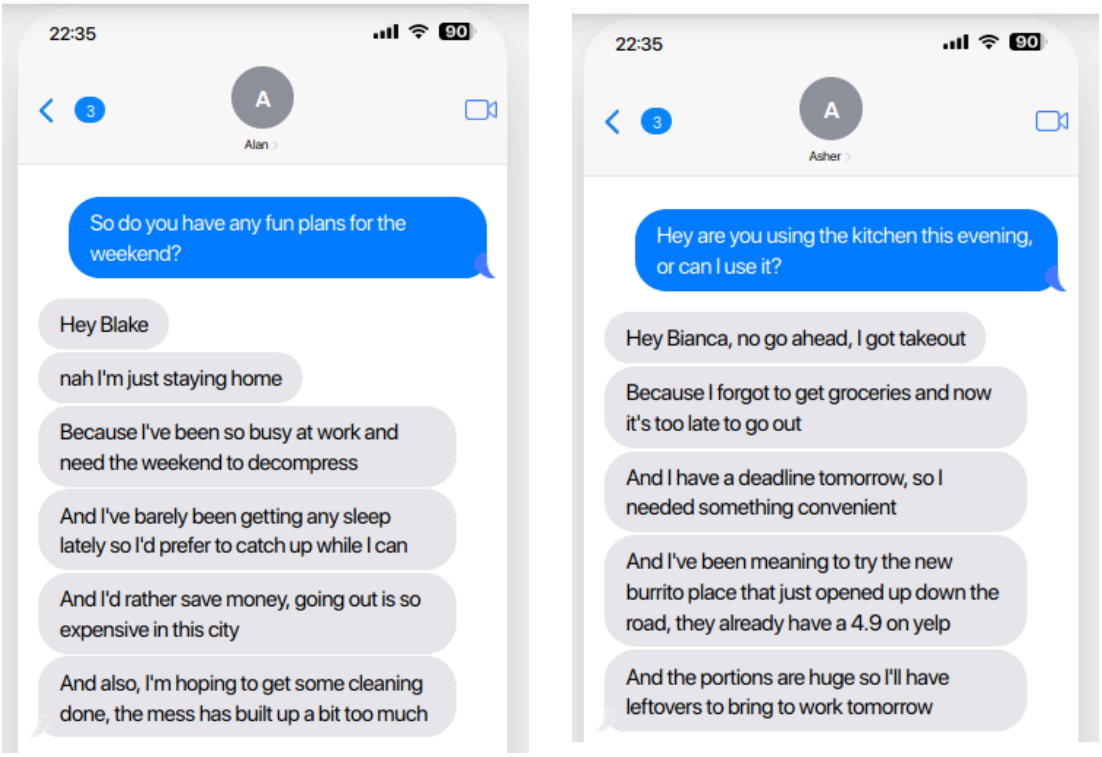


Figure 1.1: Examples of over-justification.

This effect, I propose, is predicted by the vast psycholinguistic literature on informativity. In this literature, information is defined as a measure of surprise, in keeping with computational theories of communication (following Shannon, 1948). If speakers are to make efficient use of their cognitive resources to maximize information exchange while avoiding misunderstandings, longer signals should be reserved for meanings that are more surprising, or less plausible. This prediction is borne out across all levels of linguistic structure from phonemes to full utterances (Aylett and Turk, 2004; Levy and Jaeger, 2007; Mahowald et al, 2013; Asr and Demberg, 2020; Lemke et al, 2022).

Likewise, comprehenders have been shown to interpret utterances with an expectation for informativity: they assume that speakers do not mention easily inferrable information, and they derive pragmatic inferences to rationalize why seemingly redundant utterances might be situationally informative. For example, assertion of default script continuations induces atypicality inferences ("she went to the grocery store and **paid the cashier**!" suggests that the subject *normally* doesn't pay for her groceries, as one might expect in a typical shopping script; Kravtchenko and Demberg, 2022; Ryzhova et al, 2023).

Bearing this in mind, informal arguments might indeed be weakened if more reasons than contextually appropriate are provided. Since common ground admissibility is *normally* taken for granted, a communicatively efficient speaker would prefer to justify her conclusion only when she believes that it will be particularly difficult to accept, i.e. that it is particularly controversial or implausible. If a rational listener recognizes that the speaker *expected* a disagreement, he might become suspicious of her argument even if her reasons are sensible. As a result, reasons that (extralinguistically) count in favor of the conclusion might then (pragmatically) count against it.

The experiment reported in the subsequent chapters tests the predictions of these competing hypotheses. In a behavioral study, participants read a series of text-message conversations, in which a character justifies an action with zero or four reasons. In half the conversations, the justifications are provided out of the blue (i.e. unexpectedly), and in the other half, they are elicited discoursally to suppress the informativity inference. For each conversation, participants rated whether *the speaker thought* the listener would respond favorably or unfavorably (hereafter the "pragmatic prior"), and whether the listener would actually respond favorably (the posterior). The predictions of Queen Gertrude are borne out– participants predicted a lower pragmatic prior when many reasons were provided unexpectedly, and there was a significant difference in argument strength between the expected and unexpected reasons. The findings of this study demonstrate how so-called "cooperative" pragmatics still bears upon seemingly "non-cooperative" interactions.

Chapter 2

Background

2.1 The Purposes of Argument

Argumentation has been a primary object of study in philosophy for as long as philosophy has been studied, but in linguistic pragmatics, it is a somewhat more marginal case: pragmatics is typically envisioned in terms of cooperative information exchange (e.g. Grice, 1975; Stalnaker, 1978), whereas argument is frequently envisioned adversarially (e.g. Merin, 1999). In some sense, this is not a surprise: we often think of "an argument" as something that can be won or lost, and that argument is somehow more suspicious or worthy of scrutiny than information. Whether or not this is actually true, let us grant it for now that argument involves an antagonist, whereas information involves an interlocutor who is at worst indifferent.

The function of an argument, then, is to give the interlocutor a reason to change his mind. Since our interlocutor is more inclined to accept propositions that increase the overall coherence of his beliefs, we offer reasons that "count in favor" of a controversial conclusion to make that conclusion more coherent with– i.e. inferrable from– propositions he will admit more easily (Mercier, 2020).

The "argues for" or "counts in favor" relation can be treated categorically, as is frequently the case in the study of formal/syllogistic arguments. Taking a more gradient tack, however, the strength of a reason can instead be characterized in terms of its effectiveness in bringing about the desired belief (Winterstein, 2018). The latter approach lends itself to probabilistic, and specifically Bayesian models: in brief, a set of reasons argues in favor of the conclusion if the posterior probability of the conclusion

conditioned on the reasons $P(C|R)$ is greater than the (pre-utterance) prior probability of the conclusion $P(C)$. The strength of a reason in "informal" argumentation is quantified by the extent to which it raises your *degree of belief* (credence) or commitment towards the conclusion (e.g. Hahn and Oaksford, 2007; Godden and Zenker, 2018).

Recent efforts have been made to import the Bayesian perspective into existing linguistic theories of argumentation (e.g. Anscombre and Ducrot, 1983). The central insight of such accounts is that an utterance's ability to argue in favor of a conclusion depends on its linguistic form in addition to the propositional content: for example, *Alice almost arrived on time, don't scold her!* is a legible argument while *#Alice barely arrived on time, don't scold her!* is not, even though only the latter entails *Alice arrived on time*. Winterstein (2012; 2018) adopts the Bayesian approach to formalize the argumentative profile of a number of discourse connectives and particles.

Along similar lines, it has been observed that pragmatic enrichment, due for example to approximate numeral interpretation (Cummins and Franke, 2021) or quantifier vagueness (Macuch Silva et al, 2024), can also affect the extent to which an utterance argues in favor of a conclusion. For example, *This school has a top 19 linguistics program* is logically stronger, but pragmatically and therefore argumentatively weaker, than *This school has a top 20 linguistics program* (Cummins and Franke, 2021), since 19 is typically interpreted punctually, whereas 20 is typically interpreted vaguely and might include numbers higher than 19.

Although these accounts diverge on which aspects of form and meaning contribute to argumentative strength and how precisely that strength is quantified, they all seem to make a tacit assumption that the "best" argument licenses the strongest inference from the reasons to the conclusion. Indeed, a number of them also make reference to a type of reasoning over alternatives reminiscent of Grice, where a listener might reject a favorable reason if he knows of an obviously stronger one the speaker could have provided (assuming that she would have, if it were true; see also the "weak evidence effect", Barnett et al, 2022).

As such, it seems to be the case that *many* good reasons (for some measure of goodness, and as long as they are all acceptable and consistent with one another), should always count more strongly in favor of the conclusion. The listener's degree of belief in the conclusion should increase monotonically towards certainty (i.e. $P(C|R) = 1$) as more positive reasons are provided. While first-order Bayesian conditionalization

does predict "diminishing returns" as the addressee's credence gets closer to 1, it has nothing to say about Queen Gertrude's observation, that *too many* reasons might count *against* the conclusion.

Now, I want to question the assumption that, if the speaker provides a more complete justification for any given conclusion, the listener can only be better off epistemically. When the conclusion *C* is particularly challenging or implausible to the listener, providing reasons becomes necessary to make *C* available by inference (Mercier, 2020). But in a day-to-day conversation, it seems to me that the typical number of reasons provided in favor of any given conclusion is zero; most assertions are admitted on the basis of conversational norms (i.e. Grice, 1975).

On the view of Sperber et al (2010), learning from testimony typically requires a combination of trust in the speaker and plausibility of the utterance, rather than just one or the other. Of course, it would not be rational to trust indiscriminately, and language would not be evolutionarily stable if gullible listeners always accepted the testimony of deceptive speakers (Mercier, 2020). But almost all of the knowledge required to support our modern lives can only be acquired testimonially (e.g. Levy, 2021; Mercier and Sperber, 2017), and language would be highly *inefficient* means of acquiring new knowledge if everything we learned from others was fully or even mostly deducible from known propositions (Kravtchenko and Demberg, 2022– among many others).

Then, if the purpose of an (argumentative) reason is to raise the interlocutor's credence towards some conclusion in order to ensure its acceptance, highly effortful arguments citing multiple reasons would only be worthwhile to a rational speaker when her addressee is very suspicious or strongly epistemically biased against her claim. But as the vast psycholinguistic literature on redundancy has shown, speakers are often more informative than listeners expect, and the case of "over-justification" seems no different. Drawing upon this literature, the rest of the paper will propose that the "protesting too much" inference is simply a case of a rational listener expecting informativity from a rational speaker.

2.2 Rational Redundancy

"Information" in formal pragmatics is traditionally conceptualized as the set of possible worlds at which a proposition is true: it is a measure of logical strength. In certain computational and psycholinguistic traditions, however, information is a measure of

surprise. Any linguistic unit (word, phoneme, etc) is informative when it is surprising; it is redundant when it is predictable (Levy and Jaeger, 2007).

Following Shannon (1948), the information content ("surprisal") of a linguistic unit is typically quantified by the negative log probability of that unit conditioned on its context: $-\log P(u_i|u_0...u_{i-1})$. This influential model describes the optimal rate of information, or bitrate, through a noisy channel with limited bandwidth. In order to maximize the amount of information that can be transmitted without degradation, a constant bitrate that approaches, but does not exceed, the channel's capacity must be maintained. As such, the optimal coding scheme for a set of messages M distributed according to $P(M)$ assigns $-logP(m)$ bits to each $m \in M$: longer codes are assigned to low-probability messages, and shorter codes to high-probability messages, in order to keep the bitrate fixed.

Natural languages are not optimal coding schemes, of course. However, it has long been of interest to psycholinguists whether language users might make *rational* usage of limited cognitive resources by constructing utterances efficiently. Over the past 20 years, numerous instantiations of the noisy-channel model have successfully captured patterns of signal reduction in low-surprisal contexts across various levels of linguistic structure. Early work on speech processing first demonstrated reductions in duration and articulatory detail on syllables and words in predictive contexts (Bell et al, 2003; Aylett and Turk, 2004; Pluymaekers et al, 2005). This pattern was generalized to all levels of language processing as the Uniform Information Density (UID) Hypothesis (Levy and Jaeger, 2007), and has accounted for phenomena such as contractions (Piantadosi et al, 2011), lexical reductions (e.g. chimp/chimpanzee, Mahowald et al, 2013), omission of complementizers (Jaeger, 2010), use of sentence fragments (Lemke et al, 2021), and optionality of discourse connectives (Asr and Demberg, 2020).

Comprehenders are aware of these production preferences and, correspondingly, have been shown to interpret utterances with an expectation for informativity (Sedivy, 2003; Rohde et al, 2021; Rohde and Rubio-Fernandez; 2022, Hao et al, 2025). They tend to assume that rational speakers will not mention predictable information, and often revise their prior beliefs about the context of the utterance if necessary to rationalize any apparent redundancies. For example, upon hearing the word "yellow", comprehenders anticipate that the continuation will be an object that is plausibly yellow (e.g. a shirt), but not predictably yellow (e.g. a banana; Rohde and Rubio-Fernandez, 2022), unless the scene also contains a nonprototypical category competitor (e.g. a brown banana;

Sedivy, 2003). While most of this literature deals with "over"modification with adjectives in object requests, a number of recent experiments have extended this analysis to more obstructive, utterance-level redundancies (Kravtchenko and Demberg, 2022; Rees and Rohde, 2023; Ryzhova et al, 2023; Reksnes et al, 2024). These experiments involve the assertion of a proposition that is trivially predictable from the conversational common ground and/or basic world knowledge. For example, Kravtchenko and Demberg (2022) show that assertion of default script continuations (*She went to the grocery store,* ***and she paid the cashier!***) give rise to habituality inferences: listeners infer that this particular character doesn't normally pay for her groceries.

Adapting the "wonky world" model of Degen et al (2015), Kravtchenko and Demberg argue that a rational speaker would only make an effortful assertion of *she paid the cashier* if it meaningfully increased the probability of a literal listener (in the sense of Frank and Goodman, 2012) inferring the correct world state (i.e. the one where the character pays for her groceries). A pragmatic listener, then, jointly infers the world state and the habituality of the action– considering the fact that the speaker found it worthwhile to mention the event, she must have predicted that the listener's prior on *she paid the cashier* was not high enough for him to assume it by default. As such, he learns that the character does not usually pay for her groceries (i.e. the "wonky" prior), in addition to the fact that she did on this specific occasion (the posterior).

Returning now to the case of *reasons*– recall that the purpose of an argument is to raise the listener's degree of belief towards a controversial or implausible conclusion C. Argument strength on the first-order analysis is the increase in credence from the prior $P(C)$ to the posterior $P(C|R)$. Bearing that in mind, a rational speaker would make a *more* effortful argument if she expected the listener to be *more* resistant to her conclusion (i.e. she expected his prior belief would be low, and would require numerous reasons to increase it to the level of acceptance). The pragmatic listener jointly infers this expected prior, as well as his posterior belief in C.

2.3 The Present Study

The objective of the present study is to determine whether such communicative efficiency considerations play a role in the evaluation of reasons– i.e. the argument strength. I will constrain the space of "arguments" to first-person explanations for actions, in the format of "I did/will do A, because of W, X, Y, Z". This is a sensible

candidate for potential "over-justification" because, of course, people often want to be sure they are perceived as having done "the right thing". The experiment will measure whether excessively detailed reasoning for an action can negatively influence the interlocutor's opinion of the action.

Further, it investigates whether such a change in opinion is moderated by how "over-informative" the listener actually perceives the explanation to be. Naturally, it cannot be the case that giving many reasons to justify a conclusion can never increase the hearer's credence–if that were the case, this paper has already gone on too long. Credence-raising conversational moves are known to be interpreted informatively when the prejacent is surprising or at-issue in the discourse (i.e. a cooperative communicator is *expected* to increase her epistemic/evidential standard to admit the proposition), and redundantly or even infelicitously otherwise (c.f. Lai, 2012; Beltrama, 2018). As such, this experiment also makes use of a "reasons expected" vs "reasons unexpected" manipulation, as a baseline for an informative argument.

2.3.1 RQ1: The Pragmatic Prior

Comprehenders, at first, will have their own independent assessment of the "favorability" or acceptability of the action; call this the naive prior. The naive prior will be agnostic to any justification provided by the speaker. However, upon hearing explicit reasons, comprehenders might infer how (un)favorable the action would have to be in order to merit an effortful rationale; call this the pragmatic prior. The first research question of this study asks whether over-justification can generate a pragmatic prior that is meaningfully lower (less favorable) than the naive prior. Such an inference would be in keeping with the previously mentioned studies (Degen et al, 2015; Kravtchenko and Demberg, 2022) which show that comprehenders will interpret against "wonky" priors to make sense of utterances which would not be "rational" to assert against their naive priors.

2.3.2 RQ2: Post-utterance Belief

The second component of the proposal is that the reduction of the prior, as proposed in RQ1, can counteract the strength of an argument. That is, if the listener learns that the action should *normally* be received unfavorably, then an informal (i.e. non-deductive) argument might not be sufficient to recover from that inference, even if all

of the individual reasons are strong.

2.3.3 Hypotheses

1. The **naive hypothesis** predicts that comprehenders will not reason about the speaker's motivation for justifying her action. As such, the number of reasons provided should have no effect on the prior. Furthermore, assuming the reasons are good ones, the listener's posterior belief should always be *more* favorable when reasons are provided.

2. The **pragmatic hypothesis** predicts that comprehenders *will* reason about the speaker's motivation for justifying her action (i.e. she predicted a disagreement). The number of reasons provided should yield a reduction in the prior when reasons are provided unexpectedly, and should have no effect when the reasons are solicited. As such, the listener's posterior belief should be *less* favorable when reasons are unexpectedly/redundantly provided, and *more* favorable when reasons are solicited/expectedly provided.

Chapter 3

Methods

3.1 Participants

60 participants were recruited through Prolific. All participants self-identified as native speakers of English, and were residents of the United Kingdom or United States. An additional 6 participants were recruited for a shorter pilot study (with the free-text question omitted); their data is not included in the subsequent analysis. The survey took a median time of 17 minutes, 52 seconds to complete, and participants were compensated £5.00. All participants in the main study passed the attention checks, so they were all included in the analysis.

3.2 Study Design

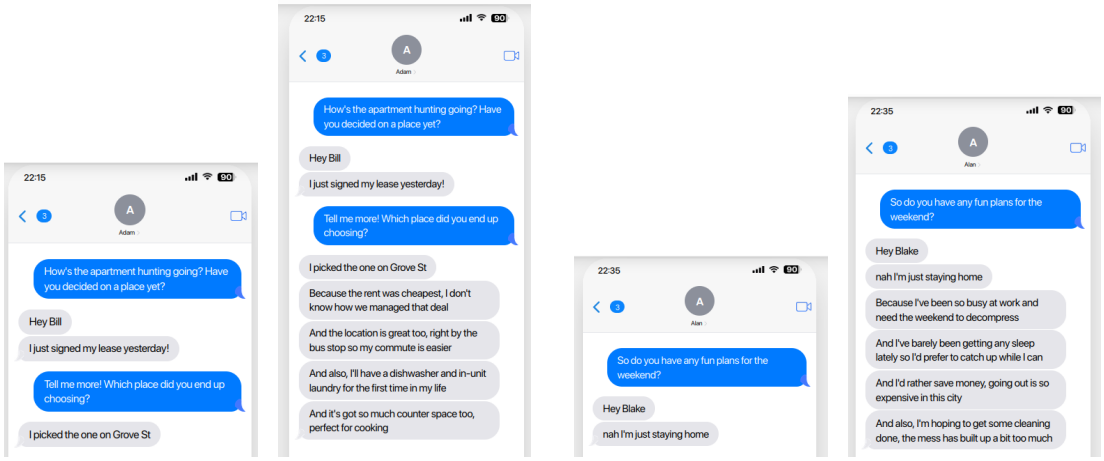
The experiment was constructed using a 2x2 Latin square design, crossing expectation (**reasonsUnexpected** vs **reasonsExpected**) with reasons (**noReasonsGiven** vs **reasonsGiven**. Items were sorted into two lists, with 4 trials in each of the 4 conditions. Each item appeared in only one condition per list. Participants also saw 4 attention checks, for a total of 20 trials.

3.3 Materials

Prior work has shown that comprehenders have a greater expectation for informativity when the utterance is produced by a salient speaker (for example, a named character), so redundancies are more easily perceived as deliberate (Reksnes et al, 2024). As

such, the stimuli for this study were presented as text message conversations between two named characters, adapting the formatting of Wallbridge et al (2021). To avoid confusion, the explainer's name always began with A, and the addressee's name always began with B.

Since the research questions target comprehenders' inferences about speakers' production choices (rather than speakers' production choices themselves), participants were made to assume the role of the addressee rather than the explainer. Therefore, the explanation always came from the "incoming"/grey bubble character, and both rating scale questions concerned the opinion of the "outgoing"/blue bubble character.



Sample **reasonsExpected** stimuli

Sample **reasonsUnexpected** stimuli

Figure 3.1: Sample stimuli in **noReasonsGiven** and **reasonsGiven** conditions.

For the **reasonsUnexpected** class of items, character A describes an "ordinary" altruistic or innocuous behavior (e.g. cleaning the kitchen for a roommate or studying in the library), which typically would not involve explicit justification. The reasons (when provided) were offered unprompted.

For the **reasonsExpected** class of items, character A describes a behavior that typically involves a deliberative process where multiple reasons would be considered (e.g. moving to a new city or making an expensive purchase), in response to a "which one"/"tell me more"-type question from character B. This was meant to create a context where extensive elaboration would be appropriate both in terms of subject-matter and surrounding discourse.

In the **noReasonsGiven** condition, character A simply mentions the behavior.

In the **reasonsGiven** condition, character A provides four reasons. The content of the

explanation was not held constant across item types, but was instead tailored to provide strong/sufficient justification for each individual behavior (to avoid confounding with the "weak evidence" effect e.g. Barnett et al, 2022; as well as to maintain the naturalness of the dialogues). All explanations were approximately the same total length (although some variability was introduced, again to maintain naturalism).

Realism of the stimuli was prioritized so that the participants might have actual intuitions about why a justification would or would not be provided in a given circumstance, following prior work on pragmatic inferences generated by strategic language. For example, while a number of studies have claimed that participants nearly-uniformly do not make lie judgments on implicit meaning, Weigmann (2023) shows that such judgments are simply suppressed in odd situations where the characters' motives for deception are unintuitive. This prediction was borne out in the responses to the free-text question, see section 5.4.

In all target trials, the messages cut off before character B responds. An additional 4 attention checks were constructed, where character B does react to character A's message with explicit approval or disapproval (2 items for each), so the participant could simply report B's perspective without inference.

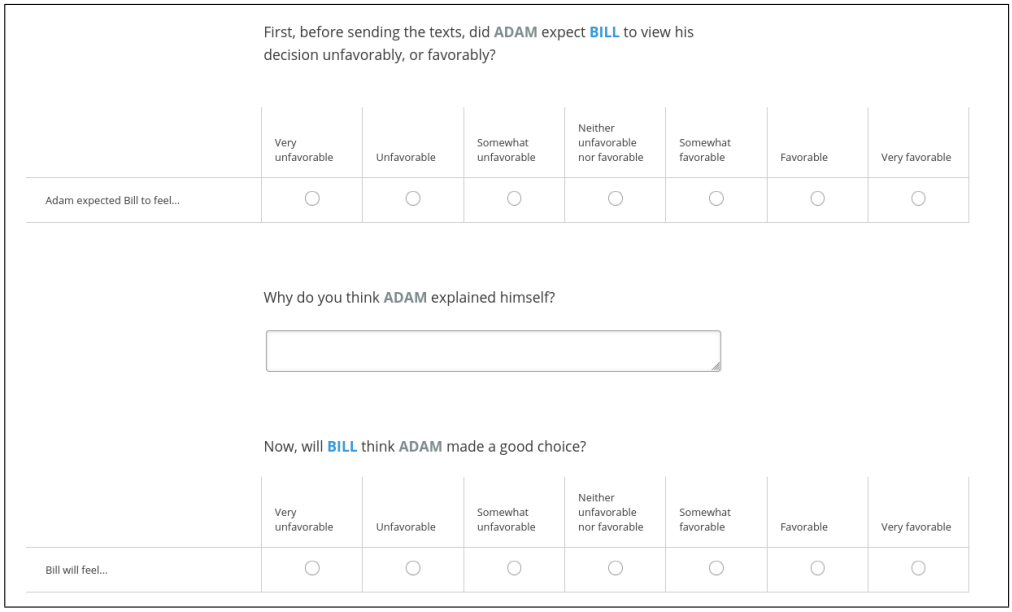


Figure 3.2: Sample questions in the **reasonsGiven** condition.

Each target trial contained two or three questions. First, participants were asked to rate the following on a 7-point Likert scale from "strongly unfavorable" to "strongly favorable":

- First, before sending the texts, did A expect B to view her decision unfavorably, or favorably? A expected B to feel...
- Now, will B think A made a good choice? B will feel...

The first rating was intended to target RQ1: whether the listener would think the speaker had explained her decision because she expected him to disagree with it. The second rating was intended to target RQ2: whether the listener's recognition of that expectation could counteract the effectiveness of the reasons in influencing his beliefs. In other words, the first question identifies whether the "pragmatic prior" is lower (less favorable) in the **reasonsGiven** condition than in the **noReasonsGiven** condition, and whether that effect is moderated by expectedness of the explanation. The second question identifies whether the lower prior produced by the explanation would then yield a reduction in *post-utterance belief* (again, a reduction in favorability rating from the **noReasonsGiven** condition to the **reasonsGiven** condition– and again, whether that effect is moderated by expectedness of the explanation).

Participants were also asked to provide a free-text response to the question "Why do you think A explained herself?" for items in the **reasonsGiven** condition, to see if participants preferred an alternative rationalization for the over-informativity besides the hypothesized judgments measured by the Likert ratings. This additional question was included in the main experiment to provide more interpretable data for analysis, after conducting a preliminary pilot study which only included the Likert ratings.

3.4 Procedure

The experiment was hosted online on Qualtrics. Before proceeding to the survey, participants were asked to read an information sheet and consent to participation (approved by the University of Edinburgh LEL Research Ethics Board) and to confirm their Prolific ID and native-speaker status. The participants were then instructed to read each set of messages closely and reason about the characters' motives for sending them. Using the Qualtrics survey randomizer, each participant was assigned to one of the two stimulus lists, and saw the items in random order. Each trial was presented on a separate page, and participants were required to answer each of the questions before clicking to advance. At the end of the survey, participants completed a "debrief", where they were asked to provide feedback and guess what the experiment was about.

Chapter 4

Results

4.1 Model selection

Data analysis for the Likert rating questions was conducted using a Bayesian cumulative logit model. Although linear mixed-effects models are standard practice in psychology for 7-point Likert scales (assumption of equidistant category boundaries is more acceptable with high-granularity ordinal scales), the Bayesian cumulative model is still best practice for ordinal data since it makes no assumptions regarding the spacing of the categories (Bürkner and Vuorre, 2019). The model was implemented using the R (R Core Team, 2025) package `brms` (Bürkner, 2017) with default priors. The fixed effects included REASONS (**noReasonsGiven** vs **reasonsGiven**), EXPECTATION (**reasonsUnexpected** vs **reasonsExpected**), and their interaction. To account for participant and item-level variability, random effects for PARTICIPANT and ITEM were also included.

4.2 Do listeners rationalize over-explanation?

The first Likert rating question (*Before sending the texts, did A expect B to view her decision unfavorably, or favorably?*) measures the pragmatic prior. It was intended to target RQ1: whether listeners would attempt to rationalize a speaker's decision to (over)explain her actions, by inferring that she predicted a disagreement. Responses are plotted in Figure 5.1.

Treating **noReasonsGiven** and **reasonsUnexpected** as the baseline conditions, the

model identified a credible main effect of REASONS (β = -0.77, 95% CI -1.09, -0.43), indicating a reduction in favorability rating when reasons were given. This effect was moderated by a credible interaction of REASONS with EXPECTATION (β = 0.83, 95% CI 0.35, 1.30), which reversed the main effect entirely when the reasons were expected. In other words, participants rated the speaker's estimation of the listener's opinion credibly lower when reasons were *unexpectedly* provided. Participants did not rate the speaker's estimation of the listener's opinion credibly lower when the reasons were *expectedly* provided. There was also no credible main effect of EXPECTATION, indicating that **reasonsExpected** items were not consistently rated higher or lower in favorability than **reasonsUnexpected** items in the **noReasonsGiven** condition. Results of the models are summarized in Table 5.1.

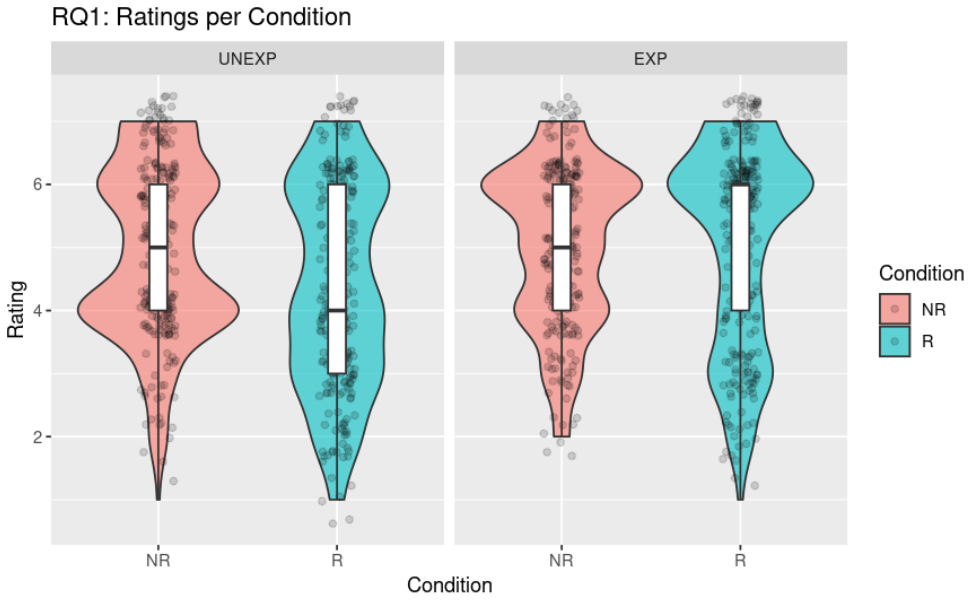


Figure 4.1: Responses to RQ1. Prior favorability ratings reduced when reasons were provided unexpectedly.

These results are in accordance with the predictions of the pragmatic hypothesis: when many reasons are unexpectedly provided, the listener rationalizes the speaker's decision to provide them. The listener infers that the speaker expected him to disapprove of her action– that he would need to be *convinced* to agree with her. The result is a reduced "pragmatic prior" favorability rating. However, in a context where many reasons are appropriate, no such inference is necessary. The literal hypothesis would instead predict no effect for either item type (expected or unexpected), since the listener does not reason about the speaker's choice to explain herself in any case.

Predictor	Estimate	Std. Error	95% CI Lower	95% CI Upper
conditionR	**-0.77**	**0.17**	**-1.09**	**-0.43**
typeEXP	0.19	0.66	-1.12	1.52
conditionR:typeEXP	**0.83**	**0.24**	**0.35**	**1.30**

Table 4.1: Log-odds coefficients from the Bayesian ordinal model of RQ1. A credible main effect of REASONS, moderated by a credible interaction with EXPECTATION, is observed.

4.3 Does over-explanation affect argument strength?

The second Likert scale question (*Now, will B think A made a good choice?*) measures the listener's posterior favorability rating. It was intended to target RQ2: whether higher-order inferences about *why the speaker felt an explanation was necessary* would mitigate that explanation's effectiveness in persuading the listener. Responses are plotted in Figure 5.2.

Once again treating **noReasonsGiven**, **reasonsUnexpected** as the baseline conditions, there was no credible main effect of REASONS for **reasonsUnexpected** items. However, the model identified a credible interaction of REASONS with EXPECTATION (β = 0.75, 95% CI 0.28, 1.22), meaning that the posterior favorability ratings improved when the reasons were *expected*. Although unexpected reasons had no effect on the listener's post-utterance favorability ratings, expected reasons significantly improved the listener's rating. There was still no credible main effect of EXPECTATION, indicating that **reasonsExpected** items were not consistently rated higher or lower than **reasonsUnexpected** items in the **noReasonsGiven** condition. Results of the models are summarized in Table 5.2.

Predictor	Estimate	Std. Error	95% CI Lower	95% CI Upper
conditionR	0.15	0.17	-0.20	0.50
typeEXP	0.25	0.76	-1.25	1.79
conditionR:typeEXP	**0.75**	**0.24**	**0.28**	**1.22**

Table 4.2: Log-odds coefficients from the Bayesian ordinal model of RQ2. A credible interaction of REASONS and EXPECTATION is observed.

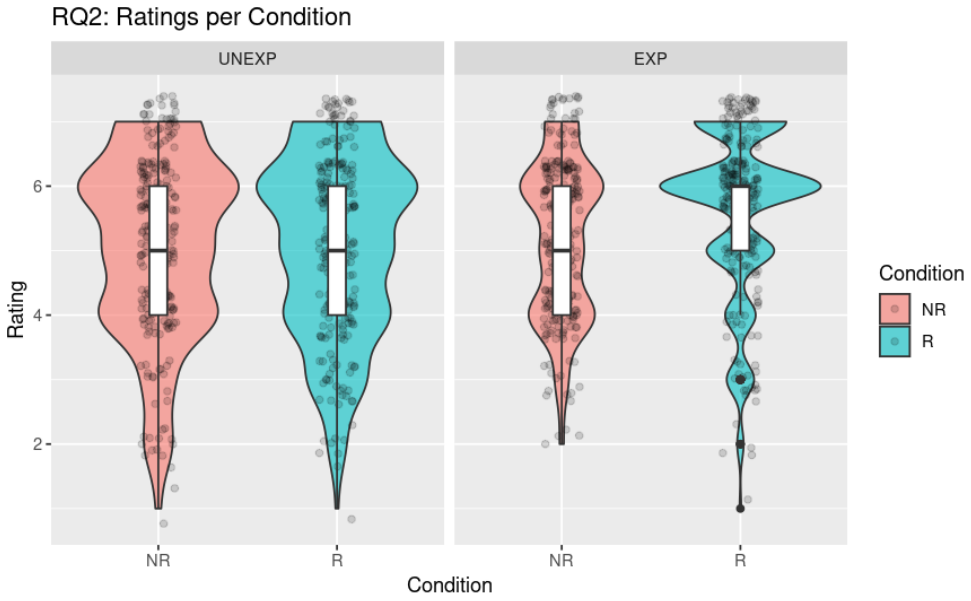


Figure 4.2: Responses to RQ2. Posterior favorability ratings increased with reasons only when the reasons were expected.

There was no credible *reduction* in listeners' posterior favorability rating in the **reasonsGiven** condition, suggesting that the higher-order inference about why the speaker wanted to explain herself was insufficient to fully counteract the argumentative strength of the reasons. However, these results still support a pragmatic analysis. It should be noted that there was no credible effect of EXPECTATION for items in the **noReasonsGiven** condition: that is, there was no significant difference between baseline approval ratings for the **reasonsExpected** and **reasonsUnexpected** items. However, the explanations improved post-utterance listener favorability ratings only for **reasonsExpected** items; they had no credible effect for **reasonsUnexpected** items. This suggests that, although there was similar "room for improvement" in both conditions (i.e. it was not the case that listeners *already* maximally approved of the **reasonsUnexpected** items), listeners were only persuaded in the **reasonsExpected** condition. If this is indeed the case, the argumentative strength of the reasons could have been dampened by the reduction of the pragmatic prior identified in Q1, as predicted by the pragmatic hypothesis.

To corroborate this analysis, a new model was constructed to examine the difference between participants' responses to the two questions (effectively Q2-Q1, but maintaining the ordinality of the data); that is, to measure the difference between the pragmatic prior (Q1 rating) and the listener's post-utterance belief (Q2 rating). The model in-

Predictor	Estimate	Std. Error	95% CI Lower	95% CI Upper
condR	-0.77	0.17	-1.11	-0.44
Q2	0.19	0.17	-0.15	0.52
typeEXP	0.24	0.68	-1.10	1.60
condR:Q2	**0.95**	**0.24**	**0.48**	**1.44**
condR:typeEXP	0.86	0.24	0.40	1.32
Q2:typeEXP	-0.03	0.24	-0.51	0.44
condR:Q2:typeEXP	-0.17	0.34	-0.87	0.49

Table 4.3: Log-odds coefficients from the three-way model. A credible interaction of QUESTION and REASONS is observed.

cluded an additional fixed effect of QUESTION number (and its interactions), with all other parameters maintained from the previous models. The new model found no credible main effect of QUESTION, meaning that the pragmatic prior (Q1 rating) and posterior (Q2 rating) were about the same when no reasons were provided. There was a credible interaction of QUESTION and REASONS: the posterior (Q2) was credibly higher than the pragmatic prior (Q1) when reasons were provided. Finally, there was no credible three-way interaction between QUESTION, REASONS, and EXPECTATION, meaning that the increase in favorability from the pragmatic prior to the listener's posterior belief was similar for both **reasonsUnexpected** and **reasonsExpected** items when reasons were provided.

Summing up, then: the Q1 model found that the that the pragmatic prior was reduced in the **reasonsGiven** condition for **reasonsUnexpected** items, but it was not reduced for the **reasonsExpected** items. The three-way model found that the *difference* between the pragmatic prior and the posterior favorability was equivalent for both **reasonsUnexpected** and **reasonsExpected** items. Therefore, items in the **reasonsExpected** set saw a net improvement from the **noReasonsGiven** condition to the **reasonsGiven** condition, whereas the items in the **reasonsUnexpected** set simply recovered from the reduced pragmatic prior and returned to "baseline", as per the findings of the Q2 model.

4.4 Notes on the free-text data

Although full quantitative analysis of the free-text responses is beyond the scope of this paper[1], some high-level analysis of the results that emerged are discussed in the following chapter.

4.5 Notes on the debrief

At the end of the survey, participants were asked to report what they thought experiment was investigating. The vast majority of respondents did not mention explanation/justification at all. Many reported that they could not even guess what the experiment was about, and several others suggested that it might be testing how the text-message medium affects tone recognition. These results indicate that participants did not trivially identify the research questions and attempt to give "correct" responses, which provides validation for the experiment design.

[1] Check back here in mid-December to see the full analysis!

Chapter 5

General Discussion

5.1 Summing up

The results of the behavioral study show that comprehenders are indeed sensitive to "over-justification", much like they are sensitive to over-informativity. They identify when more reasons are provided than expected, and derive inferences about what prior beliefs the speaker predicted them to have held, in order to make an argument worthwhile. Furthermore, these pragmatic inferences do seem to affect the strength of an argument: when more reasons are given than expected, the argument fails to influence the listener's posterior beliefs in the desired direction.

5.2 Limitations and future work

5.2.1 Non-monotonicity

Although the present findings are in support of a pragmatic analysis of argument strength, there is no evidence for true argumentative non-monotonicity. While it seems that the reduced prior did, in fact, weaken the argument for the **reasonsUnexpected** items in the **reasonsGiven** condition in comparison to the **reasonsExpected** items, it was not enough to produce post-utterance favorability ratings lower than in the **noReasonsGiven** condition. Impressionistic evaluation of the free-text data suggests some potential avenues for further work: some participants commented that it was difficult to determine whether the character made a good decision when no reasons at all were given, while others commented that many of the explanations seemed excessive or un-

necessary (as designed). Non-monotonicity might be more easily identified, then, if a third condition with just one or two reasons was added to the experiment. If participants are unconvinced when no reasons are given, and suspicious when too many are given, perhaps the optimal number is somewhere between the two extremes. A reduction in argument strength might then be observed between the "few reasons" and "many reasons" conditions.

5.2.2 Discourse vs subject-matter manipulation

Another useful qualitative finding from the free-text responses was that the discourse manipulation seemed much more effective in dismissing the "defensive" interpretation for **reasonsExpected** items than the subject-matter manipulation. Many participants commented that the speaker had explained herself because the addressee seemed curious, or wanted to know more, although some did also comment on the stakes or typicality of the decision being explained. Since it seems that the **reasonsExpected** condition might indeed be induced fully-discoursally, a more tightly controlled replication of the present study becomes possible. In brief, the expected/unexpected sets could be replaced by a manipulation of the question under discussion (QUD; Roberts, 2012). If the (explicit) QUD is "Why [action]?" or similar, at least some amount of explanation is expected, which might be enough to suppress the informativity inference, even if the action described does not usually require multiple reasons in the deliberative process. As such, the same actions and reasons could be used in both conditions, allowing for better control over item-level variability.

5.2.3 Content of the reasons

This study was primarily concerned with the absolute quantity of reasons. However, the free-text results also suggest that the *type* of reason might have an effect on the inferences that arise. Since the scenarios used in this experiment involved justifications for actions, many different "flavors" of reason are possible: the character might *want* to perform the action, or she might feel *obligated* to do it, or else she might insist that it is *permissible*, and these all concievably might generate different inferences. For example, a character over-explaining why she *wanted* to make a charitable donation generated inferences of "virtue signaling", and perhaps trying to recruit the listener to contribute as well. By contrast, a character over-explaining why he *had* to go to the library repeatedly generated the inference that he was avoiding spending time with the

listener.

5.3 Concluding Remarks

The findings of this study lend support to the idea that informative and argumentative speech are not so radically different. "Informativity expectations" are traditionally imagined as a feature of cooperative language use, going back at least to Grice, but this study shows that these same informativity expectations can give rise to inferences that allow comprehenders to recognize strategic, persuasive goals. So, while formal theories of information updating (e.g. Stalnaker, 1978) have developed quite separately from formal theories of argumentation (e.g. Anscombre and Ducrot, 1983), perhaps evidence from psycholinguistic research can provide a foundation to bridge the gap between these traditions.

Bibliography

Anscombre, J-C., & Ducrot, O. (1983). *L'argumentation dans la langue*. Mardaga.

Aylett, M., & Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence, and Duration in Spontaneous Speech. *Language and Speech*, *47*(1), 31–56. https://doi.org/10.1177/00238309040470010201

Barnett, S. A., Griffiths, T. L., & Hawkins, R. D. (2022). A Pragmatic Account of the Weak Evidence Effect. *Open Mind*, *6*, 169–182. https://doi.org/10.1162/opmi_a_00061

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, *113*(2), 1001–1024. https://doi.org/10.1121/1.1534836

Beltrama, A. (2018). Totally Between Subjectivity and Discourse. Exploring the Pragmatic Side of Intensification. *Journal of Semantics*, *35*(2), 219–261. https://doi.org/10.1093/semant/ffx021

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal Regression Models in Psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Cummins, C., & Franke, M. (2021). Rational Interpretation of Numerical Quantity in Argumentative Contexts. *Frontiers in Communication*, *6*. https://doi.org/10.3389/fcomm.2021.662027

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *37*(0). https://escholarship.org/uc/item/9wn4w9zk

Florian Jaeger, T. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62. https://doi.org/10.1016/j.cogpsych.2010.02.002

Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998–998. https://doi.org/10.1126/science.1218633

Godden, D., & Zenker, F. (2018). A probabilistic analysis of argument cogency. *Synthese*, *195*(4), 1715–1740. JSTOR. https://doi.org/10.2307/26750710

Grice, H. P. (1975). Logic and conversation. *Communications*, *30*(1), 41–58. https://doi.org/10.3406/comm.1979.1446

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, *114*(3), 704–732. https://doi.org/10.1037/0033-295x.114.3.704

Hao, H., He, M., & Fuchs, Z. (2024). Greta is a female director: When gender stereotypes interact with informativity expectations. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46. Retrieved from https://escholarship.org/uc/item/49d0685w

Kravtchenko, E., & Demberg, V. (2022). Informationally redundant utterances elicit pragmatic inferences. *Cognition*, *225*, 105159. https://doi.org/10.1016/j.cognition.2022.105159

Lemke, R., Reich, I., Schäfer, L., & Heiner Drenhaus. (2021). Predictable Words Are More Likely to Be Omitted in Fragments–Evidence From Production Data. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.662125

Lai, Catherine. (2012). Rises All the Way Up: The Interpretation of Prosody, Discourse Attitudes and Dialogue Structure. PhD thesis, University of Pennsylvania

Levy, N. (2021). *Bad Beliefs*. Oxford University Press.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems*, 849–856. https://doi.org/10.7551/mitpress/7503.003.0111

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318. https://doi.org/10.1016/j.cognition.2012.09.010

McCready, E. (2015). *Reliability in Pragmatics*. OUP Oxford.

Mercier, H. (2022). *NOT BORN YESTERDAY : the science of who we trust and what we believe.* Princeton Univ Press.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

Merin, A. (1999). "Information, Relevance, and Social Decision-Making: Some Principles and Results of Decision-Theoretic Semantics," in *Logic, Language, and Computation*. Editors L. S. Moss, J. Ginzburg, and M. de Rijke (Stanford, CA: CSLI Publications), Vol. 2, 179–221.

Oey, L. A., Schachner, A., & Vul, E. (2022). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001277

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529. https://doi.org/10.1073/pnas.1012551108

Pluymaekers, M., Mirjam Ernestus, & R. Harald Baayen. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *Journal of the Acoustical Society of America*, *118*(4), 2561–2569. https://doi.org/10.1121/1.2011150

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Reksnes, VRS, Rees, A, Cummins, C & Rohde, H 2024, Tell me something I don't know:Speaker salience and style affect comprehenders' expectations for informativity. in R Lemke, L Schäfer & I Reich (eds), Information Structure and Information Theory: Topics at the Grammar-Discourse Interface 10. Topics at the Grammar-Discourse Interface, Language Science Press, Berlin, pp. 177-214. https://doi.org/10.5281/zenodo.12784266

Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, *5*. https://doi.org/10.3765/sp.5.6

Rohde, H., Futrell, R., & Lucas, C. G. (2021). What's new? A comprehension bias in favor of informativity. *Cognition*, *209*, 104491. https://doi.org/10.1016/j.cognition.2020.104491

Rohde, H., & Rubio-Fernandez, P. (2022). Color interpretation is guided by informativity expectations, not by world knowledge about colors. *Journal of Memory and Language*, *127*, 1–10. https://doi.org/10.1016/j.jml.2022.104371

Ryzhova, M., Mayn, A., & Demberg, V. (2023). What inferences do people actually make upon encountering informationally redundant utterances? An individual differences study. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45). https://escholarship.org/uc/item/88g7g5z0

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations . *Journal of Psycholinguistic Research*, *32*(1), 3–23. https://doi.org/10.1023/a:1021928914454

Shannon, C. E. (1948). A Mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Silva, V. M., 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

Sperber, D., Clement, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Stalnaker, R. (1978). Assertion. Formal Semantics, pages 147–161.

Torabi Asr, F., & Demberg, V. (2020). Interpretation of Discourse Connectives Is Probabilistic: Evidence From the Study of But and Although. *Discourse Processes*, *57*(4), 376–399. https://doi.org/10.1080/0163853x.2019.1700760

Wallbridge, S., Bell, P., & Lai, C. (2021). It's not what you said, it's how you said it: discriminative perception of speech as a multichannel communication system. ArXiv.org. https://arxiv.org/abs/2105.00260v1

Wiegmann, A. (2023). Does lying require objective falsity? *Synthese (Dordrecht. Print)*, *202*(2). https://doi.org/10.1007/s11229-023-04291-3

Winterstein, G. (2012). What but-sentences argue for: An argumentative analysis of but. *Lingua*, *122*(15), 1864–1885. https://doi.org/10.1016/j.lingua.2012.09.014

Winterstein, G. (2018). *A Bayesian approach to Argumentation within Language.* https://semanticsarchive.net/Archive/mFkZGM0N/Argumentation.pdf