Modeling the prompt in inference judgment tasks

Julian Grove & Aaron Steven White[*]

Abstract. We show that when analyzing data from inference judgment tasks, it can be important to incorporate into one's data analysis regime an explicit representation of the semantics of the natural language prompt used to guide participants on the task. To demonstrate this, we conduct two experiments within an existing experimental paradigm focused on measuring factive inferences, while manipulating the prompt participants receive in small but semantically potent ways. In statistical model comparison couched within the framework of probabilistic dynamic semantics, we find that probabilistic models structured, in part, by the semantics of the prompt fit better to data collected using that prompt than models that ignore the semantics of the prompt.

Keywords. presupposition, factivity, dynamic semantics, probabilistic models

1. Introduction. When collecting inference judgments in formal experiments, it is common for trials to consist of the following pieces: (i) a target linguistic expression whose inferential affordances one is interested in measuring; (ii) a description of some context which the expression might be used in; (iii) a natural language prompt guiding participants on the relevant task; and (iv) a response instrument, such as an ordinal or slider scale. When analyzing the data from such experiments, one generally incorporates some representation of components (i), (ii), and (iv); but it is relatively rare to incorporate a representation of component (iii)—likely because this component is typically constant across all experimental items.[1]

We show that it can be important to incorporate an explicit representation of the semantics of natural language prompts used in inference judgment experiments into one's data analysis regime. To demonstrate this, we fix components (i), (ii), and (iv) of an experimental paradigm focused on measuring factive inferences—such as the inference from (1a) to (1b)—and manipulate component (iii)—the natural language prompt—in small but semantically potent ways.

(1) a. Jo {loves, doesn't love} that Mo left. b. Mo left.

We collect data using two such manipulated prompts and show, through statistical model comparison, that probabilistic models structured, in part, by the semantics of the prompt fit better to data collected using that prompt than models that ignore its semantics.

In section 2, we describe how we incorporate natural language prompts into statistical models using the probabilistic dynamic semantics (PDS) framework (Grove & White 2024a,b). We then review, in section 3, the experimental paradigm and prior models of data collected under that paradigm within PDS. In section 4, we describe our two modifications to this paradigm and the

[*] We thank the organizers of ELM 3, as well as four anonymous reviewers. Authors: Julian Grove, University of Rochester (julian.grove@gmail.com) & Aaron Steven White, University of Rochester (aaron.white@rochester.edu).

[1] One might conceive of the way that hypotheses are modeled in the natural language inference (NLI) literature (Cooper et al. 1996, Dagan, Glickman & Magnini 2006, MacCartney 2009 *et seq*) as an exception; but most NLI models assume a single ground truth label aggregated from possibly multiple participants responses (cf. Gantt, Kane & White 2020), rather than modeling the distribution of reponses themselves, as we do here.

two corresponding experiments. In section 5, we describe our model implementations and report model comparisons before concluding in section 6.

2. Probabilistic dynamic semantics. We couch our modeling within the recently developed probabilistic dynamic semantics (PDS) framework (Grove & White 2024a,b). The core idea we implement within this framework is that an inference judgment task can be characterized as a kind of discourse: a target sentence and its context of interpretation can be analyzed as updates to a common ground; and the prompt can be regarded as contributing a new question under discussion (QUD; Ginzburg 1996, Roberts 2012). The upshot of implementing this idea within PDS is that, once we fix these formal components, we need only make linking assumptions—i.e. assumptions about how the probability distribution over possible answers to the QUD (given a common ground updated with the assertion) should manifest as a particular distribution of responses, given a particular response instrument. Our approach is therefore analogous to standard analysis regimes, in that one is always making linking assumptions via one's choice of statistical model (see Jasbi, Waldon & Degen 2019), but it has the additional benefit of providing an explicit interface between the formal semantics of an experimental item and these linking assumptions. We touch on some crucial formal details below. For a fuller treatment, see Grove & White 2024b.

2.1. Distributions on discourse constructs. In PDS, an ongoing discourse is represented by a map from an input state to a probability distribution over output states. States are simply tuples of *metalinguistic parameters*—including, e.g., the common ground (Stalnaker 1978) and the QUD, along with other conversationally relevant features of discourse, such as the meanings of individual lexical items. One can view the state as akin to the context state of Farkas & Bruce (2010), though our state is, in principle, less constrained, in the sense that it is compatible with alternative representations of the context.[2] Notating the type of this state tuple 's_μ', an ongoing discourse is of type $s_\mu \rightarrow \mathtt{P}s_\mu$, where for any type α, $\mathtt{P}\alpha$ is the type of *probability distributions* over values of type α. $\mathtt{P}s_\mu$ is thus the type of probability distributions over output states.

To perform updates of various types to some ongoing discourse, we use two operators—*bind* and *return*—which allow probability distributions to be sampled (bind) and ordinary, non-probabilistic values to be returned as *degenerate* distributions over those values (return).[3] To illustrate, suppose we have some categorical distribution $\mathsf{mammal} : \mathtt{P}e$ on mammals. We can represent a distribution on mammals' mothers as:

$$\begin{array}{c} x \sim \mathsf{mammal} \\ \boxed{\mathsf{mother}(x)} \end{array}$$

where $\mathsf{mother} : e \rightarrow e$ maps an entity to its mother. Here, x is sampled from mammal using bind, and then $\mathsf{mother}(x)$ is returned, as indicated by the orange box.

We represent the common ground as a probability distribution over tuples of possible worlds and linguistic parameters (the latter of which we refer to as 'contexts')—something of type $\mathtt{P}s_{w,\kappa}$.

[2] As Grove & Bernardy (2023) show, this framework can also be used to represent the Rational Speechs Acts (RSA) framework, as laid out in Frank & Goodman 2012, Goodman & Frank 2016. The frameworks are distinct, however, in the sense that PDS has strong commitments to computational purity that often do not hold in practical applications of RSA (see discussion in Grove & Bernardy 2023, Grove & White 2024b).

[3] We do not focus on this fact here, but these operators importantly conform to the monad laws. See Grove & Bernardy 2023, Grove & White 2024b for why this fact is useful.

The idea is that the first part $w : s_w$ of this tuple represents facts about the relevant world—e.g., how tall a particular individual is—while the second part $\kappa : s_\kappa$ represents the values of linguistic parameters—e.g., the vague threshold of height past which one's height is considered tall.

2.2. MAKING AN ASSERTION. Similar to discourses, the meanings of expressions are dynamic. We therefore represent them as functions of type $s_\mu \to \texttt{P}(\alpha \times s_\mu)$; that is, given an input state, the meaning of an expression produces a probability distribution over pairs of ordinary meanings (of type α) and possible output states. Given a sentence whose meaning ϕ is of type $s_\mu \to \texttt{P}((s_{w,\kappa} \to t) \times s_\mu)$, we represent an *assertion* of that sentence as an update to the common ground of an ongoing discourse. Given a discourse $m\ : s_\mu \to \texttt{P}s_\mu$, we may update its common ground with ϕ as follows:

$$\texttt{update}_{\mathsf{cg}} : (s_\mu \to \texttt{P}((s_{w,\kappa} \to t) \times s_\mu)) \to (s_\mu \to \texttt{P}s_\mu) \to s_\mu \to \texttt{P}s_\mu$$

$$\begin{aligned}\texttt{update}_{\mathsf{cg}}(\phi)(m) \overset{\text{def}}{=} \lambda\mu.\ & \mu' \sim m(\mu)\\ & \langle\psi,\mu''\rangle \sim \phi(\mu')\\ & cg \sim \boxed{\begin{array}{l}\langle w,\kappa\rangle \sim \mathsf{cg}(\mu'')\\ \texttt{observe}(\psi(w,\kappa))\\ \boxed{\langle w,\kappa\rangle}\end{array}}\\ & \boxed{\texttt{set}_{\mathsf{cg}}(cg,\mu'')}\end{aligned}$$

Intuitively, $\texttt{observe} : t \to \texttt{P}\diamond$ (where $\diamond$ is the unit type) takes a truth value and either keeps or throws out the distribution represented by the expression which follows it, depending on whether this truth value is T or F.[4] For instance, we may constrain our earlier mother program to describe a distribution over only dogs' mothers as:

$$\begin{aligned}\texttt{dog-mother} \overset{\text{def}}{=}\ & x \sim \mathsf{mammal}\\ & \texttt{observe}(\mathsf{dog}(x))\\ & \boxed{\mathsf{mother}(x)}\end{aligned}$$

In $\texttt{update}_{\mathsf{cg}}$, $\mathsf{cg}(\mu'')$ is that projection of the state μ'' identifying the common ground; meanwhile, $\texttt{set}_{\mathsf{cg}} : \texttt{P}s_{w,\kappa} \times s_\mu \to s_\mu$ takes a common ground together with a state and gives back a new state just like the old, but with the new common ground instead of the old one—i.e., $\texttt{set}(cg, \langle ..., cg', ...\rangle) = \langle ..., cg, ...\rangle$. $\texttt{update}_{\mathsf{cg}}(\phi)$ thus modifies an ongoing discourse so that its probability distribution over output states involves common grounds in which the proposition returned by ϕ has been observed.

2.3. ASKING A QUESTION. We follow a categorial grammar tradition by analyzing questions as denoting (given a world-context pair) sets of (true) short answer meanings (Hausser & Zaefferer

[4] $\texttt{observe}$ is defined in terms of a more primitive operation, $\texttt{factor}$,

$$\texttt{observe} : t \to \texttt{P}\diamond$$
$$\texttt{observe}(\phi) \overset{\text{def}}{=} \texttt{factor}(\mathbb{1}(\phi))$$

where $\mathbb{1} : t \to r$ is an *indicator* function, i.e., taking T onto 1 and F onto 0. See Grove & Bernardy 2023 for details.

1978, Hausser 1983; cp. Karttunen 1977, Groenendijk & Stokhof 1984). For simplicity (but not by necessity), we assume that all questions are degree questions—i.e., that at a given world-context pair, they denote sets of degrees of type r (real numbers). Questions—taking their dynamic potential into account—are thus of type $s_\mu \to \texttt{P}((s_{w,\kappa} \to r \to t) \times s_\mu)$.

Asking a question is a matter of pushing a question meaning onto the QUD stack (Roberts 2012, Farkas & Bruce 2010). We therefore define $\texttt{update}_{\textsf{qud}}$ as follows:

$$\begin{aligned}
\texttt{update}_{\textsf{qud}} &: (s_\mu \to \texttt{P}((s_{w,\kappa} \to r \to t) \times s_\mu)) \to (s_\mu \to \texttt{P}s_\mu) \to s_\mu \to \texttt{P}s_\mu \\
\texttt{update}_{\textsf{qud}}(q)(m) &\overset{\text{def}}{=} \lambda\mu.\ \mu' \sim m(\mu) \\
&\qquad\quad \langle q', \mu'' \rangle \sim q(\mu') \\
&\qquad\quad \boxed{\texttt{set}_{\textsf{qud}}(q' {::} \textsf{qud}(\mu''), \mu'')}
\end{aligned}$$

Here, $\textsf{qud}(\mu'')$ is the QUD stack associated with the state μ''.

2.4. Responding to a question. Given an ongoing discourse, one can respond to the QUD at the top of the QUD stack based on their prior knowledge. Thus we can assume that a given responder knows some prior $k : \texttt{P}s_\mu$ over *starting* states, which they use—in conjuction with the interim updates to the discourse—to derive a probability distribution over answers to the QUD. Given this prior k and an ongoing discourse m, the answer distribution—of type $\texttt{P}r$—can be defined as:

$$\begin{aligned}
&\mu \sim k \\
&\mu' \sim m(\mu) \\
&\langle w, \kappa \rangle \sim \textsf{cg}(\mu') \\
&\boxed{\textsf{max}(\textsf{qud}(\mu')_0(w, \kappa))}
\end{aligned}$$

Here, $\textsf{qud}(\mu')_0$ is the QUD at the top of μ''s QUD stack. The returned value $\textsf{max}(\textsf{qud}(\mu')_0(w, \kappa))$ takes the maximum degree value of which this QUD is true, given the world-context pair $\langle w, \kappa \rangle$. Thus answers to the QUD are fundamentally based on the common ground: to compute their probability distribution, a world-context pair is sampled from the common ground, and the QUD is evaluated at that world-context pair.

2.5. Linking assumptions. In practice (e.g., in the setting of a formal experiment), an answer needs to be given using a particular testing instrument. We assume that a given testing instrument may be modeled in terms of a family f of distributions representing the likelihood, which is then fixed by a collection Φ of nuisance parameters. Thus we may define a family of *response functions*, parametric in the particular testing instrument (i.e., likelihood function), each of which takes a distribution representing one's prior knowledge k, along with an ongoing discourse m, to produce a distribution over answers to the current QUD given the testing instrument:

$$\begin{aligned}
\texttt{respond}^{f_\Phi} &: \texttt{P}s_\mu \to (s_\mu \to \texttt{P}s_\mu) \to \texttt{P}r \\
\texttt{respond}^{f_\Phi}(k)(m) &\overset{\text{def}}{=} \ \mu \sim k \\
&\qquad\ \mu' \sim m(\mu) \\
&\qquad\ \langle w, \kappa \rangle \sim \textsf{cg}(\mu') \\
&\qquad\ f(\textsf{max}(\textsf{qud}(\mu')_0(w, \kappa)), \Phi)
\end{aligned}$$

The testing instrument employed in our experiments, for example, is a slider scale that records responses on the unit interval $[0, 1]$. A suitable $\texttt{respond}^{f_\Phi}$ would therefore be:

$$\begin{aligned}\texttt{respond}^{\mathcal{N}(\cdot,\sigma)\,\mathsf{T}[0,1]}(k)(m) = \quad & \mu \sim k \\ & \mu' \sim m(\mu) \\ & \langle w, \kappa \rangle \sim \mathsf{cg}(\mu') \\ & \mathcal{N}(\mathsf{max}(\mathsf{qud}(\mu')_0(w, \kappa)), \sigma)\,\mathsf{T}[0, 1]\end{aligned}$$

Here, the likelihood takes a normal distribution with standard deviation σ (the single nuisance parameter), truncated to the unit interval. This likelihood can be viewed as allowing some distribution of response errors, given the intended target response (i.e., the answer to the question).

These are the ingredients we need to model the effects of the fine-grained semantics of the target and question prompt, given a particular inference task. We provide further details in section 5.

3. Factive inferences. To investigate how natural language prompts modulate the distribution of participants responses, we modify an experimental paradigm developed by Degen & Tonhauser (2021, 2022), which they use to experimentally investigate the projective inferences triggered by factive predicates—henceforth, *factive inferences*. We describe the paradigm (subsection 3.1) then discuss prior modeling of their data within PDS (subsection 3.2). The paradigm forms the basis for our experiments in section 4, and we build directly on the prior modeling in section 5.

3.1. MEASURING FACTIVE INFERENCES. Degen & Tonhauser's (2021) main aim is to characterize the influence of world knowledge on factive inferences. To achieve this, they measure factive inferences in the presence of a background fact whose content they manipulate (their experiment 2b). For example, participants are given trials of the form in (2) and asked to respond using a slider scale, with *no* on one end and *yes* on the other.

(2) a. **Fact (which Elizabeth knows):** Zoe is a math major.

b. **Elizabeth asks:** "Does Tim know that Zoe calculated the tip?"

c. Is Elizabeth certain that Zoe calculated the tip?

They focus on the set of twenty clause-embedding predicates listed in (3).[5]

(3) Twenty clause-embedding predicates (Degen & Tonhauser 2022, p. 559, ex. 13)

a. canonically factive: *be annoyed*, *discover*, *know*, *reveal*, *see*

b. non-factive
(i) non-veridical non-factive: *pretend*, *say*, *suggest*, *think*
(ii) veridical non-factive: *be right*, *demonstrate*

c. optionally factive: *acknowledge*, *admit*, *announce*, *confess*, *confirm*, *establish*, *hear*, *inform*, *prove*

[5] The grouping in (3) is due to Degen & Tonhauser 2022 and is based on the prior literature on factivity (Kiparsky & Kiparsky 1970, Karttunen 1971: *et seq*).

Each embedded clause in their experiment is paired with one of two facts: either a fact intended to make the clause likely to be true (as in the example above), or a fact intended to make the clause unlikely to be true. To validate the use of these background facts for this purpose, Degen & Tonhauser (2021) conduct a norming experiment, in which the prior certainties about the truth of the complement clauses featured in their projection experiment are assessed independently, given the same background facts (their experiment 2a). Trials in this experiment ask participants to judge how likely the relevant clause is to be true, given one of the two background facts constructed for it. For example, participants are given trials of the form in (4) and asked to respond using a slider scale, with *impossible* on one end and *definitely* on the other.

(4) a. **Fact:** Zoe is a math major.

b. How likely is it that Zoe calculated the tip?

Degen & Tonhauser find that the by-item means for the forty pairs of complement clauses and background facts, as assessed in their norming experiment, are a good predictor of the inference ratings for items featuring the same complement clauses and facts which they obtain in their experiment investigating projective inferences.

3.2. MODELING FACTIVE INFERENCES. Degen & Tonhauser (2022) and others (White & Rawlins 2018) observe that measures of a predicate's factivity derived from the sort of judgment data Degen & Tonhauser (2021) collect display gradience when the data is aggregated.[6]

Using models developed in PDS, Grove & White ask whether this apparent gradience arises due to *metalinguistic uncertainty*—uncertainty about whether a predicate is factive or not—or *occasional uncertainty*—uncertainty inherently associated with predicate meanings.[7] Under a metalinguistic uncertainty account, different predicates differ in the frequencies with which they trigger factive inferences across uses. Under an occasional uncertainty account, predicates would license inferences with varying degrees of certainty on particular uses, similar to the manner in which a vague predicate, such as *tall*, can license uncertain inferences about the heights of individuals of which it is predicated.

Grove & White fit four models to Degen & Tonhauser's data, varying whether uncertainty about either background world knowledge or factivity is encoded as metalinguistic or occasional. We extend their models here by adding an explicit model of the semantics of the natural language prompt. In the case of Degen & Tonhauser's (2021) original paradigm, this prompt is as in (5).

(5) Is PERSON certain that CLAUSE?

[6] Degen & Tonhauser (2022) argue that this gradience is evidence that factive inferences are fundamentally gradient—following a proposal by Tonhauser, Beaver & Degen (2018)—and that there is no distinct classes of factive predicates. This argument is based on an apparent lack of clear clustering in the aggregate measures of different predicates' derived from inference judgment data collected using this and similar paradigms (White & Rawlins 2018, Ross & Pavlick 2019). It turns out this lack of clear clustering is likely a product of measurement noise: when such noise is appropriately modeled, distinct classes of factive predicates reveal themselves (Kane, Gantt & White 2022).

[7] See Grove & White 2024a for the full formal details of their models.

Following their suggestion that no extant proposal posits that world knowledge should display metalinguistic uncertainty, we focus specifically on two of their models: the *discrete-factivity* model (DF), which regards uncertainty about factivity as metalinguistic and uncertainty about world knowledge as occasional, and the *wholly-gradient* model (WG), which regards both kinds of uncertainty as occasional.

Grove & White find that DF performs the best in a model comparison pitting all four models against each other, as assessed by expected log pointwise predictive densities (ELPDs). They argue that this finding lends support to a view of factivity wherein it is a fundamentally discrete phenomenon, and they discuss how both conventionalist and conversationalist accounts might approach this sort of discreteness.

4. Modifying the prompt. While Grove & White's results are promising, they are consistent with the possibility that the nature of Degen & Tonhauser's prompt biased experimental participants toward making discrete 'yes' or 'no' judgments, even while the contribution to inference judgments made by factive predicates may be gradient. Because the prompt is a polar question, and *yes* and *no* label the slider scale, participants may effectively treat their response as a binary forced choice by providing an answer near *yes* if they are sufficiently certain about the relevant inference, and an answer near *no* if they are not. If so, an *a priori* advantage is conferred on models regarding the contribution to inference of factive predicates as discrete and, thus, models which regard uncertainty about factive inferences as metalinguistic.

To assess the effect the prompt has on participants' responses, we conduct two experiments identical to Degen & Tonhauser's, but which vary the prompt. In both, participants are provided with a *degree* question, which is either about the speaker's degree of certainty (6a) or the degree of *likelihood* that the speaker is certain (6b).

(6) a. How certain is PERSON that CLAUSE?

b. How likely is it that PERSON is certain that CLAUSE?

The prompt in (6a) was paired with a slider labeled *not at all certain* on the left and *completely certain* on the right, while the prompt in (6b) was paired with *impossible* and *definitely* (like Degen & Tonhauser's norming experiment).

The idea behind using a degree questions in both variants is that, insofar as factivity is fundamentally gradient in nature, the degree question should encourage participants to contact that fundamentally gradient representation, whatever it may consist in. Or at the very least, a degree question should not discourage participants from contacting such a gradient representation, and it should not encourage them to discretize it.

All aspects of the experimental materials and methods (besides the prompts) match Degen & Tonhauser's. We recruited 300 participants to give judgments for each prompt. Data from 15 participants was removed in the experiment employing (6a), and data from 7 participants was removed in the experiment employing (6b); for both, we followed Degen & Tonhauser's criteria. Both groups of participants were recruited through Amazon Mechanical Turk and paid $2.

5. Modeling the prompt.

5.1. Adding a prompt model. We fit both the DF and WG models of Grove & White, while manipulating the semantics of the question prompt for both.[8] Specifically, to model the prompt in (6a), we assume that the degree introduced by *certain* ranges over degrees of confidence rather than degrees of probability (following, e.g., Klecha 2012), and thus that its scale is truncated relative to that of *likely*. We refer to this model as the *confidence scale* model.

$$[\![\textit{certain}]\!] = \lambda\mu.\ f \sim \boxed{\lambda\phi.\mathbb{P}\left(\begin{array}{l}\langle w, \kappa\rangle \sim \mathsf{cg}(\mu) \\ \boxed{\phi(w, k)}\end{array}\right)}$$

$$\boxed{\langle\lambda\phi, x, \langle w, \kappa\rangle.\frac{\mathsf{max}(0, f(\phi) - \theta_{\textit{certain}}(\mu))}{1 - \theta_{\textit{certain}}(\mu)} \geq \mathsf{d}_{\textit{certain}}(\kappa)(x), \mu\rangle}$$

To model the prompt in (6b), we assign a semantics to *likely* on which it introduces a degree corresponding to a probability, and where this degree is computed based on the corresponding semantics for *certain*. We refer to this model as the *probability of confidence scale* model.

$$[\![\textit{likely}]\!] = \lambda\mu.\boxed{\left\langle\lambda\phi, \langle w, \kappa\rangle.\mathbb{P}\left(\begin{array}{l}\langle w', \kappa'\rangle \sim \mathsf{cg}(\mu) \\ \boxed{\phi(w', k')}\end{array}\right) \geq \mathsf{d}_{\textit{likely}}(\kappa), \mu\right\rangle}$$

Both models contrast with Grove & White's original model, which encodes a meaning for the prompt (5) according to which it asks for a value on a probability scale. We refer to this model as the *confidence scale* model.

5.2. Model fitting. We fit all models using Hamiltonian Monte Carlo sampling as implemented in STAN—specifically, `CmdStanR` (Gabry & Češnovar 2023). Four chains were sampled for each model to assess convergence, with at least 6,000 warmup samples and 6,000 samples kept per chain. All convergence diagnostics implemented in CmdStanR were conducted. In cases where a model did not converge for reasons that can be solved by drawing more samples, the number of samples was increased until convergence was reached. Even after substantial increases in the number of samples, we were unable to fit to convergence the probability of confidence scale model encoding gradience as occasional in nature. We do not show the results for this model here for this reason.[9]

5.3. Results. We follow Grove & White in reporting model comparisons using ELPD. Figure 1 shows the ELPD for the models of the data collected using prompt (6a) and (6b). Error bars indicate standard errors of the pointwise difference from the best model in each facet.

The DF models are those which regard factive inferences as giving rise to *metalinguistic* gradience, while the WG models are those regarding it as giving rise to *occasional* gradience. The y-axes of each plot provide the three variants of the semantics of the prompt discussed above.

Across the board, the DF models continue to perform the best on both datasets. Meanwhile, we find that confidence scale model performs the best on the dataset containing the prompt in (6a),

[8] We specifically use the variants of Grove & White's models that employ a truncated normal likelihood and that incorporate an antiveridicality component. See their Appendix C for details.

[9] The fits we obtained show extremely poor performance on both datasets, though we cannot read too much into this performance, given that these models did not converge.

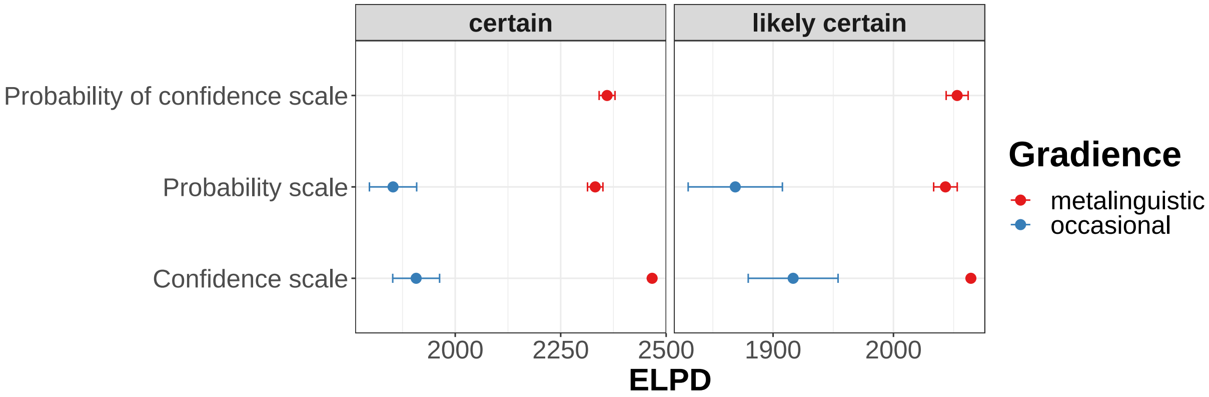

Figure 1: ELPDs for the models of the data collected using prompt (6a) and (6b). We do not show the probability of confidence scale models encoding gradience as occasional in nature, since they do not converge. Error bars indicate standard errors of the pointwise difference from the best model in each facet.

as expected, while the confidence scale and probability of confidence scale models perform about equally on the dataset containing the prompt in (6b) and better than Grove & White's original prompt model (the probability scale model). This result may suggest that, while we approximate the denotation of *certain* well, the denotation we give for *likely* (encoded in the probability of confidence scale model) is not correct. We take this result to highlight the benefits of developing models in the PDS framework, since it makes it straightforward to quantitatively evaluate alternative denotations not considered in the prior literature.

6. Conclusion. Our results suggest that, when analyzing data from an inference judgment task, it can be important to incorporate into one's data analysis regime an explicit representation of the semantics of the natural language prompt used to guide participants on the task. They additionally confirm (i) that the model comparisons obtained by Grove & White do not reflect an *a priori* bias conferred on the discrete models by the experimental task, but rather these models' abilities to capture the distributions of degrees of certainty associated with the inferences generated for the predicates and complement clauses tested; and (ii) that while prior work has potentially provided a good approximation to the correct semantics for predicates like *certain* and *likely*, better approximations may help in studying their semantic contributions to complex inferences involving both.

Future research in this line will aim to leverage PDS to explore the space of possible denotations for such predicates via explicit model comparison of the form we employ here. Beyond improving the our understanding of the semantics of predicates like *likely*, we believe that the approach we have taken here can help us to better understand the fine-grained semantics of attitude predicates like *certain*, as well—a point supported by our success in modeling (6a). One clear opportunity to improving the denotation of predicates like *certain* is to address the fact that our current denotation does not take the attitude holder's epistemic state into account: it merely assumes that the contextual standard for the positive form of the adjective might be attitude-holder-specific. Investigating how an attitude holder's epistemic state might be incorporated into the semantics of *certain* is thus a potentially interesting future direction that is imminently feasible within PDS.

References

Cooper, Robin et al. 1996. *FraCaS: A Framework for Computational Semantics*. Tech. rep. LRE 62-051. The FRACAS Consortium.

Dagan, Ido, Oren Glickman & Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. en. In Joaquin Quiñonero-Candela et al. (eds.), *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, 177–190. Berlin, Heidelberg: Springer. https://doi.org/10.1007/11736790_9.

Degen, Judith & Judith Tonhauser. 2021. Prior Beliefs Modulate Projection. *Open Mind* 5. 59–70. https://doi.org/10.1162/opmi_a_00042. https://doi.org/10.1162/opmi_a_00042 (25 April, 2023).

Degen, Judith & Judith Tonhauser. 2022. Are there factive predicates? An empirical investigation. en. *Language* 98(3). Publisher: Linguistic Society of America, 552–591. https://doi.org/10.1353/lan.0.0271. https://muse.jhu.edu/article/864635 (28 November, 2022).

Farkas, Donka F. & Kim B. Bruce. 2010. On Reacting to Assertions and Polar Questions. *Journal of Semantics* 27(1). 81–118. https://doi.org/10.1093/jos/ffp010. https://doi.org/10.1093/jos/ffp010 (28 April, 2024).

Frank, Michael C. & Noah D. Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science* 336(6084). Publisher: American Association for the Advancement of Science, 998–998. https://doi.org/10.1126/science.1218633. https://www.science.org/doi/10.1126/science.1218633 (20 June, 2022).

Gabry, Jonah & Rok Češnovar. 2023. *CmdStanR*. Tech. rep. https://mc-stan.org/cmdstanr/index.html.

Gantt, William, Benjamin Kane & Aaron Steven White. 2020. Natural language inference with mixed effects. In Iryna Gurevych, Marianna Apidianaki & Manaal Faruqui (eds.), *Proceedings of the ninth joint conference on lexical and computational semantics*, 81–87. Barcelona, Spain (Online): Association for Computational Linguistics. https://aclanthology.org/2020.starsem-1.9.

Ginzburg, Jonathan. 1996. Dynamics and the semantics of dialogue. In Jerry Seligman & Dag Westerståhl (eds.), *Logic, Language, and Computation*, vol. 1, 221–237. Stanford: CSLI Publications.

Goodman, Noah D. & Michael C. Frank. 2016. Pragmatic Language Interpretation as Probabilistic Inference. en. *Trends in Cognitive Sciences* 20(11). 818–829. https://doi.org/10.1016/j.tics.2016.08.005. https://www.sciencedirect.com/science/article/pii/S136466131630122X (11 February, 2021).

Groenendijk, Jeroen & Martin Stokhof. 1984. *Studies on the Semantics of Questions and the Pragmatics of Answers*. University of Amsterdam dissertation.

Grove, Julian & Jean-Philippe Bernardy. 2023. Probabilistic Compositional Semantics, Purely. en. In Katsutoshi Yada et al. (eds.), *New Frontiers in Artificial Intelligence* (Lecture Notes in

Computer Science), 242–256. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36190-6_17.

Grove, Julian & Aaron Steven White. 2024a. *Factivity, presupposition projection, and the role of discrete knowlege in gradient inference judgments*. LingBuzz Published In: submitted. https://ling.auf.net/lingbuzz/007450 (25 April, 2024).

Grove, Julian & Aaron Steven White. 2024b. Probabilistic dynamic semantics. University of Rochester.

Hausser, Roland & Dietmar Zaefferer. 1978. Questions and Answers in a Context-Dependent Montague Grammar. en. In F. Guenthner & S. J. Schmidt (eds.), *Formal Semantics and Pragmatics for Natural Languages*, 339–358. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-9775-2_12. https://doi.org/10.1007/978-94-009-9775-2_12 (19 April, 2024).

Hausser, Roland R. 1983. The Syntax and Semantics of English Mood. en. In Ferenc Kiefer (ed.), *Questions and Answers*, 97–158. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-7016-8_6. https://doi.org/10.1007/978-94-009-7016-8_6 (19 April, 2024).

Jasbi, Masoud, Brandon Waldon & Judith Degen. 2019. Linking hypothesis and number of response options modulate inferred scalar implicature rate. *Frontiers in Psychology* 10. https://doi.org/10.3389/fpsyg.2019.00189. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00189.

Kane, Benjamin, Will Gantt & Aaron Steven White. 2022. Intensional Gaps: Relating veridicality, factivity, doxasticity, bouleticity, and neg-raising. en. *Semantics and Linguistic Theory* 31(0). 570–605. https://doi.org/10.3765/salt.v31i0.5137. https://journals.linguisticsociety.org/proceedings/index.php/SALT/article/view/31.029 (11 May, 2023).

Karttunen, Lauri. 1971. Some observations on factivity. *Paper in Linguistics* 4(1). Publisher: Routledge _eprint: https://doi.org/10.1080/08351817109370248, 55–69. https://doi.org/10.1080/08351817109370248. https://doi.org/10.1080/08351817109370248 (26 June, 2023).

Karttunen, Lauri. 1977. Syntax and semantics of questions. *Linguistics and Philosophy* 1(1). 3–44.

Kiparsky, Paul & Carol Kiparsky. 1970. FACT. en. In *Progress in Linguistics*, 143–173. De Gruyter Mouton. https://doi.org/10.1515/9783111350219.143 (9 June, 2023).

Klecha, Peter. 2012. Positive and Conditional Semantics for Gradable Modals. en. *Proceedings of Sinn und Bedeutung* 16(2). Number: 2, 363–376. https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/433 (14 December, 2023).

MacCartney, Bill. 2009. *Natural language inference*. English. ISBN: 9781109240887. United States – California: Stanford University Ph.D. https://www.proquest.com/docview/305018371/abstract/91568AF09DF74C63PQ/1 (9 June, 2024).

Roberts, Craige. 2012. Information Structure: Towards an integrated formal theory of pragmatics. en. *Semantics and Pragmatics* 5. 6:1–69. https://doi.org/10.3765/sp.5.6. https://semprag.org/index.php/sp/article/view/sp.5.6 (26 July, 2023).

Ross, Alexis & Ellie Pavlick. 2019. How well do NLI models capture verb veridicality? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2230–2240. Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1228. https://aclanthology.org/D19-1228 (26 June, 2023).

Stalnaker, Robert. 1978. Assertion. In Peter Cole (ed.), *Pragmatics*, vol. 9, 315–332. New York: Academic Press.

Tonhauser, Judith, David I. Beaver & Judith Degen. 2018. How Projective is Projective Content? Gradience in Projectivity and At-issueness. en. *Journal of Semantics* 35(3). 495–542. https://doi.org/10.1093/jos/ffy007. (25 June, 2019).

White, Aaron Steven & Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. In Sherry Hucklebridge & Max Nelson (eds.), *NELS 48: Proceedings of the Forty-Eighth Annual Meeting of the North East Linguistic Society*, vol. 48, 221–234. University of Iceland: GLSA (Graduate Linguistics Student Association), Department of Linguistics, University of Massachusetts.