Towards a computational account of projection inferences in polar interrogatives with clause-embedding predicates

Dingyi Pan (dpan3@stanford.edu)
Symbolic Systems Program, Stanford University
Stanford, CA, 94305

Judith Degen (jdegen@stanford.edu)
Department of Linguistics, Stanford University
Stanford, CA, 94305

Abstract

Projection inferences are inferences about speaker commitment to a content embedded under an entailment-canceling operator, for example in polar interrogatives with clause-embedding predicates (*Does John know that Julian dances salsa?*). Speaker commitment to embedded content is modulated by multiple factors, including the predicate, interlocutors' prior beliefs about the content, and its at-issueness. We propose an RSA model of projection inferences in such environments. Crucially, we take the interpretive procedure to involve inferring a speaker's and attitude holder's belief in the content. In a behavioral study, we investigate inferred beliefs about contents embedded under the predicates "think" and "know" that listeners ascribe to the speaker and a potential attitude holder. We use the empirical data to parametrize the model. The resulting predictions mirror some, but not all, of the qualitative empirical patterns. This is a first step towards a systematic analysis of projection inferences using probabilistic pragmatic models.

Keywords: projection; clause-embedding predicates; experimental pragmatics; probabilistic pragmatics

Introduction

When a speaker uses a factive clause-embedding predicate like "know" in (1a) in a conversation, the listener can infer that the speaker is committed to the truth of the proposition *p*: that *Julian dances salsa*. The inference about the speaker commitment to *p* persists even when the sentence is in interrogative form, as in (1b). In this case, what is questioned is the belief of the attitude holder (i.e., John), and the ascribed speaker belief in *p* projects through the entailment-canceling environment (Kiparsky & Kiparsky, 1970).

1 a) John knows that Julian dances salsa.
b) Does John know that Julian dances salsa?

The inference that the speaker is committed to the truth of the embedded content is commonly referred to as a *projection inference*. Projection inferences have been shown to be modulated by a variety of factors, including the predicate itself (Kiparsky & Kiparsky, 1970), the at-issueness of the embedded content (Tonhauser, Beaver, & Degen, 2018; Stevens, de Marneffe, Speer, & Tonhauser, 2017), prosodic focus (Djärv & Bacovcin, 2020), and prior beliefs about the likely truth of the embedded content (Mahler, 2020; Degen & Tonhauser, 2021; Lorson, 2021).

Classic semantic accounts of projection inferences maintain that predicates categorically require projection if they are factive (e.g., "know") but not if they are non-factive (e.g., "think"), while allowing for cases that may be optionally factive (e.g., "acknowledge" Kiparsky & Kiparsky, 1970; Karttunen, 1971; Heim, 1983). Recent experimental investigations of projection inferences across a variety of lexical triggers have prompted accounts of projection inferences that treat them instead as gradient, probabilistic inferences (Tonhauser et al., 2018; Degen & Tonhauser, 2022).

For instance, Tonhauser et al. (2018) propose the Gradient Projection Principle, which predicts that more not-at-issue embedded content (content less likely to address a salient Question Under Discussion, QUD) is more likely to project. More recent work shows that the effect of the QUD is variable across predicates, suggesting that the lexical semantics of the predicate may constrain QUD effects on projection (Tonhauser & Degen, under review). The effect of prior beliefs on projection inferences is also gradient, and appears to be stable and independent of predicate or QUD (Tonhauser & Degen, under review).

As the empirical landscape on projection inferences grows denser, there is to date no formal account of how the various factors introduced above interact to generate the observed probabilistic projection patterns. In this paper, we take a first step towards providing such an account, couched in the Rational Speech Act (RSA) framework (Frank & Goodman, 2012; Goodman & Frank, 2016). RSA is a Bayesian framework that captures the probabilistic nature of belief updates and provides a systematic way to analyze factors that affect pragmatic inferences. We empirically investigate patterns of belief attributions to speakers and attitude holders in response to interrogatives with two clause-embedding predicates – "think" and "know" – as a first step towards informing a formal account of projection inferences in these environments. We use the collected data to inform an interpretation model for polar interrogatives with clause-embedding predicates. We show that systematic effects of prior beliefs on projection inferences fall naturally out of the model, and that reasoning about alternative utterances can give rise to some of the observed variability in projection inferences.

Computational model

The Rational Speech Act (RSA) framework (Frank & Goodman, 2012; Goodman & Frank, 2016; Degen, 2023) includes a family of probabilistic models that formalize language pro-

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

duction and interpretation as recursive reasoning between speaker and listener. Interlocutors are assumed to be rational and soft-maximize the utility of utterances and interpretations, respectively, calculated against alternative interpretations a listener might arrive at or alternative utterances the speaker could have produced, respectively. RSA has been used to model a variety of production and interpretation side phenomena in semantics and pragmatics, including implicature (Bergen, Goodman, & Levy, 2012), non-literal language use phenomena (Kao, Bergen, & Goodman, 2014; Kao, Wu, Bergen, & Goodman, 2014), the interpretation of gradable adjectives (Lassiter & Goodman, 2017), and quantifier scope ambiguity (Attali, Scontras, & Pearl, 2021). Despite its success on a wide range of phenomena, applications to presupposition projection are few and far between. Notable exceptions include work on projective content under negation in utterances with the change-of-state verb "stop" (e.g., "John did not stop smoking" presupposes that John used to smoke, Qing, Goodman, & Lassiter, 2016) and in utterances with manner adverbs (e.g., "Masha didn't run quickly" presupposes that Masha ran, Stevens et al., 2017).

In a standard RSA model, the *pragmatic speaker* produces an utterance proportional to its utility, compared to alternatives. An utterance's utility is based on the informativeness and the cost of that utterance, where informativeness is defined via the probability that a *literally interpreting listener* would correctly infer the intended meaning. Upon observing an utterance, the *pragmatic listener* chooses an interpretation of the utterance probabilistically by integrating their prior beliefs about likely meanings with their expectations about the pragmatic speaker using Bayes' rule.

Here, we take a first step towards providing a predictive model of projective content embedded under clause-embedding predicates in interrogative utterances. In particular, we provide a partial model of beliefs attributed to speakers and attitude holders upon observing polar interrogatives containing the canonically factive verb "know" and the non-factive verb "think", as well as the unembedded polar interrogative without a clause-embedding predicate.

The choice to model speaker and attitude holder beliefs instead of formalizing the notion of *speaker commitment* typically invoked to explain projection phenomena is motivated by the following considerations. First, the notion of a speaker commitment itself is unclear. There is ample evidence from the experimental pragmatics literature that listeners track speaker beliefs; what it would mean to track a speaker commitment is less clear. Speaker commitments can be thought of as propositions added to the common ground that a speaker's utterance suggests they are taking for granted. Common ground update is indeed the main mechanism of the only fully worked out RSA model of presupposition projection to date (Qing et al., 2016). However, this model does not straightforwardly generalize or extend to the case of clause-embedding predicates. We thus instead choose to model listener inferences about speaker beliefs about the embedded content directly. Future work should explore whether a common ground update model better captures the data.

However, inferences about speaker beliefs alone will not suffice to capture the observed range of factors modulating projection inferences. In particular, effects of content at-issueness on projection inferences suggest that representing the possibility for addressing variable contextual QUDs is important. The plausible QUDs a speaker might address using an interrogative with a clause-embedding predicate are plentiful. For the example in (1), one plausible QUD targets the truth of the embedded content, i.e., whether Julian dances salsa. Another QUD targets John's attitude towards the embedded content. Other QUDs are possible by focusing one of the lexical items or constituents in the embedded clause. To put a lid on the overflowing cup of possible QUDs, we suggest that two that are most relevant for the purpose of understanding the behavior of the cognitive predicates "know" and "think" are whether the speaker and the attitude holder believe the embedded content p. While we do not model the effect of QUDs directly, these considerations inspire our assumptions about the relevant space of meanings.

Any RSA model requires being explicit about two crucial components: the *set of utterance alternatives* a speaker is assumed to make a choice between, and the possible *set of meanings* to communicate in context. To our knowledge, the literature on presupposition projection, by virtue of typically considering projection a semantic phenomenon triggered by a specific class of (factive) predicates, does not discuss plausible alternatives to the observed utterance that a listener may take into account in their reasoning about what beliefs to attribute to the speaker. Our first substantial contribution is thus to make an explicit assumption in that regard. In particular, we assume that the set of alternatives is $U =$ {"know p", "know not p", " think p", "think not p", and "p"}, where each u in U is an interrogative with a particular predicate ("know," "think") and an embedded clause in either affirmative or negated form (e.g., "Does John think that Julian dances salsa?", "Does John know that Julian doesn't dance salsa?"). We also include the simple polar interrogative without embedding (e.g., "Does Julian dance salsa?").[1]

We assume that the space of meanings consists of belief tuples $\langle b_{SP}, b_{AH} \rangle$, where $b_{SP}, b_{AH} \in \{p, \neg p, ?\}$: the speaker and the attitude holder can either believe p, its negation, or be uncertain. For instance, $\langle p, \neg p \rangle$ represents that the speaker believes p but the attitude holder believes its negation. In addition, because the unembedded polar interrogative does not introduce an attitude holder, we assume the attitude holder's belief can also be undefined ($\varnothing$). There are a total of 12 possible states: $M = \{\langle p,p\rangle, \langle p,\neg p\rangle, \langle p,\varnothing\rangle, \langle p,?\rangle, \langle \neg p,p\rangle, \langle \neg p,\neg p\rangle, \langle \neg p,\varnothing\rangle, \langle \neg p,?\rangle, \langle ?,p\rangle, \langle ?,p\rangle, \langle ?,\varnothing\rangle, \langle ?,?\rangle \}$.

[1] We do not include the negation of the unembedded polar interrogative, because doing so requires making a judgment call about whether to use high or low negation (e.g., "Doesn't Julian dance salsa?" vs. "Does Julian not dance salsa?"), which have been argued to convey different speaker biases with respect to whether Julian dances salsa (Ladd, 1981; Buring & Gunlogson, 2000).

Given an utterance $u \in U$, the pragmatic listener infers a meaning (belief state) $m \in M$ using Bayes' rule:

$$\underbrace{P_{L_1}(\langle b_{SP}, b_{AH}\rangle | u)}_{\text{empirically elicited}} \propto \underbrace{P_{S_1}(u | \langle b_{SP}, b_{AH}\rangle)}_{\text{inferred (BDA)}} \cdot \underbrace{P(b_{SP}) \cdot P(b_{AH})}_{\text{re-used norms}} \quad (1)$$

The pragmatic listener L_1 considers their expectations about how likely a speaker S_1 with a belief about p and a belief about an attitude holder's belief about p would be to produce the observed utterance, and multiplies that probability with the assumed prior beliefs of speaker and attitude holder.

The usual next step in building an RSA model is to spell out the generative model that characterizes how the speaker is assumed to reason about a literal listener who interprets utterances according to their lexical truth-conditional semantics. However, in this domain the empirical landscape regarding key components to include in the model is vastly underdetermined: neither the set of utterance alternatives nor the relevant aspects of the lexical semantics of clause-embedding predicates have been investigated in enough detail to allow for implementing a model with well motivated assumptions.[2]

We thus take a different route here and ask: for a restricted set of predicates, under the meaning space laid out above, is there a pragmatic speaker distribution *in principle* that, when combined with known prior beliefs in the way specified by the pragmatic listener rule, can generate a pragmatic listener distribution that conforms with empirical interpretation judgments? If we discover such a distribution, we can investigate and develop the additional model components in future work.

To assess whether such a pragmatic speaker distribution exists, we collected pragmatic listener probabilities in an experiment that elicited judgments of ascribed speaker and attitude holder beliefs. We then used these probabilities, in conjunction with previously empirically collected prior belief norms (Degen & Tonhauser, 2021), to infer a hard-coded pragmatic speaker distribution via Bayesian data analysis (BDA). We first report the experiment, followed by the BDA.

Experiment: belief ascription

Method

Participants We recruited 360 participants on Prolific. Based on pre-registered exclusion criteria, 345 participants were included in the analysis.[3]

Materials and procedure In addition to the two critical predicates "think" and "know" as well as the unembedded polar interrogative, "say" and "inform" were included as control items, which are communicative predicates with similar projection contrasts as the cognitive predicates of interest (Anand & Hacquard, 2014; Schlenker, 2008).

[2]Not to mention the vast number of possible QUDs the listener might consider in interpreting a sentence like "Does John know that Julian dances salsa?"; or the variability in production costs one might want to ascribe to utterances with and without embedded clauses, and with affirmative vs. negated embedded clauses.

[3]The pre-registration is available at `https://osf.io/gtdw5`. All data, materials, and analysis scripts can be accessed at `https://github.com/pennydy/Projectivity_RSA`.

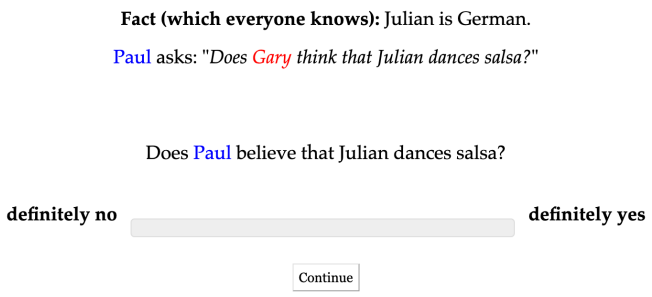

Figure 1: Example of a trial asking about speaker belief, given a low probability fact that makes the embedded content unlikely.

Eighteen critical items from Degen and Tonhauser (2021) were used as the embedded contents (e.g., Julian dances salsa). These were paired with facts that made the content either likely or unlikely a priori (e.g., "Julian is Cuban" vs. "Julian is German," respectively; norms also taken from Degen & Tonhauser, 2021). For each participant, half of the critical items were paired with the high probability fact and half with the low probability fact. For each type of prior, the unembedded polar interrogative was presented once, and each predicate ("know", "think", "say", and "inform") was presented twice, once with the affirmative embedded clause ("p") and once with the negated embedded clause ("not p"). We randomized the pairing between content and predicate, such that each participant saw each of the 18 critical items exactly once, and each utterance was randomly paired with a speaker name and an attitude holder name. In addition, we included 6 control items from Degen and Tonhauser (2021), which were unembedded interrogatives with presumably unbiased prior content probabilities. Each participant completed 24 trials.

Participants were instructed to imagine that they walked into a kitchen and overheard somebody asking another person a question. The uttered interrogative was displayed with a fact that was presented as common knowledge (see Fig. 1). To assess ascribed speaker and the attitude holder belief, participants were instructed to provide a rating on a slider with endpoints labeled "definitely no" (coded as 0) and "definitely yes" (coded as 1) in response to two questions that used the carrier sentence "Does SPEAKER/ATTITUDE HOLDER believe . . . ?". The name of the speaker and the name of the attitude holder were color-coded and matched in utterance and rating question to minimize referential confusion. To avoid a potential effect of question order, we randomized question order across participants, such that some participants rated the speaker's belief first, whereas others rated the attitude holder's belief first. Attitude holder beliefs were not rated for unembedded polar interrogatives because they lack an attitude holder to ask about.

Results

While we are primarily interested in the empirical patterns for the purpose of informing the cognitive model (see next sec-

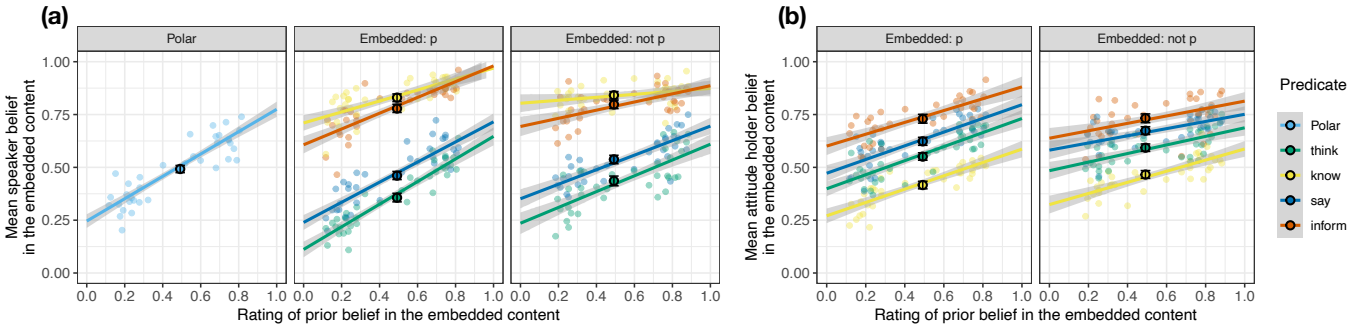

Figure 2: Mean speaker belief ratings (a) and mean attitude holder belief ratings (b) against the prior rating of the embedded content, by predicate and embedded clause type. Each translucent dot represents an item mean, and each larger solid dot with a black border represents the grand mean of corresponding ratings in that condition. The error bars on grand means (very small) and the shaded ribbon represent bootstrapped 95% confidence intervals.

tion), we nevertheless describe the effects of the experimental manipulations on ascribed speaker and attitude holder belief ratings, respectively. We limit ourselves to the discussion of just the critical predicates "think" and "know" and the unembedded polar interrogative.

Speaker beliefs Fig. 2(a) shows mean belief ratings ascribed to the speaker. We conducted two Bayesian mixed effect linear regressions to assess whether predicate, embedded content form, and prior affected ascribed speaker beliefs.[4]

One regression was conducted on just the affirmative embedded content subset of the data (2070 observations). The model predicted belief rating from fixed effects of predicate (reference level: "Polar"), centered prior belief rating, and their interaction, as well as the maximal random effects structure that allowed the model to converge (random by-item and by-participant intercepts and slopes for predicate and prior). The model included weakly informative default priors.

We consider an effect significant if 0 is not included in the CrI. Compared to the unembedded polar interrogative, participants ascribed greater belief in p to the speaker when p was embedded under "know" ($\beta = 0.34, CrI = [0.31, 0.36]$) and lower belief in p when it was embedded under "think" ($\beta = -0.13, CrI = [-0.15, -0.10]$). In addition, there was a significant main effect of prior belief, such that participants ascribed greater belief in p to the speaker, the more a priori likely p was ($\beta = 0.54, CrI = [0.46, 0.61]$). A significant negative interaction between the "know" predicate contrast and prior suggests the prior effect was slightly smaller with "know" than with the unembedded polar interrogative ($\beta = -0.28, CrI = [-0.36, -0.20]$).

A second regression was conducted on just the subset of the data without the unembedded polar interrogatives (2760 observations). The model predicted belief rating from fixed effects of predicate (reference level: "know"), centered prior belief rating, centered embedded clause type (reference level before centering: "affirmative") and their interactions, as well as the maximal random effects structure that allowed the model to converge (random by-item and by-participant intercepts and slopes for predicate, prior, and embedded clause type). The model included weakly informative default priors.

Replicating the predicate effect from the first regression, the belief ascribed to the speaker was significantly lower for "think" than for "know" ($\beta = -0.45, CrI = [-0.47, -0.42]$). Also replicating the prior effect from the first regression, there was a significant effect of the prior ($\beta = 0.18, CrI = [0.11, 0.24]$), which was larger for "think" (evidenced in a significant interaction between the "think" contrast with prior, $\beta = 0.29, CrI = [0.22, 0.36]$). There was no significant effect of embedded clause type when the predicate was "know," but there was a significant interaction between the "think" contrast and embedded clause type, such that participants ascribed greater belief in the embedded content to the speaker under "think" when the embedded clause was negated ($\beta = 0.05, CrI = [0.01, 0.08]$). Moreover, there was a significant interaction between prior and embedded clause type, such that the effect of the prior was smaller when the embedded clause was negated ($\beta = -0.19, CrI = [-0.28, -0.09]$).

There was a significant effect of predicate, such that ascribed belief in p was higher for "think" than for "know" ($\beta = 0.13, CrI = [0.11, 0.15]$). This was the opposite of the ascribed speaker belief pattern. There was also a significant effect of the prior, such that attitude holders were judged to be more likely to believe contents with higher prior probability ($\beta = 0.29, CrI = [0.24, 0.35]$). Finally, there was a marginal effect of embedded clause type, such that negated embedded clauses received higher belief ratings ($\beta = 0.03, CrI = [0.00, 0.05]$). None of the interactions reached significance.

Attitude holder beliefs Fig. 2(b) shows mean belief ratings ascribed to the attitude holder. We conducted the same analysis type as we did on speaker ratings, but were able to include all three predictors in a single analysis because attitude holder beliefs were not rated for polar interrogatives.

[4]Running two models per dataset was necessary because of the principled choice to only include affirmative unembedded interrogatives. We thus couldn't run one model with the full three-way interactions. Models were run using the `brms` package (Bürkner, 2021).

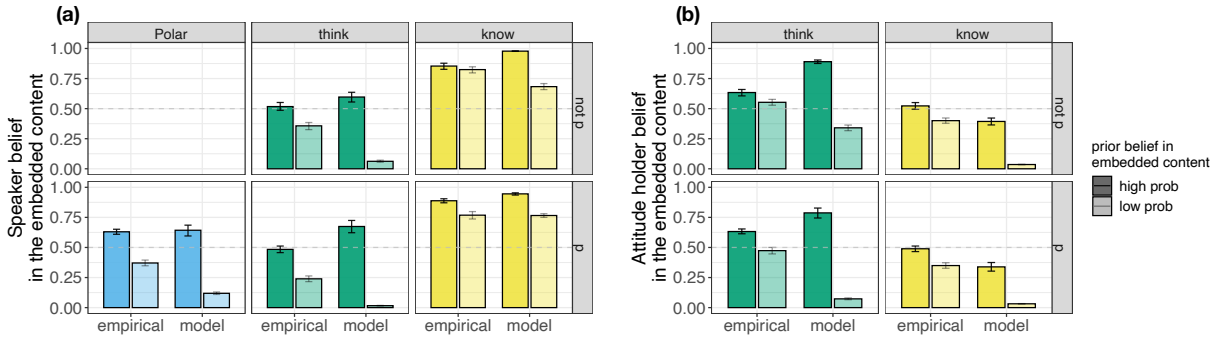

Figure 3: Mean empirical and model predicted belief ratings for speaker (a) and attitude holder (b) with high and low prior probability facts. Rows indicate embedded clause type. Error bars represent bootstrapped 95% confidence intervals.

Discussion

If we interpret ascribed speaker beliefs as standing proxy for speaker *commitments* typically taken to be indicative of projection inferences, these results replicate previously established intuitions and empirical results showing that content embedded under the canonically factive verb "know" is more projective than under the non-factive "think," and that prior beliefs modulate projection inferences (Tonhauser et al., 2018; Degen & Tonhauser, 2021, 2022).

These results also contain novel data points: compared to the unembedded polar baseline, "think" appears to have an anti-veridical effect – speakers are judged to be less likely to believe the embedded content than if use of "think" simply indicated complete uncertainty, in which case participants should have defaulted to the prior. Interestingly, embedding a negated clause eliminates that anti-veridical bias. This may be the result of pragmatic competition of interrogative "think p" with "know p," which allows the speaker to directly signal belief in p, whereas competition between "think not p" and "know not p" may be reduced due to the cost or more complex licensing conditions of the negated embedded clause.

This experiment is the first to investigate the beliefs ascribed to attitude holders. The result that belief attributions were flipped for "think" and "know," compared to speaker belief attributions, is surprising under standard semantic accounts which hold that knowing p entails thinking p. An interesting avenue for future work is to explore whether these patterns can be derived from considerations of the lexical semantics of these predicates and their pragmatic licensing conditions in interrogative rather than declarative sentences.[5]

Calibrating the computational model

To reverse-engineer a pragmatic speaker distribution that, if reasoned about by a pragmatic listener according to Equation (1), would give rise to the observed patterns in the empirical experiment, we conducted a Bayesian data analysis (BDA). We used the speaker and attitude holder belief ratings from the behavioral experiment and the prior norms for items that were used in the experiment from Degen and Tonhauser (2021) as proxy for the pragmatic listener $P_{L_1}(\langle b_{SP}, b_{AH}\rangle|u)$ and prior $P(b_{SP/AH})$[6] terms, respectively.[7]

To facilitate inference, we discretized both the prior belief ratings and the belief ratings collected in the behavioral experiment into three bins corresponding to the primitives of the meaning space ($b < .4$: $\neg p$, $.4 \leq b \leq .6$: ?, $b > .6$: p).

Given the assumption about the meaning space and the proposed structure of the pragmatic speaker as defined in Equation (1), a total of 60 parameters (12 possible belief states $\times$ 5 utterances) were estimated, each of which represents the production probability of an utterance given a particular belief state. Each parameter was sampled from a uniform prior over the interval [0,1] with a drift kernel. We used Markov Chain Monte Carlo (MCMC) sampling to collect 1000 samples with a burn-in of 500 and a lag of 10 and ran the analysis with 5 different random seeds.

While it needn't have been so, there is indeed a pragmatic speaker distribution that generates the qualitative patterns in the pragmatic listener judgments observed in the experiment. We first discuss the resulting pragmatic listener predictions, followed by the reverse-engineered production distribution.

Pragmatic listener The model-predicted mean beliefs attributed to speaker and attitude holder under the different predicates, embedded clause type, and prior belief are shown alongside the empirical means in Fig. 3. The model qualitatively captures the overall patterns in the behavioral results: the speaker is considered to be more committed to content when it is embedded under "know" than under "think," consistent with the behavioral results. Moreover, the prior modulates both speaker and attitude holder beliefs in the direction observed empirically. However, the empirical effect is systematically smaller than predicted by the model across all conditions. The discrepancy between predicted and actual

[5] Support for this idea comes from a separate experiment with declarative sentences, which yielded a significant effect of predicate in both ascribed speaker and attitude holder beliefs in the expected direction: higher ratings for "know" than for "think."

[6] We assumed that speaker and attitude holder share prior beliefs.

[7] The possible attitude holder beliefs and speaker beliefs differ slightly because the unembedded polar interrogative does not introduce an attitude holder. Thus, $b_{SP} \in \{p, \neg p\}$ whereas $b_{AH} \in \{p, \neg p, \varnothing\}$, where $\varnothing$ represents the lack of an attitude holder. We treat $b_{AH} = \varnothing$ as very unlikely with a probability of 0.05.

prior effect is greatest for attitude holder beliefs; and greater for negated than affirmative embedded clauses. Possible explanations for this weaker than predicted empirical effect include: 1) the listener may be uncertain about the prior knowledge of the speaker (and even more so, of the attitude holder), and thus the ultimately ascribed belief state may be a mixture of the empirical prior and a uniform prior representing uncertainty (see Degen, Tessler, & Goodman, 2015, for discussion of a similar issue in the domain of scalar implicature); 2) the prior is just one of many sources of information that listeners combine in the interpretation of complex sentences with clause-embedding predicates. It is possible that the more additional reasoning listeners must engage in (e.g., to explain away the use of a negation, which can be cognitively costly, Wales & Grieve, 1969; Kaup & Zwaan, 2003), the more resources are taken away from processing prior information.

Pragmatic speaker The pragmatic speaker distribution that was used to generate the pragmatic listener predictions is shown in Fig. 4. It has at least four interesting explanatory features. First, when attitude holder belief is undefined ($\varnothing$), the speaker strongly prefers to use the unembedded polar interrogative, especially when the speaker is uncertain about the embedded content, in line with expected licensing conditions.

Second, the speaker prefers to use "know" when they believe the embedded content and are uncertain about the attitude holder's belief. This is what allows the pragmatic listener to faithfully recover the speaker's belief from the observed use of "know." In cases where the speaker is uncertain about the proposition, "know" is strongly dispreferred.

In all other situations, the speaker is more likely to use "think" than any other utterance alternative. The preference for "think p" compared to alternatives when $b_{SP} = \neg p$ (second row) gives rise to the anti-veridical effect at the pragmatic listener level. Moreover, the overall preference for "think" across the board (with the above exceptions) explains the overall greater observed uncertainty about speaker beliefs for "think" compared to "know." The widespread use of "think" is consistent with there being many reasons for a speaker to produce "think p" – they may be incredulous that John thinks that Julian dances salsa because they know he doesn't ($\langle \neg p, p \rangle$); they may be uncertain about whether Julian dances salsa and expect that John is at least somewhat of an authority on the matter ($\langle ?, p \rangle$,$\langle ?, ? \rangle$); or they may already believe that Julian dances salsa and be fairly certain that John believes so as well, and simply seek confirmation ($\langle p, p \rangle$).

Overall, the marginal probability of producing a negated embedded clause is lower than that of producing an affirmative embedded clause, possibly reflecting a production cost on negation. The effects of this anti-negation preference on the pragmatic listener are to be explored further.

General discussion

In this paper, we took a first step towards proposing a probabilistic interpretation model of projection inferences about content embedded under clause-embedding predicates occurring in polar interrogative sentences. To inform the model, we collected attributed speaker and attitude holder beliefs in a behavioral experiment that varied predicate, embedded clause type, and prior beliefs about the embedded content. The results replicated the effect of prior beliefs on projection inferences and predicate-based variability in inferred speaker commitment. In addition, speakers who produce an interrogative with "think" are considered to believe more strongly in the opposite of the embedded content, suggesting that "think" may be inherently anti-veridical.

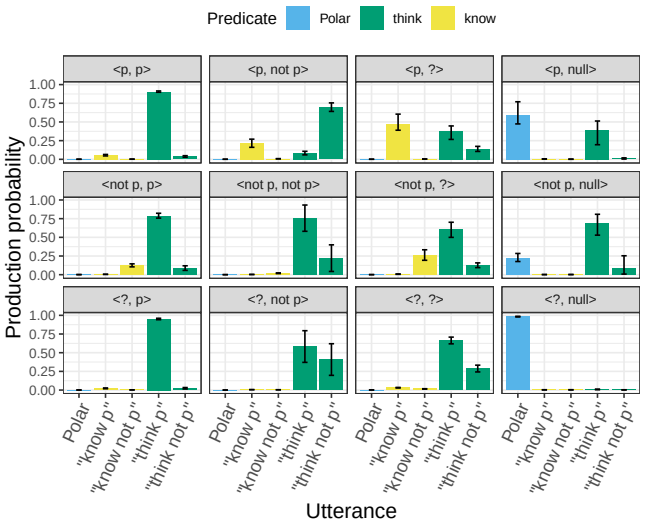

Figure 4: Reverse-engineered pragmatic speaker production probabilities, faceted by meaning states $\langle b_{SP}, b_{AH} \rangle$. Error bars indicate 95% bootstrapped confidence intervals based on results of inference chains from 5 different random seeds.

Using the belief ratings from the behavioral experiment and previously elicited content prior belief norms, we conducted a Bayesian data analysis to reverse-engineer the pragmatic speaker distribution that could have given rise to the observed listener probabilities. The explained qualitative patterns include the high baseline projection rate of "know," the anti-veridicality of "think," and speaker uncertainty as a licensing condition for the unembedded polar interrogative. However, the pragmatic listener model predictions overpredicted the effect of the prior.

A key limitation of this work is that we have only provided a proof of concept that a production distribution exists in principle that can give rise to the observed interpretation side inferences. Future work should use production tasks to validate the inferred distribution. Additional work should investigate whether and how the reverse-engineered or empirically elicited production distribution can be modeled as a speaker pragmatically choosing from among a set of utterance alternatives to address a contextually salient QUD. Implementing different hypotheses about the lexical semantics of the utterance alternatives under consideration and the possible QUDs interlocutors might address are important next steps in providing a unified account of projection inferences.

Acknowledgements

We would like to thank Judith Tonhauser, Brandon Waldon, Todor Koev, and the ALPS Lab for helpful comments and discussion. We are also grateful to Brandon Waldon for help with the Bayesian data analysis.

References

Anand, P., & Hacquard, V. (2014). Factivity, belief and discourse. In L. Crnič & U. Sauerland (Eds.), *The art and craft of semantics: A festschrift for irene heim* (Vol. 1, pp. 69–90). Cambride, MA: MIT Working Papers in Linguistics.

Attali, N., Scontras, G., & Pearl, L. S. (2021). Pragmatic factors can explain variation in interpretation preferences for quantifier-negation utterances: A computational approach. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Bergen, L., Goodman, N., & Levy, R. (2012). That's what she (could have) said: How alternative utterances affect language use. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34, p. 120-125).

Buring, D., & Gunlogson, C. (2000). Aren't positive and negative polar questions the same? Retrieved from `http://hdl.handle.net/1802/1432`

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. doi: 10.18637/jss.v100.i05

Degen, J. (2023). The Rational Speech Act Framework. *Annual Review of Linguistics*, *9*, 519-540.

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*(2), 548–553.

Degen, J., & Tonhauser, J. (2021). Prior beliefs modulate projection. *Open Mind*, *5*, 59-70. doi: 10.1162/opmi_a_00042

Degen, J., & Tonhauser, J. (2022). Are there factive predicates? an empirical investigation. *Language*, *98*(3), 552-591. doi: doi:10.1353/lan.0.0271.

Djärv, K., & Bacovcin, H. A. (2020). Prosodic effects on factive presupposition projection. *Journal of Pragmatics*, *169*, 61-85. doi: https://doi.org/10.1016/j.pragma.2020.04.011

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818-829.

Heim, I. (1983). On the projection problem for presuppositions. In D. F. et al (Ed.), *Proceedings of the second west coast conference on formal linguistics (wccfl)* (pp. 114–125). Stanford, Cali: Stanford University Press.

Kao, J. T., Bergen, L., & Goodman, N. (2014). Formalizing the pragmatics of metaphor understanding. In *Proceedings of the 36th annual meeting of the cognitive science society* (Vol. 36).

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Karttunen, L. (1971). *The logic of English predicate complement constructions.* Indiana University Linguistics Club, Bloomington.

Kaup, B., & Zwaan, R. A. (2003). Effects of negation and situational presence on the accessibility of text information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(3), 439.

Kiparsky, P., & Kiparsky, C. (1970). Fact. In *Progress in linguistics* (pp. 143–173). Berlin, Boston: De Gruyter Mouton. doi: 10.1515/9783111350219.143

Ladd, D. R. (1981). A first look at the semantics and pragmatics of negative questions and tag questions. In *Papers from the... regional meeting. chicago ling. soc. chicago, ill* (pp. 164–171).

Lassiter, D., & Goodman, N. D. (2017). Adjectival vagueness in a bayesian model of interpretation. *Synthese*, *194*, 3801–3836.

Lorson, A. (2021). *The influence of world knowledge on projectivity.* (Unpublished master's thesis, The University of Potsdam)

Mahler, T. (2020). The social component of the projection behavior of clausal complement contents. *Proceedings of the Linguistic Society of America*, *5*(1), 777–791.

Qing, C., Goodman, N. D., & Lassiter, D. (2016). A rational speech-act model of projective content. In *Proceedings of the thirty-eighth annual conference of the cognitive science society.*

Schlenker, P. (2008). Be articulate: A pragmatic theory of presupposition projection.

Stevens, J., de Marneffe, M.-C., Speer, S. R., & Tonhauser, J. (2017). Rational use of prosody predicts projection in manner adverb utterances. In *Proceedings of the thirty-ninth annual conference of the cognitive science society.*

Tonhauser, J., Beaver, D. I., & Degen, J. (2018). How projective is projective content? gradience in projectivity and at-issueness. *Journal of Semantics*, *35*(3), 495–542.

Tonhauser, J., & Degen, J. (under review). Prior beliefs and at-issueness independently modulate projection. Retrieved from `https://ling.auf.net/lingbuzz/006771`

Wales, R., & Grieve, R. (1969). What is so difficult about negation? *Perception & Psychophysics*, *6*, 327–332.