Modeling the prompt in inference judgment tasks

Introduction. A major question in the literature on presupposition projection is whether factive inferences (e.g., *Jo {loves, doesn't love} that Mo left* $\rightsquigarrow$ *Mo left*) are necessary, as classically assumed (Kiparsky and Kiparsky 1970; Karttunen 1971), or not (Tonhauser, Beaver, and Degen 2018). Recent work by Grove and White (2023) addresses this question by fitting statistical models encoding these two assumptions about factive inferences to inference judgment data aimed at capturing factive inferences' strength (Degen and Tonhauser 2021). Grove and White find that models characterizing factive inferences as necessary (henceforth, *discrete models*) fit the inference judgment data better than models that assume they are not (*gradient models*).

Contribution 1. We address a potential flaw in Grove and White's use of Degen and Tonhauser's data for comparing their models: the way participants were asked to respond may artificially improve the discrete models' performance. With the aim of putting the discrete and gradient models on more equal footing, we present two new datasets that keep all other aspects of Degen and Tonhauser's materials constant but which manipulate the natural language prompt participants are given. Consistent with Grove and White 2023, we find that discrete models fit the data better than gradient models for both datasets, supporting Grove and White's claim.

Contribution 2. We show that jointly modeling both the compositional semantics of the target sentence—i.e., the sentence containing the presupposition trigger—and the compositional semantics of the natural language prompt within Grove and White's framework substantially improves fit to response distributions. This finding suggests that it is important to model the interaction between the meaning of a target sentence and the meaning of a prompt when analyzing experimental data.

Degen and Tonhauser's data. Degen and Tonhauser provide experimental participants with a background fact, paired with a predicate taking a complement clause related to that fact.

(1) a. **Fact (which Elizabeth knows):** Zoe is a math major.
Elizabeth asks: "Did Tim discover that Zoe calculated the tip?"
b. Is Elizabeth certain that Zoe calculated the tip?

Participants are asked to provide an answer to the prompt in (1b) on a sliding scale with 'yes' on the left and 'no' on the right. Degen and Tonhauser collect responses for twenty clause-embedding predicates taking one of twenty possible embedded clauses, each paired with either a "high prior" fact or a "low prior" fact. ((1a) illustrates the high prior fact for the given clause.)

Grove and White's models. The aggregate measures of different predicates' factivity derived from inference judgment data show substantial gradience (White and Rawlins 2018; Degen and Tonhauser 2022), and hence constitute potential evidence for variation among predicates in the strength of such inferences. Grove and White ask if this gradience arises due to *metalinguistic uncertainty*—uncertainty about whether a predicate is factive or not—or *contextual uncertainty*—uncertainty inherently associated with predicate meanings. If the uncertainty is metalinguistic, factive inferences may nevertheless be discrete; different predicates would in turn differ in the frequencies with which they trigger such inferences. If it is contextual, predicates would license inferences with varying degrees of certainty, 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 and White fit four models to Degen and Tonhauser's data, varying whether uncertainty about either background world knowledge or factivity is encoded as metalinguistic or contextual. Their models are the *discrete-factivity* model (DF), which regards uncertainty about factivity as metalinguistic and uncertainty about world knowledge as contextual; the *wholly-gradient* model (WG), which regards both kinds of uncertainty as contextual; the *discrete-world* model (DW), which regards uncertainty about factivity as contextual and uncertainty about world knowledge as on a

par with metalinguistic uncertainty; and the *wholly-discrete* model (WD), which regards both kinds of uncertainty as (on a par with) metalinguistic uncertainty. They find that DF performs the best, as assessed by expected log pointwise predictive densities (ELPDs), lending support to the classical view of factivity as a fundamentally discrete phenomenon.

While Grove and White's results are promising, they are consistent with the possibility that the nature of the question prompt exemplified in (1b) biases 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 in (1b) is a polar question, and 'yes' and 'no' label the slider response, 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. Our manipulations of the prompt address this concern, while our new models explicitly target the semantics of the question prompt.

Varying the prompt. We conduct two experiments identical to Degen and 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 (2a) or degree of *likelihood* that the speaker is certain (2b).

(2) a. How certain is Elizabeth that Zoe calculated the tip?
b. How likely is it that Elizabeth is certain that Zoe calculated the tip?

The prompt in (2a) was paired with a slider labeled 'not at all certain' on the left and 'completely certain' on the right, while the prompt in (2b) was paired with 'impossible' and 'definitely'.

Modeling. We obtained the Stan code used to fit each of the four models of factivity from Grove and White, and we constructed two additional models which extend DF, in order to implement a semantics for *certain* and *likely* which allows them to attend to distinct lexical scales. Specifically, to model the prompt in (2a), 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* (yielding the *discrete-factivity-certain* model (DF+C)). To model the prompt in (2b), 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* (yielding the *discrete-factivity-likely-certain* model (DF+LC)).

Results. We compare the (rounded) ELPDs (s.e. in parentheses) of the four original models of Grove and White with our models of the prompts in (2), each fit to the two new datasets.

Experiment	n	DF+C	DF+LC	DF	WG	DW	WD
(2a)	285	2466 (67)	2360 (64)	2183 (65)	1653 (66)	1837 (63)	2000 (56)
(2b)	292	2064 (56)	2052 (56)	1966 (57)	1821 (60)	1524 (48)	1540 (44)

Among the original models, DF continues to perform the best on both datasets. Meanwhile, we find that DF+C performs the best on the dataset containing the prompt in (2a), as expected, while DF+C and DF+LC perform about equally on the dataset containing the prompt in (2b).

Conclusions. Our results (i) confirm that the model comparisons obtained by Grove and 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) suggest that it is important to develop explicit, semantically-motivated linking hypotheses when modeling inference data, not only about the nature of the natural language expression under investigation, but about the question prompt used to elicit an inference. Future research in this line will aim to understand why the model of the prompt in (2a) performs equally well on the dataset containing (2b).