Projection without lexically-specified presupposition: A model for *know*[1]

Gregory SCONTRAS — *University of California, Irvine*
Judith TONHAUSER — *University of Stuttgart*

Abstract. We present an analysis—formalized as a computational cognitive model—of the projection of the content of the clausal complement of *know* in utterances of negated declaratives. Our model, formulated within the Bayesian Rational Speech Act framework (Frank and Goodman, 2012), derives projection from lexical entailments of *know* and sensitivity to the Question Under Discussion (QUD; as do Abrusán, 2011, and Simons, Beaver, Roberts, and Tonhauser, 2017), as well as reasoning about utterance informativity relative to private speaker assumptions (Qing, Goodman, and Lassiter, 2016; Warstadt, 2022). Crucially, our model predicts projection for *know* without encoding the inference via a lexically-specified constraint on the Common Ground. The model goes beyond existing analyses by also making predictions about the projection of the content of the clausal complement of nonfactive *think*, as well as other types of inferences. We find support for three qualitative predictions of our model in two experiments that measured the projection inferences participants draw about utterance content.

Keywords: projection inferences, presupposition, factivity, Rational Speech Act models.

1. Introduction

From the underlined sentence in the naturally-occurring example in (1), interpreters may infer that Estonia includes hundreds of islands, even though this content is contributed by the clausal complement of *know* in the scope of sentential negation. This so-called "projection" inference is classically derived by coding the content of the clausal complement of *know* as content that is required to be entailed by or satisfied in the Common Ground of the interlocutors prior to the interpretation of the negated *know*-sentence (e.g., Heim, 1983; van der Sandt, 1992). On such analyses, the content projects because it is a lexically-specified constraint on the Common Ground that must be satisfied prior to the interpretation of the negated *know*-utterance.

(1) Geographically, Europe can be divided in two parts: Europe of the hare and Europe of the squirrel. Estonia is the latter: a squirrel could, without too much trouble, go from one end of the country to the other, jumping from tree to tree. Of course, the squirrel doesn't know that Estonia includes hundreds of islands, small and big.[2]

Examples like (1) challenge analyses on which the content of the clausal complement of *know* is lexically coded as a constraint on the Common Ground because many readers of (1) do not know that Estonia includes hundreds of islands prior to interpreting the negated *know*-sentence. In other words, the relevant content is not part of the common ground of the interlocutors prior to interpretation. To address this challenge, analyses like Heim (1983) and van der Sandt (1992) resort to a rescue strategy: The presupposed content may be globally accommodated to the common ground of the interlocutors prior to interpretation. Global accommodation is assumed to be the default, with local accommodation (e.g., in (1) in the scope of negation) applying only

[1]For helpful feedback, we thank the audiences at the University of Stuttgart and at *Sinn und Bedeutung* 29, as well as the reviewers for *Sinn und Bedeutung* 29.

[2]https://www.theguardian.com/world/2004/apr/26/eu.politics6

when global accommodation would lead to inconsistencies or problems with binding.

Investigations of presuppositions in naturally-occurring discourse, like Delin (1992) and Spenader (2002), have shown, however, that presuppositions contributed by a variety of expressions are frequently not part of the common ground at the time at which the triggering expression is encountered. For factive predicates like *know*, Spenader found that the contents of the clausal complements of the majority of the utterances of sentences with factive verbs (i.e., 81 out of 109) had to be accommodated. That accommodation, a rescue strategy, has to apply in the majority of examples renders analyses like Heim (1983) or van der Sandt (1992) unsatisfying.

In this paper, we present an analysis that derives the projection of the content of the clausal complement of *know* without coding it as a presupposition in the lexical entry of *know*. Instead, the projection inference is derived from the lexical entailments of *know* and sensitivity to the Question Under Discussion (QUD; see also Abrusán, 2011, and Simons et al., 2017), as well as reasoning about utterance informativity relative to private speaker assumptions. Our analysis, which is formulated as a computational cognitive model in the Bayesian Rational Speech Act (RSA) framework, draws significant inspiration from the model of Qing et al. (2016) (see also Warstadt, 2022, as well as Roberts and Simons, 2024, for a suggestion along these lines). Our model goes beyond existing projection analyses by also making predictions about the effects of prior beliefs on projection inferences (see, e.g., Mahler, 2020, 2022; Degen and Tonhauser, 2021), about the projection of the content of the clausal complement of nonfactive *think* (see, e.g., Simons et al., 2017; Degen and Tonhauser, 2022), as well as inferences other than projection inferences.

In Section 2, we begin by establishing the empirical targets that we aim to hit with our analysis. We then present our analysis in Section 3, showing that the model makes qualitatively-appropriate predictions regarding the empirical targets established in Section 2. Section 3 also compares our analysis to prior ones and points out that none of the prior analyses hit all three of the empirical targets we set out for ourselves. In Section 4, we show that a extension of the analysis presented in Section 3 makes adequate predictions for a broader range of inferences than projection inferences. We take this result to be a further advantage of our approach to deriving projection inferences. Section 5 concludes.

2. Establishing the empirical targets

We begin with the empirical targets we aim to hit. In (2) we present three hypotheses developed on the basis of prior theoretical and empirical research on projection.

(2) Empirical hypotheses:
   a. The content of the complement of *know* is more projective than that of *think*.
   b. The contents of the complements of *know* and *think* are more projective when they have a higher prior probability than when they have a lower prior probability.
   c. The contents of the complements of *know* and *think* are more projective when they are not-at-issue than when they are at-issue.

The hypothesis in (2a) derives from the long-standing assumption in the literature that the content of the clausal complement of *know* is more projective than that of *think* (Kiparsky and Kiparsky, 1970, i.a.), as well as empirical support from experiments (e.g., Djärv and Bacovcin,

2020; Degen and Tonhauser, 2022). Regarding the hypothesis in (2b), Degen and Tonhauser (2021) provided empirical evidence that the projection of the contents of the clausal complements of *know* and *think* are sensitive to the prior beliefs of the interlocutors (see Mahler, 2022, for further discussion). Finally, regarding the hypothesis in (2c), Tonhauser et al.'s (2018) Gradient Projection Principle leads to the expectation that projection is sensitive to at-issueness, such that content that is more not-at-issue is also more projective.[3]

To validate these empirical hypotheses and establish the qualitative empirical targets for our analysis, we ran two behavioral experiments.[4] In both experiments, participants read a short story that ended with one of six utterances of interest, like those in (3), including the negated declarative with *know* in (3b).

(3) a. Cole knows that Charley speaks Spanish.
b. Cole doesn't know that Charley speaks Spanish.
c. Cole thinks that Charley speaks Spanish.
d. Cole doesn't think that Charley speaks Spanish.
e. Charley speaks Spanish.
f. Charley doesn't speak Spanish.

The story manipulated the at-issueness of the content C (that Charley speaks Spanish in (3)) via the QUD (BEL?: whether Cole believes that Charley speaks Spanish vs. C?: whether Charley speaks Spanish) and, only in Exp. 1, the prior probability (more vs. less likely) of C. Participants rated the likelihood of the content after reading the story and the target utterance.

2.1. Experiment 1

Exp. 1 was designed to investigate all three hypotheses in (2): (a) comparing the projection for *know* vs. *think*, (b) comparing the projection for higher vs. lower prior probability of content, and (c) comparing the projection for at-issue vs. not-at-issue content.

2.1.1. Methods

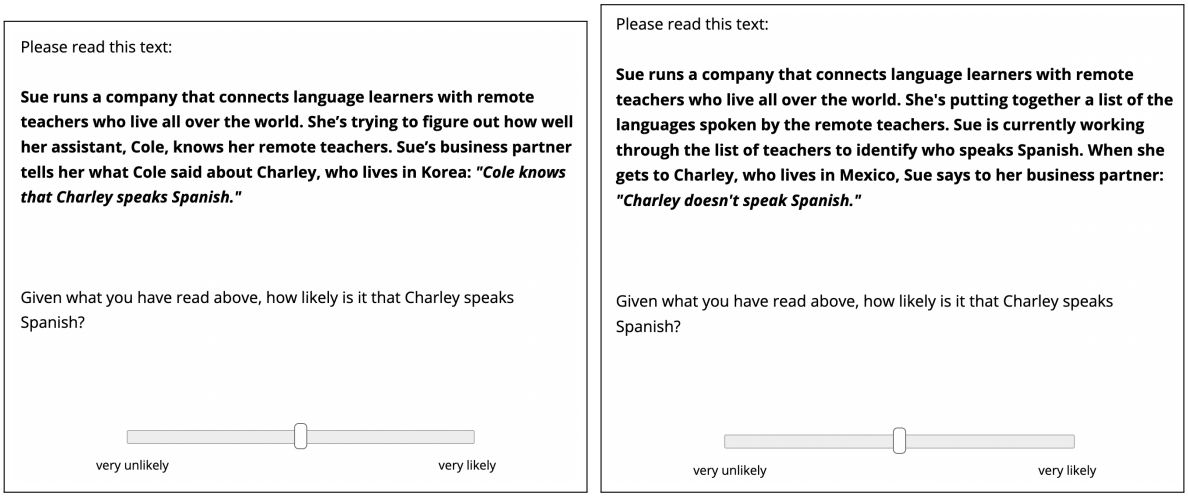
Participants. We recruited 883 participants via prolific.com (ages: 18-82; mean age: 43; 443 female, 420 male, 15 nonbinary, 5 preferred to not disclose). These participants had registered on the platform as monolingual native speakers of English who lived in the USA. They had at least 100 previous submissions and an approval rate of at least 97%.

Materials and procedure. Stimuli consisted of a short story that ended with one of the six utterance types in (3), as shown in the sample trials in Fig. 1. The stories manipulated the QUD (BEL? vs. C?) and the prior probability of the content C (higher vs. lower). The sample trial in Fig. 1a privileges the QUD BEL? (here: Does Cole believe that Charley speaks Spanish?) and the content that Charley speaks Spanish has a lower prior probability, given that Charley lives in Korea. By contrast, the sample trial in Fig. 1b privileges the QUD C? (here: Does

[3]The results of Djärv and Bacovcin (2017) and Mahler et al. (2020) suggest that the Gradient Projection Principle may not hold for some clause-embedding predicates. We return to this point in Section 2.3.

[4]The experiments, data, and R code for generating the figures and analyses of the experiments reported on in this paper are available at https://github.com/judith-tonhauser/SuB29-Scontras-Tonhauser. This repository also contains the appendix with the full set of stimuli and the models presented in Sections 3 and 4.

Charley speak Spanish?) and the content has a higher prior probability, given that Charley lives in Mexico. Stories that privileged the QUD C? were combined with all six utterance types in (3), whereas stories that privileged the QUD BEL? were combined only with the four utterance types with *think* or *know* (which can address the QUD BEL?). The 20 conditions were realized with two items (which denote the contents C): the one about Charley speaking Spanish, and another one about Jackson running ten miles (lower prior probability: Jackson is obese; higher prior probability: Jackson is training for a marathon). The prior probabilities of the contents were normed in Degen and Tonhauser (2021). The full set of stimuli (also for Exp. 2) is given in an appendix, which is provided in the repository linked in footnote 4.



(a) QUD BEL?, lower prior probability, utterance 'Cole knows that Charley speaks Spanish'.

(b) QUD C?, higher prior probability, utterance 'Charley doesn't speak Spanish'.

Figure 1: Sample trials in Exp. 1, with the item 'Charley speaks Spanish'.

Each participant completed a single trial, which consisted of a random combination of an item, a QUD, a prior, and an utterance type. Participants were asked to read the short story and then rate the likelihood of the content of the item. They gave their response on a slider marked 'very unlikely' at one end (coded as 0) and 'very likely' on the other end (coded as 1); see Fig. 1. We assume that the higher the rating, the stronger the inference to the content C. After completing the trial, participants filled out a short optional demographic survey (age, gender, native languages, speaker of American English, education). To encourage truthful responses, participants were told that they would be paid no matter what answers they gave in the survey.

Data exclusion. We excluded the data from 10 participants who did not self-identify as native speakers of American English in our demographic survey. Data from 873 participants entered into the analysis reported below (ages: 18-82; mean age: 43; 436 female, 417 male, 15 nonbinary, 5 preferred to not disclose).

Results. Fig. 2 shows the full set of results aggregated by condition. Visual inspection reveals the expected qualitative pattern of responses to our six utterance types (with positive utterances receiving higher ratings than negated ones, and *know*-utterances receiving higher ratings than *think*-utterances), as well as the expected effect of prior for all of our utterances.

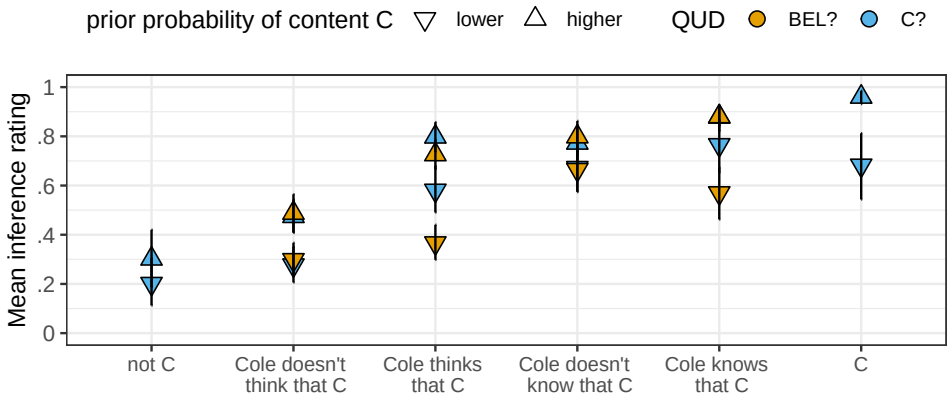
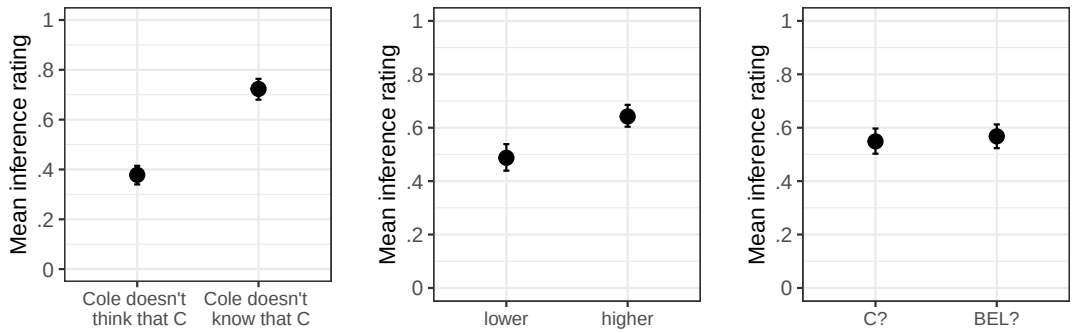


Figure 2: Mean inference ratings in Exp. 1 by utterance type, QUD, and prior probability of content C. Error bars indicate bootstrapped 95% confidence intervals.

To evaluate the three empirical hypotheses in (2), Fig. 3 plots the relevant comparisons, with the mean inference ratings by (a) utterance (negated *know* vs. negated *think*), (b) prior probability (collapsing across negated *know* and negated *think*), and (c) QUD (collapsing across negated *know* and negated *think*). In support of hypothesis (2a), we see in Fig. 3a that participants rated C as more likely when it realized the content of the complement of a negated *know*-utterance than when it realized the content of the complement of a negated *think*-utterance. In support of hypothesis (2b), we see in Fig. 3b that participants rated C as more likely when it has a higher prior probability than when it has a lower one. However, in contrast to our expectation from hypothesis (2c), we fail to see in Fig. 3c the predicted difference between the QUD conditions, whereby C would project more with QUD BEL? (C is not at-issue) than with QUD C? (C is at-issue). Thus, the results of Exp. 1 do not provide support for hypothesis (2c). We suspect that the QUD-manipulation for the complex utterance types was not sufficiently salient to participants. In Exp. 2, we follow up on this result with a more overt QUD manipulation.



(a) By utterance. (b) By prior probability. (c) By QUD.

Figure 3: Mean inference rating in Exp. 1 for C (a) by utterance, aggregated across items, QUDs, and prior probabilities, (b) by prior probability, aggregated across utterances, items and QUDs, and (c) by QUD, aggregated across utterances, items, and prior probabilities. Error bars indicate 95% bootstrapped confidence intervals.

As shown in Table 1, these observations are supported by mixed-effects linear regression mod-

els that predict participants' ratings from utterance type (negated *know* vs. negated *think*) for hypothesis (2a), from the prior probability of C (higher vs. lower) for hypothesis (2b) for negated *know* and *think* together, and from the QUD (BEL? vs. C?) for hypothesis (2c) for negated *know* and *think* together. The models included the maximal random-effects structure that allowed the models to converge, namely random by-prior and by-item intercepts for hypothesis (2a), random by-utterance and by-item intercepts for hypothesis (2b), and by-utterance, by-prior, and by-item intercepts for hypothesis (2c). Analyses were conducted using the `lme4` package (Bates et al., 2015) and *p*-values were obtained using the `lmerTest` package (Kuznetsova et al., 2017).

	β	SE	t	p
Hypothesis (2a); reference level: negated *think*	.35	.03	12.2	$<.001$
Hypothesis (2b); reference level: lower prior probability	.16	.03	5.5	$<.001$
Hypothesis (2c); reference level: C?	.009	.03	.3	.75

Table 1: Model results in Experiment 1.

2.2. Experiment 2

We hypothesize that the QUD effect did not materialize in Exp. 1 because the QUD manipulation was not sufficiently salient and any QUD effects might have been overwhelmed by the observed prior effects. Exp. 2 omitted the manipulation of the prior and manipulated the QUD in a more salient fashion. Because the QUD manipulation only concerns the four complex utterances with *think* and *know*, Exp. 2 did not include the simple utterance types (3e) and (3f).

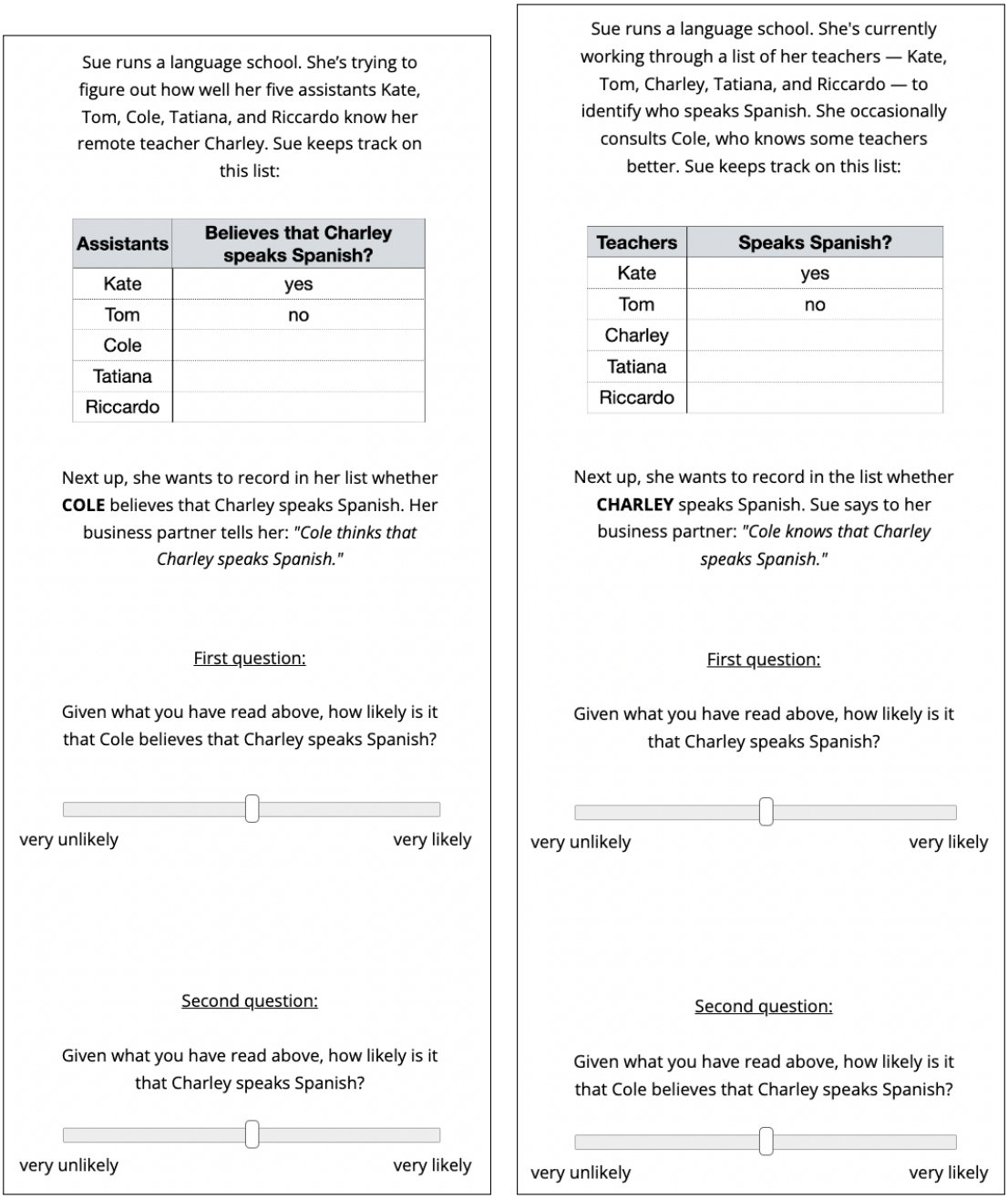
2.2.1. Methods

Participants. We recruited 405 participants via prolific.com who had not participated in Exp. 1 (ages: 20-74; mean age: 41.8; 232 female, 162 male, 5 nonbinary, 6 preferred to not disclose). These participants had registered on the platform as monolingual native speakers of English who lived in the USA. They had at least 100 previous submissions and an approval rate of at least 97%.

Materials and procedure. Stimuli consisted of a short story that ended with one of the four biclausal utterance types in (3), as shown in the sample trials in Fig. 4. The stories manipulated the QUD via the following five measures: (a) by setting up the protagonist's general goal (Fig. 4a: to figure out how well her assistants know her remote teacher Charley; Fig. 4b: to figure out who speaks Spanish), (b) by presenting a visual representation of that goal to the participants in the form of a list (cf. Song et al., 2021), (c) by highlighting, after the visual representation, the protagonist's immediate goal, which corresponds to the QUD (Fig. 4a: QUD BEL? 'Does Cole believe that Charley speaks Spanish?'; Fig. 4b: QUD C? 'Does Charley speak Spanish'), (d) by bold-facing the name of the individual that the respective QUD is about,[5] and (e) by asking the participants to first rate the likelihood of the content that addresses the respective QUD (Fig. 4a: BEL? Does Cole believe that Charley speaks Spanish?; Fig. 4b: C? Does Charley speak Spanish?). This last manipulation means that participants in the QUD

[5]This manipulation was inspired by Lu et al. (2024), where boldfacing was also used to indicate prosodic prominence as a cue to the QUD.

BEL? condition rated the likelihood of content C as the second question, whereas participants in the QUD C? condition rated it as the first question. The 8 conditions were realized with the same two items as in Exp. 1. For the full set of stimuli, see the aforementioned appendix.



(a) QUD BEL?, utterance 'Cole thinks that Charley speaks Spanish'.

(b) QUD C?, utterance 'Cole knows that Charley speaks Spanish'.

Figure 4: Sample trials in Exp. 2, with item 'Charley speaks Spanish'.

Each participant completed a single trial, which consisted of a random combination of an item, a QUD, and an utterance type. Participants were asked to read the short story and then to rate the likelihood of the belief-content and the content C (with the order dependent on the

QUD-condition, as detailed above). Participants gave their responses on a slider marked 'very unlikely' at one end (coded as 0) and 'very likely' on the other end (coded as 1); see Fig. 4. After completing the trial, participants filled out a short optional demographic survey (age, gender, native languages, speaker of American English, education). As before, participants were told that they would be paid no matter what answers they gave in the survey.

Data exclusion. We excluded the data from 8 participants who did not self-identify as native speakers of American English. We also excluded the data from an erroneously-coded condition (additional 70 participants). The data from 327 participants entered into the analysis (ages: 20-73; mean age: 42.4; 179 female, 138 male, 4 nonbinary, 6 preferred to not disclose).

2.2.2. Results

Fig. 5 plots the full results, which largely replicate the results from Exp. 1. To evaluate the two hypotheses in (2a) and (2c), Fig. 6 shows the mean inference ratings by (a) utterance (negated *think* vs. negated *know*), and by (b) QUD (collapsing across negated *think* and negated *know*). In support of hypothesis (2a), we see in Fig. 6a that participants rated C as more likely when it realized the content of the complement of a negated *know*-utterance than when it realized the content of the complement of a negated *think*-utterance. This result replicates the result of Exp. 1. Further, in support of hypothesis (2c), we see in Fig. 6b that participants rated C as more likely under QUD BEL? (C is not at-issue) than under QUD C? (C is at-issue).

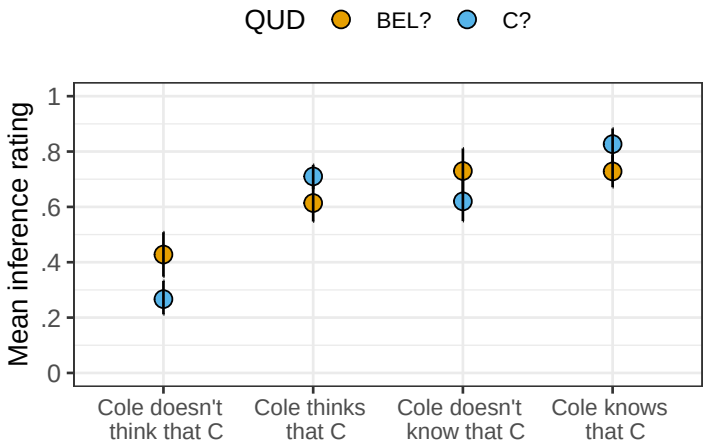
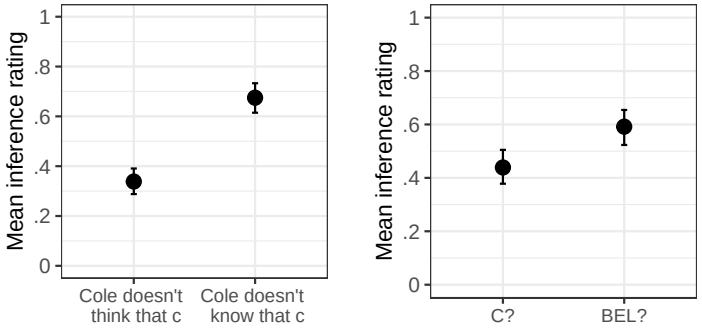


Figure 5: Mean inference ratings in Exp. 2 by utterance type and QUD. Error bars indicate bootstrapped 95% confidence intervals.

As shown in Table 2, these observations are supported by mixed-effects linear regression models that predict participants' ratings from utterance type (negated *know* vs. negated *think*) for hypothesis (2a) and from the QUD (BEL? vs. C?) for hypothesis (2c) for negated *know* and negated *think*. The models included the maximal random-effects structure that allowed the models to converge, namely random by-QUD and by-item intercepts for hypothesis (2a), and random by-utterance and by-item intercepts for hypothesis (2c).[6]

[6]For hypothesis (2c), the effect is significant when collapsing across negated *know* and negated *think*, but it does not reach significance for negated *know* on its own, although the ratings trend in the expected direction.



(a) By utterance. (b) By QUD.

Figure 6: Mean inference rating in Exp. 2 for C (a) by utterance, aggregating across items and QUDs, and (b) by QUD, aggregating across items and negated *think*- and negated *know*-utterances. Error bars indicate 95% bootstrapped confidence intervals.

	β	SE	t	p
Hypothesis (2a); reference level: negated *think*	.34	.04	8.8	$<$.001
Hypothesis (2c); reference level: C?	.14	.04	3.6	$<$.001

Table 2: Model results in Experiment 2.

2.3. Summary and discussion

The results of the two experiments presented in this section provided empirical support for the three empirical hypotheses in (2), repeated here for convenience. The model presented in the next section will be evaluated against its ability to predict these three empirical targets.

(2) Empirical hypotheses:
   a. The content of the complement of *know* is more projective than that of *think*.
   b. The contents of the complements of *know* and *think* are more projective when they have a higher prior probability than when they have a lower prior probability.
   c. The contents of the complements of *know* and *think* are more projective when they are not-at-issue than when they are at-issue.

As mentioned above, the hypothesis in (2c) is derived from Tonhauser et al.'s (2018) Gradient Projection Principle, which leads to the expectation that projection is sensitive to at-issueness, such that content that is more not-at-issue is also more projective. The results of our Exp. 2 support this hypothesis for both negated *know*- and negated *think*-utterances. These results are in contrast to those of Djärv and Bacovcin (2017), who observed the opposite effect for *believe*-utterances, as well as those of Mahler et al. (2020), who did not observe a QUD effect for nonfactive predicates as a group. The three investigations differ, however, on a number of factors, such as the predicates investigated, the entailment-canceling operators considered, whether at-issueness was manipulated (and, if yes, how) or merely measured, and the type of stimuli. Identifying the conditions under which a QUD effect on projection emerges is an important task for future research. For now it suffices to show that with negated *think* and negated *know*, there is an effect of at-issueness as hypothesized in (2c).

3. A Rational Speech Act model of projection from under negation

Our model follows the basic RSA architecture: A literal listener interprets utterances according to their semantics, a speaker reasons about the hypothetical literal listener in choosing utterances, and a pragmatic listener infers the state of the world that would have been most likely to lead the speaker to produce the observed utterance (Scontras et al., electronic; Degen, 2023).[7]

To handle inferences for the various utterance types in (3), we assume a set of four possible world states $\mathscr{W}$ (i.e., ways in which the world might be), as shown in (4). These four world states differ on whether Cole believes that Charley speaks Spanish (BEL:1 (true) / BEL:0 (false)) and whether Charley speaks Spanish (C:1 or C:0). The model assumes the six possible utterances in (3), whose literal meanings are each compatible with a subset of $\mathscr{W}$, as shown in (4).[8] As shown in (4), the literal meanings of negated utterances are the complements of the literal meanings of their positive variants. Moreover, (4a) entails the truth of the clausal complement, but there is no lexical coding of the projection inference for (4b).

(4) Utterance types and their literal meanings
$\mathscr{W}$ = {⟨BEL:1, C:1⟩, ⟨BEL:1, C:0⟩, ⟨BEL:0, C:1⟩, ⟨BEL:0, C:0⟩}

a.	*Cole knows that Charley speaks Spanish*	{⟨BEL:1, C:1⟩}
b.	*Cole doesn't know that Charley speaks Spanish*	{⟨BEL:0, C:0⟩, ⟨BEL:0, C:1⟩, ⟨BEL:1, C:0⟩}
c.	*Cole thinks that Charley speaks Spanish*	{⟨BEL:1, C:1⟩, ⟨BEL:1, C:0⟩}
d.	*Cole doesn't think that Charley speaks Spanish*	{⟨BEL:0, C:0⟩, ⟨BEL:0, C:1⟩}
e.	*Charley speaks Spanish*	{⟨BEL:1, C:1⟩, ⟨BEL:0, C:1⟩}
f.	*Charley doesn't speak Spanish*	{⟨BEL:1, C:0⟩, ⟨BEL:0, C:0⟩}

QUDs are modeled as partitions over $\mathscr{W}$, following Kao et al., 2014. The QUD BEL? (whether Cole believes that Charley speaks Spanish) partitions $\mathscr{W}$ into the set of two world states in which Cole believes (BEL:1), and the two in which he doesn't (BEL:0). The QUD C? (whether Charley speaks Spanish) partitions $\mathscr{W}$ into the states where Charley speaks Spanish (C:1) and those where he doesn't (C:0). Thus, the content that Charley speaks Spanish is not-at-issue with respect to the QUD BEL?, and at-issue with respect to the QUD C?.[9]

Finally, we follow Qing et al. (2016) in assuming that the literal listener interprets utterances relative to the speaker's private assumptions *A*, modeled as a non-empty subset of $\mathscr{W}$.[10] For instance, a speaker who has no assumptions operates with respect to A = $\mathscr{W}$ (i.e., no world states are assumed to be ruled out), whereas a speaker who assumes that Charley speaks Spanish operates with respect to A = {⟨BEL:1, C:1⟩, ⟨BEL:0, C:1⟩} (i.e., the speaker assumes that world states in which Charley does not speak Spanish are ruled out).

[7] A video that introduces the model and how it derives projection inferences can be found at https://youtu.be/8xjCuqEvVTA.

[8] Qing et al. (2016) assume additional utterance alternatives that exhaustively describe the possible world states uniquely. Including such alternatives in our model does not improve the model predictions for *know* and *think*; see model `1a-original-more-alternatives` in the repo.

[9] In contrast to Qing et al. (2016), we do not assume a maximal QUD. Including such a QUD does not improve the model predictions for *know* and *think*; see model `1b-original-with-maxQUD` in the repo.

[10] Qing et al. (2016) call these subsets the "common ground," but we think "private assumptions" better captures this component of the model because the pragmatic listener does not know which *A* is assumed.

The literal listener L_0, defined in (5), observes an utterance u, the speaker's assumption about possible world states A, and the QUD Q assumed by the speaker, and returns a distribution over answers to the QUD.

(5) Literal listener: $P_{L_0}(Q(w) \mid u, A, Q) \propto \sum_{w' \in A \cap [\![u]\!]} \delta_{Q(w)=Q(w')} \cdot P(w')$

The speaker S_1, defined in (6), observes the true state of the world w and wants to convey the correct answer to the QUD Q, taking into consideration their assumptions A. The speaker evaluates the six possible utterances with respect to how likely the literal listener is to infer the correct answer given a specific utterance u, their assumptions A, and the QUD Q.

(6) Speaker: $P_{S_1}(u \mid w, A, Q) \propto exp(\alpha(log(P_{L_0}(Q(w) \mid u, A, Q) - C(u))))$

The pragmatic listener L_1, defined in (7), observes an utterance u and QUD Q and updates their prior beliefs about the world state w and the speaker assumptions A by reasoning about the world state w and the private assumption A that the speaker used in choosing their utterance.[11]

(7) Pragmatic listener: $P_{L_1}(w, A \mid u, Q) \propto P_{S_1}(u \mid w, A, Q) \cdot P(w) \cdot P(A)$

3.1. Evaluating the predictions of the RSA model

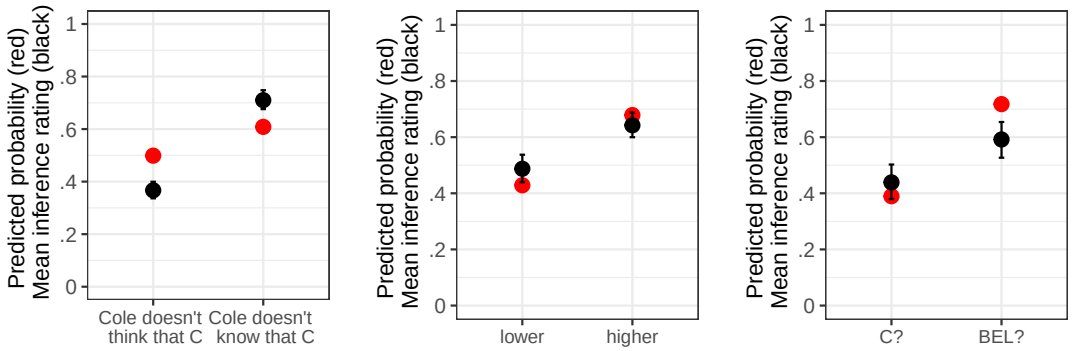
We evaluate the model predictions for the negated *know*- and *think*-utterances assuming similar parameter settings as in Qing et al. (2016).[12] We assume that the complex utterances in (3) are twice as costly as the simple ones. For the prior probability of content, we assume that content is twice as likely to be true than to be false in a higher prior probability context, and twice as likely to be false than to be true in a lower prior probability context.

The three panels of Fig. 7 plot the predictions of the model (in red) for each of the three empirical targets in (2) against the human data from Exps. 1 and 2 (in black). As shown in panel (a), the model predicts that the inference to content C is stronger when the clausal complement that contributes C is embedded under negated *know* than when it is under negated *think*. As shown in panel (b), the model predicts that the inference to C is stronger when C has a higher prior probability than when it has a lower one. Finally, as shown in panel (c), the model predicts that the inference to C is stronger under QUD BEL? (C is not-at-issue) than under QUD C? (C is at-issue). As shown in each panel, the model predictions are matched qualitatively by the human data, which suggests that the model is making empirically adequate predictions.

We now discuss how the model derives the three empirical targets in (2), starting with the empirical target in (2a), according to which an interpreter (modeled as a pragmatic listener L_1) is more likely to infer C for negated *know*-utterances than for negated *think*-utterances. What

[11] Projection inferences were defined in Section 1 as inferences about the truth of non-asserted content under an entailment-canceling operator. This definition is motivated by the fact that we evaluate the predictions of our model based on the marginal posterior probability of w, specifically, those w in which C is true. A different definition of projection inferences as inferences about speaker belief in the truth of non-asserted content under an entailment-canceling operator (see, e.g., Degen and Tonhauser, 2022; Pan and Degen, 2023) is also compatible with our model, if we evaluated its predictions based on the marginal posterior probability of A, specifically those A that entail C.

[12] We assume the same prior over the speaker's private assumptions as in Qing et al. (2016) and $\alpha = 10$. The predictions shown here are stable across a number of other parameter settings. For the full model specification, see model `1-original-model` in the repo.



(a) By utterance. (b) By prior probability. (c) By QUD.

Figure 7: Predicted probability of C (in red) against mean inference rating for C from Exps. 1 and 2 (in black) (a) by utterance (data pooled from Exps. 1 and 2), aggregated across QUDs (Exps. 1 and 2), prior probabilities (Exp. 1 only) and items, (b) by prior probability, aggregated across QUDs, items and utterances (data from Exp. 1), and (c) by QUD, aggregated across utterances, items and prior probabilities (data from Exp. 2). Error bars on mean inference ratings indicate 95% bootstrapped confidence intervals.

is crucial here is that the negated *think*-utterance is semantically stronger than the negated *know*-utterance, as the former excludes more states. Thus, if the speaker makes no private assumptions (e.g., A = $\mathcal{W}$), the negated *think*-utterance is more informative than the negated *know*-utterance and, hence, more useful; more useful utterances are more likely to be selected by the speaker. On the other hand, the negated *know*-utterance is just as informative as the negated *think*-utterance if the speaker assumes C (i.e., A entails C) and the QUD is BEL?, as now the negated *know*-utterance fully answers the QUD. So, an interpreter is more likely to infer that the speaker assumes C upon observing that the speaker produced the negated *know*-utterance; with this assumption, uttering the negated *know*-utterance becomes much more likely. In other words, given that the pragmatic listener infers the conditions that led the speaker to produce the utterance that they heard, upon hearing negated *know* the listener infers that the speaker assumed C when making the utterance.

The next empirical target is (2b), according to which interpreters are more likely to infer C when it has a higher prior probability than when it has lower prior probability. Generalizing across possible speaker private assumptions A, neither the negated *know*-utterance nor the negated *think*-utterance resolve whether C is true. Without information from the utterance, interpreters are left to rely on their prior beliefs to infer the world state. When prior beliefs favor C, posterior beliefs will, too.

The final empirical target is (2c), according to which interpreters are more likely to infer C when the QUD is BEL? (C is not-at-issue) than when it is C? (C is at-issue). As noted above for (2a), if the QUD is BEL?, it is reasonable to infer from both the negated *know*- and *think*-utterances that the speaker assumes that C is true, as both utterances are informative in this constellation (that is, they resolve the QUD to BEL:0). On the other hand, if the QUD is C?, interpreters are unlikely to infer that the speaker assumes that C is true from either of the two utterances, as neither would resolve the QUD.

The quantitative predictions of the model are also quite good already, given that we are assuming parameters of Qing et al. (2016), who did not evaluate the predictions of their model against human data. There are, however, some apparent mismatches in the quantitative predictions. First, as shown in Fig. 7a, the predicted probability of C is too high for negated *think*-utterances, compared to the human data pooled from Exps. 1 and 2. This result suggests that the model does not derive the neg-raising inference for such utterances (see also Pan and Degen, 2023). Second, as shown in Fig. 7b, the effect of the prior beliefs is stronger in the model than in the data collected in Exp. 1. Further, as shown in Fig. 7c, the effect of the QUD is stronger in the model than in the data collected in Exp. 2. Both of these latter mismatches between the model predictions and the human data may be due to model error or because the manipulation in the experiments was not sufficiently clear. Teasing these two possibilities apart is an important goal for future research.

3.2. Discussion

This section has presented a formal analysis for the projection of the content of the clausal complement of negated *know*-utterances. The model predicts that interpreters are more likely to derive a projection inference from negated *know*-utterances than negated *think*-utterances, when the prior probability of the content is higher than when it is lower, and when the content is not-at-issue with respect to the QUD than when it is at-issue. These model predictions are matched qualitatively by the human data collected in Exps. 1 and 2.

The analysis shares with the approaches in Simons (2001), Abusch (2002, 2010), Abrusán (2011, 2016), Romoli (2015), Simons et al. (2017), and Roberts and Simons (2024) the assumption that the projection of content is not derived from the lexical specification of the content as a presupposition, but from the lexical meanings of expressions as well as reasoning. Our approach contrasts with analyses like Heim (1983), van der Sandt (1992), and Djärv and Bacovcin (2020), where projection inferences arise from the lexical specification of content as a constraint on the Common Ground. Our analysis also captures the spirit of some of these prior analyses of projection inferences. First, projection inferences are derived in our analysis from the lexical entailments of *know* and *think*, in positive and negated utterances, just as in Abrusán (2011, 2016), and Simons et al. (2017). Second, our analysis assumes that projection is sensitive to the QUD addressed by the utterance, as do Abrusán (2011, 2016), and Simons et al. (2017). Our analysis also assumes that projection is sensitive to interpreters' prior beliefs about content, an assumption that resembles Schlenker's (2021) assumption that not just lexical entailments may project but also content that is contextually entailed. Finally, our analysis assumes that utterance alternatives are central to deriving projection inferences, an assumption also found in works like Abusch (2002, 2010), and Romoli (2015), although the role of alternatives in these analyses is different.

Despite shared assumptions between our analysis and prior ones, none of the prior analyses are able to predict the three empirical targets in (2), as summarized in Table 3. While all analyses predict that content is more projective from under negated *know* than from under negated *think* (hypothesis (2a)), none of the contemporary projection analyses except Schlenker's (2021) predict that projection is sensitive to the prior probability of the content (hypothesis (2b)). With respect to hypothesis (2c), only the analyses in Djärv and Bacovcin (2020), Abrusán (2011, 2016), Simons et al. (2017), and Beaver et al. (2017) predict that projection is sensitive

Empirical hypothesis: Content C projects more...	Heim, 1983; van der Sandt, 1992	Djärv and Bacovcin, 2020	Abrusán, 2011, 2016	Simons et al., 2017; Beaver et al., 2017	Romoli, 2015	Schlenker, 2021	Analysis presented here
(2a): ...from under negated *know* than negated *think*.	✓	✓	✓	✓	✓	✓	✓
(2b): ...with higher than with lower prior probabilities.	×	×	×	×	×	✓	✓
(2c): ...with QUD BEL? than with QUD C?.	×	✓	✓	✓	×	×	✓

Table 3: Predictions of formal projection analyses with respect to the three empirical hypotheses in (2): ✓ indicates that the analysis captures the empirical target, × that it does not.

to the QUD, in addition to our analysis.

Of course, a pressing task for future research is to extend the analysis presented in this section to other entailment-canceling operators besides sentential negation.

4. Predictions for non-projection inferences

We have seen that the analysis presented in Section 3 makes empirically-adequate predictions about projection inferences, that is, inferences to non-asserted content in the scope of an entailment-canceling operator. The analysis also makes predictions about other types of inferences—specifically, it makes predictions about the strength of the inference to the asserted content of the positive and negated declarative sentences in (3e) and (3f), respectively, and it makes predictions about the strength of the inference to the non-asserted content embedded under positive *know* and *think* in (3a) and (3c), respectively. That our analysis of projection inferences is couched in a general analysis of inferences stands as an additional advantage.

Fig. 8 plots (in red) the predictions of the model presented in Section 3 for the full range of utterances. For the current purposes of model refinement, we limit our focus to what we consider the default QUDs for the individual utterances (for simple utterances the default QUD is C?; for complex utterances with *think* and *know* it is BEL?) and to the 'higher' prior condition. While the model captures some of the response patterns in the human data (plotted in black; e.g., lowest inference ratings for simple negated *not C* and highest ratings for simple positive *C* and positive *know*; generally higher ratings for *know* compared to *think*), there are several points where the quantitative predictions deviate substantially. To improve the quantitative predictions of our model, in this section we more systematically explore the parameter space. We also introduce a change to the model that allows for less categorical behavior.

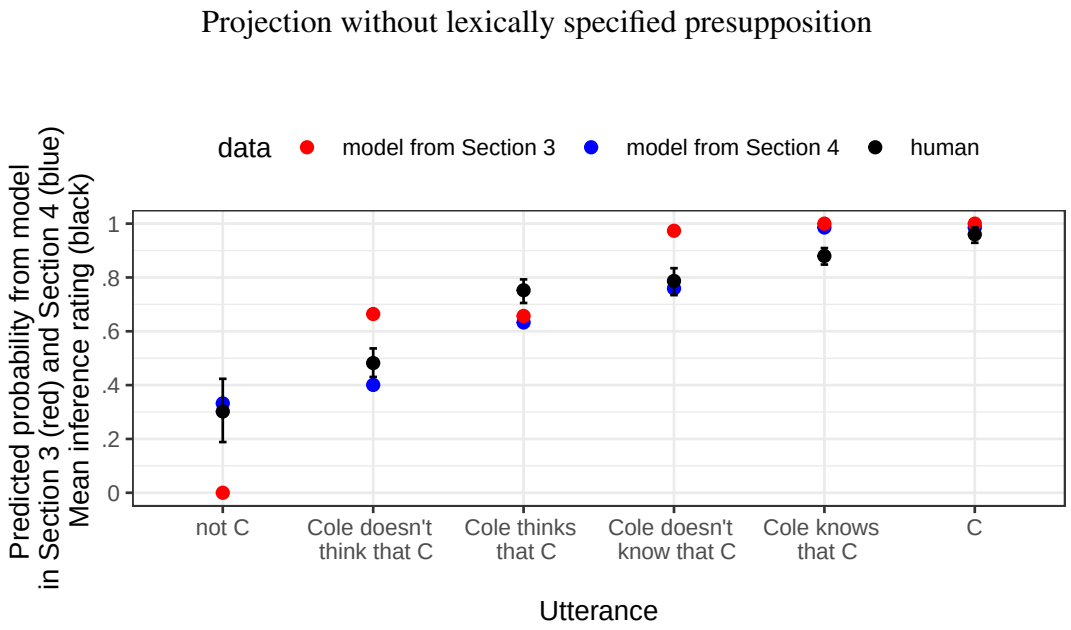


Figure 8: Predicted probabilities for content C (for the model in Section 3 in red and for the model in Section 4 in blue) when C has higher prior probability, under QUD BEL? for complex utterances and QUD C? for simple utterances, compared to mean inference ratings for C (in black) from Exp. 1 when C has higher prior probability, under QUD BEL? for complex utterances and QUD C? for simple utterances, aggregated across items. Error bars indicate 95% bootstrapped confidence intervals.

Perhaps the most obvious shortcoming of the model predictions in Fig. 8 comes for the negated declarative *not C*, where the model follows the literal semantics and predicts zero inference to the truth of C, yet humans seemingly dismiss or downweight the semantics and provide much higher ratings. We find similar behavior, although less extreme, for positive *know* and the simple positive *C*, where the literal semantics leads the model to predict certainty in the truth of C, yet participants are less sure. To improve the quantitative fit of our model we therefore need to introduce more flexibility than is allowed for by the literal semantics.

We find inspiration from the so-called "wonky worlds" model of Degen et al. (2015), which provides listeners with the means of disregarding an unlikely interpretation, choosing to back off to their prior beliefs instead. We suspect that the human response patterns in Fig. 8 arise from a similar process: When a literal interpretation is too unlikely, for example hearing that *Charlie doesn't speak Spanish* despite living in Mexico, a pragmatic listener disregards the utterance and instead uses their prior beliefs to complete the task. Importantly, this process is not categorical: Interpreters downweight the literal interpretation to the extent that it deviates from what they take to be likely.

We implement this reasoning process in our model at the level of the pragmatic listener: The listener first interprets the utterance in the manner described in Section 3. To evaluate plausibility, the L_1 listener then compares the marginal posterior distribution over C (e.g., the extent to which they believe that Charlie speaks Spanish after interpretating the utterance) with their marginal prior over C (what they believed about C before hearing the utterance). We implement this comparison as the Kullback–Leibler (KL) divergence of the marginal posterior from the marginal prior. We convert this divergence value into a probability, $e^{-divergence}$, such that a

divergence value of 0 (i.e., the posterior perfectly matches the prior) yields probability 1. The pragmatic listener returns a sample from their posterior in proportion to this probability.

The change described above allows the interpreter to systematically deviate from the literal semantics in a way that loosens up interpretations in line with the behavior of our participants. To further improve the fit of our model, we introduce the following changes. For our world state prior, C is four times as likely to be true than to be false. Instead of assuming that the speaker's private assumptions A may be any non-empty subset of $\mathscr{W}$, we assume that A may only be those nine non-empty subsets of $\mathscr{W}$ in (8) that are obtained by observing nothing, that BEL is true/false, that C is true/false, or observing the truth or falsity of both BEL and C. We further assume that having knowledge of C is four times as likely has having knowledge of BEL, and that having full knowledge of both BEL and C is half as likely as having only knowledge of BEL. We also assume that having no beliefs about either BEL or C is half as likely as having full knowledge. Finally, we set the speaker's optimality parameter α to 4.

(8) a. $\{\langle$BEL:1, C:1$\rangle$, $\langle$BEL:1, C:0$\rangle$, $\langle$BEL:0, C:1$\rangle$, $\langle$BEL:0, C:0$\rangle$ $\}$
b. $\{\langle$BEL:1, C:1$\rangle$, $\langle$BEL:1, C:0$\rangle\}$
c. $\{\langle$BEL:0, C:1$\rangle$, $\langle$BEL:0, C:0$\rangle\}$
d. $\{\langle$BEL:1, C:1$\rangle$, $\langle$BEL:0, C:1$\rangle\}$
e. $\{\langle$BEL:1, C:0$\rangle$, $\langle$BEL:0, C:0$\rangle\}$
f.-i. $\{\langle$BEL:1, C:1$\rangle\}$, $\{\langle$BEL:1, C:0$\rangle\}$, $\{\langle$BEL:0, C:1$\rangle\}$, $\{\langle$BEL:0, C:0$\rangle\}$

Fig. 8 compares the predictions from this new version of our model (in blue) with the human responses from Exp. 1 (in black). While there are still opportunities for improvement (e.g., the predictions for negated *think* and positive *know*), the amendments to our model in Section 3 indeed yield a better fit to the data. For the simple utterances, this result suggests that participants incorporate their prior beliefs about the truth of the content in determining their posterior beliefs, rather than just relying on the fact that the speaker asserted the content or its negation.

5. Conclusions

This paper has presented an analysis of the projection of the content of the clausal complement of negated declaratives with *know* that goes beyond contemporary projection analyses (e.g., Heim, 1983; van der Sandt, 1992; Djärv and Bacovcin, 2020; Abrusán, 2011, 2016; Simons et al., 2017; Schlenker, 2021) in that it predicts both sensitivity to the QUD and to interpreters' prior beliefs. The analysis also goes beyond most contemporary analyses in making predictions about the projection of the content of the clausal complement of negated declaratives with *think* (though see, e.g., Simons et al., 2017). By relying on general reasoning about utterance alternatives and lexical entailments (rather than lexically-coded constraints on the Common Ground), our analysis has the potential to extend more broadly to predicates with non-entailed clausal complements. A further benefit of our analysis is that it is couched in a general analysis of inferencing, and also makes empirically adequate predictions about other types of inferences.

Methodologically, the experiments presented here suggest that the QUD can be successfully manipulated if the story goal and the relevant alternatives are made very explicit. Future research will need to investigate whether spoken stimuli further enhance the QUD manipulation.

References

Abrusán, M. (2011). Predicting the presuppositions of soft triggers. *Linguistics & Philosophy 34*, 491–535.

Abrusán, M. (2016). Presupposition cancellation: Explaining the 'soft-hard' trigger distinction. *Natural Language Semantics 24*, 165–202.

Abusch, D. (2002). Lexical alternatives as a source of pragmatic presupposition. *Semantics and Linguistic Theory 12*, 1–19.

Abusch, D. (2010). Presupposition triggering from alternatives. *Journal of Semantics 27*, 37–80.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.

Beaver, D., C. Roberts, M. Simons, and J. Tonhauser (2017). Questions Under Discussion: Where information structure meets projective content. *Annual Review of Linguistics 3*, 265–284.

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

Degen, J., M. Tessier, and N. Goodman (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, pp. 548–553. Austin, TX: Cognitive Science Society.

Degen, J. and J. Tonhauser (2021). Prior beliefs modulate projection. *Open Mind 5*, 59–70.

Degen, J. and J. Tonhauser (2022). Are there factive predicates? An empirical investigation. *Language 98*, 552–591.

Delin, J. (1992). Properties of *it*-cleft presuppositions. *Journal of Semantics 9*, 289–306.

Djärv, K. and H. Bacovcin (2017). Prosodic effects on factive presupposition projection. *Semantics and Linguistic Theory 27*, 116–133.

Djärv, K. and H. Bacovcin (2020). Prosodic effects on factive presupposition projection. *Journal of Pragmatics 169*, 61–85.

Frank, M. C. and N. D. Goodman (2012, 5). Predicting pragmatic reasoning in language games. *Science 336*(6084), 998.

Heim, I. (1983). On the projection problem for presuppositions. *WCCFL 2*, 114–125.

Kao, J. T., J. Wu, L. Bergen, and N. Goodman (2014). Nonliteral understanding of number words. *PNAS 111*, 12002–7.

Kiparsky, P. and C. Kiparsky (1970). Fact. In M. Bierwisch and K. Heidolph (Eds.), *Progress in Linguistics*, pp. 143–173. The Hague: Mouton.

Kuznetsova, A., P. B. Brockhoff, and R. H. B. Christensen (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software 82*, 1–26.

Lu, J., D. Pan, and J. Degen (2024). Evidence for a discourse account of manner-of-speaking islands. Manuscript under review.

Mahler, T. (2020). The social component of projection behavior of clausal complements. *Linguistic Society of America 5*, 777–791.

Mahler, T. (2022). *Social identity information in projection inferences: A case study in social and semantic-pragmatic meaning*. Ph. D. thesis, The Ohio State University.

Mahler, T., M.-C. de Marneffe, and C. Lai (2020). The prosody of presupposition projection in naturally-occurring utterances. *Sinn und Bedeutung 24*, 20–37.

Pan, D. and J. Degen (2023). Towards a computational account of projection inferences in polar

interrogatives with clause-embedding predicates. *Annual Meeting of the Cognitive Science Society 45*, 1079–1085.

Qing, C., N. D. Goodman, and D. Lassiter (2016). A rational speech-act model of projective content. *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, 1110–1115.

Roberts, C. and M. Simons (2024). Preconditions and projection: Explaining non-anaphoric presupposition. *Linguistics & Philosophy 47*, 703–748.

Romoli, J. (2015). The presuppositions of soft triggers are obligatory scalar implicatures. *Journal of Semantics 32*, 173–291.

Schlenker, P. (2021). Triggering presuppositions. *Glossa 6*, 1–28.

Scontras, G., M. H. Tessler, and M. Franke (electronic). Probabilistic language understanding: An introduction to the Rational Speech Act framework. Retrieved from https://www.problang.org.

Simons, M. (2001). On the conversational basis of some presuppositions. *Semantics and Linguistics Theory 11*, 431–448.

Simons, M., D. Beaver, C. Roberts, and J. Tonhauser (2017). The best question: Explaining the projection behavior of factives. *Discourse Processes 54*(3), 187–206.

Song, Y., A. Hernandez Jimenez, and G. Scontras (2021). Cross-linguistic scope ambiguity: An investigation of English, Spanish, and Mandarin. *Proceedings of the Linguistic Society of America 6*(1), 572–586.

Spenader, J. (2002). *Presuppositions in Spoken Discourse*. Ph. D. thesis, Stockholm University.

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

van der Sandt, R. (1992). Presupposition projection as anaphora resolution. *Journal of Semantics 9*, 333–377.

Warstadt, A. (2022). Presupposition triggering reflects pragmatic reasoning about utterance utility. *Amsterdam Colloquium 23*, 444–451.