ProPres: Investigating the Projectivity of Presupposition with Various Triggers and Environments

Anonymous ACL submission

Abstract

A presupposition of a sentence refers to information taken for granted by a speaker and projectivity (e.g., *the boy did not shed tears again* presupposes *the boy had shed tears before*) is what makes it distinct from entailment. Although the projectivity might vary depending on the combination of presupposition triggers and environments, previous studies evaluate the performance of models without a human baseline or include only negation as an entailment-canceling environment. Hence, it is necessary to both collect human judgments to obtain a baseline and include various environments to investigate the projectivity of presuppositions comprehensively. In this study, we first reevaluate a previous dataset with recent models and humans, then introducing a new dataset, projectivity of presupposition (ProPres), which includes 12k premise–hypothesis pairs crossing six new triggers with five environments. Our large-scale human judgment experiments provide evidence for variable projectivity, but our model evaluation shows that the models do not capture it. This indicates that the models and humans behave differently in the processing of presuppositions. These results cannot be obtained without the human experiments or the combination of various triggers and environments, suggesting that researchers working on the model performance on pragmatic inferences need to take extra care of the annotation process and the combination of various items.

1 Introduction

There is an open question as to whether language models learn human-like pragmatic inferences (Pavlick, 2022). A speaker does not always explicitly say everything in an utterance but a hearer can readily understand implicit information in it. Investigating whether models can do the same or not is important to develop a better language processing system. In this study, we focus on one type of pragmatic inference: *presupposition*.

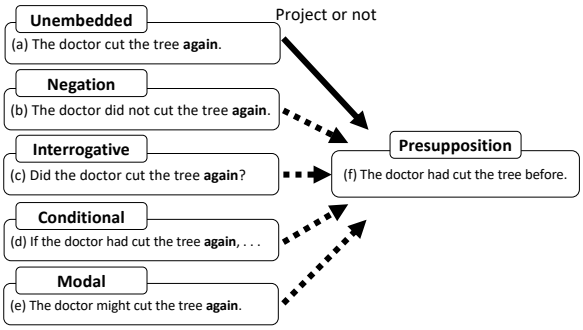

Figure 1: Projectivity of presupposition. A presupposition projects out of entailment-canceling environments. However, it is possible that the projectivity of presupposition can vary depending on the combination of triggers and environments as indicated by the dashed arrows.

Presupposition refers to information taken for granted by a speaker (Stalnaker, 1974; Beaver, 1997). It is often triggered by linguistic items called presupposition triggers such as *again* in (a) in Figure 1. A presupposition of (a) is *the doctor had shed tears before* (f). Presupposition is different from entailment (in this case, *the doctor cut the tree one more time*) as the former is assumed to project out of entailment-canceling environments (e.g., negative (b), interrogative (c), conditional (d), and modal (e) sentences) while the latter does not. In other words, the presupposition (f) holds in the entailment-canceling environments (b–e) but the entailment (*the doctor cut the tree one more time*) does not.

Previous natural language processing studies examine models' performance on presuppositions with a natural language inference (NLI) task (Ross and Pavlick, 2019; Jeretic et al., 2020; Parrish et al., 2021). In the NLI task, one classifies premise–hypothesis pairs into three classes: entailment, contradiction, and neutral (Dagan et al., 2006; Bowman et al., 2015). However, previous studies have some limitations. For instance, Jeretic et al. (2020) do not conduct a human evaluation as a baseline,

Trigger Type	Example Triggers	Example Premise
Iterative	*again*	The assistant split the log **again**.
Aspectual verb	*stop*, *quit*, *finish*	The assistant **stopped** splitting the log.
Manner adverb	*quietly*, *slowly*, *angrily*	The assistant split the log **quietly**.
Factive verb	*remember*, *regret*, *forget*	The assistant **remembered** splitting the log.
Comparative	*better than*, *earlier than*	The assistant split the log **better than** the girl.
Temporal adverb	*before*, *after*, *while*	The assistant split the log **before** bursting into the room.

Table 1: Presupposition triggers with an affirmative (unembedded) premise in ProPres.

Environment	Premise	Hypothesis (target and control)	Label
Unembedded	The doctor shed tears again.	Target: The doctor had (not) shed tears before.	E (C)
Negation	The doctor did not shed tears again.		
Interrogative	Did the doctor shed tears again?	Control: The doctor (did not) shed tears again.	E, C, or N
Conditional	If the doctor had shed tears again, ...		
Modal	The doctor might shed tears again.		

Table 2: Environments used in ProPres. *E* = *Entailment*, *C* = *Contradiction*, and *N* = *Neutral*. The labels in the target conditions are defined based on projectivity. The correct labels in the control conditions depend on the environment.

making models' performance difficult to interpret. Considering that the projectivity of presupposition can vary (Karttunen, 1971; Simons, 2001; Seveganni et al., 2021; Tonhauser et al., 2018, 2019; Degen and Tonhauser, 2021b), we should not define correct labels for the sentence pairs involving presupposition without a large-scale human judgment experiment. Additionally, Parrish et al. (2021) use only negation sentences for entailment-canceling environments; hence, it remains unclear about models' performance on other environments.

To address these concerns, we first evaluate the performance of two transformer-based models, RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020), on an implicature and presupposition diagnostic dataset (IMPPRES; Jeretic et al., 2020) against human judgments on its subset (900 pairs). We find that the best-performed model, DeBERTa, and humans show not only similar but also different projectivity patterns.

Since the nine triggers analyzed in Experiment 1 are not exhaustive (e.g., (Levinson, 1983) and (Potts, 2015) list 27 types of triggers in total), we introduce a novel evaluation dataset, projectivity of presupposition (ProPres), which crosses six new triggers (Table 1) with five environments (Table 2), consisting of 12,000 sentence pairs. We evaluate four models (bag-of-words, InferSent (Conneau et al., 2017), RoBERTa, and DeBERTa) with ProPres against human judgments on its subset (600 pairs). We discover that humans show variable projectivity but the best-performed model, DeBERTa, does not capture it. This finding cannot be obtained without additional triggers combined with various environments.

The results from the two experiments collectively suggest that researchers evaluating NLI systems and creating datasets targeting pragmatic inferences need to take extra care of the annotation process and the combination of various items.

In conclusion, this study makes the following contributions:[1]

- We introduce ProPres using the novel presupposition triggers embedded under various entailment-canceling environments to conduct a comprehensive investigation of presupposition projectivity.
- Our large-scale human judgment experiments provide new evidence for variable projectivity depending on the combination of triggers and environments.
- Model evaluation with human results reveals that the models and humans behave differently in the processing of presuppositions.

2 Background

2.1 Presupposition in Linguistics

Presuppositions are triggered by linguistic items or constructions called presupposition triggers such as *again* in Figure 1 (Stalnaker, 1974; Beaver, 1997). One property that makes presuppositions distinct

[1]We will make our dataset and codebase publicly available.

from other pragmatic inferences such as entailment is projection: presuppositions survive in entailment-canceling environments such as negation (Karttunen, 1973; Heim, 1983). For instance, a presupposition of the affirmative sentence with *again* ((f) given (a)) holds when embedded under negation (b). In the same environment, an entailment (here, *the doctor cut the tree one more time*) is canceled.

Importantly, previous studies show that the projectivity of presupposition can vary depending on factors such as context, lexical items, prior beliefs, the speaker's social identity, and prosodic focus in the utterance (Karttunen, 1971; Simons, 2001; Stevens et al., 2017; Tonhauser et al., 2018, 2019; Degen and Tonhauser, 2021b). One remaining question here is whether variable projectivity is associated with the interaction of triggers and environments. For instance, it is possible that a presupposition triggered by *again* is more likely to project over the negation (b) than the conditional (d) or vice versa. To investigate this question comprehensively, this study collects human judgments on presuppositions using a wide range of triggers and environments.

2.2 Presuppositions in NLI

NLI datasets have been introduced to evaluate model performance on pragmatic inferences (Ross and Pavlick, 2019; Jeretic et al., 2020; Parrish et al., 2021).

IMPPRES (Jeretic et al., 2020) is a template-based dataset designed to investigate presupposition (and implicature). Using this dataset, Jeretic et al. (2020) find that models (e.g., BERT (Devlin et al., 2019)) learn the projection of presuppositions triggered by *only*, cleft existence, possessive existence, and question. However, they do not conduct a human evaluation. As discussed in Section 2.1, it is possible that projectivity can vary depending on the combination of triggers and environments. In addition, humans are known to make seemingly unsystematic judgments about projection on both natural (Ross and Pavlick, 2019; de Marneffe et al., 2019) and controlled (White and Rawlins, 2018) items. Hence, it is unclear whether model performance on the projectivity reported by Jeretic et al. (2020) aligns with actual human judgments. Following Parrish et al. (2021), we collect human judgments on a subset of IMPPRES and ProPres to obtain a baseline for model evaluation.

NOPE (Parrish et al., 2021) includes naturally-occurring data with presupposition triggers. With this dataset, Parrish et al. (2021) evaluate transformer-based models against human performance, finding that models behave similarly to humans despite the fact that the training data MNLI (Williams et al., 2018) includes few presupposition cases. One limitation of NOPE is that it includes only one entailment-canceling environment: negation. To make a more general conclusion about the models' performance, it is necessary to include various types of environments. Following Jeretic et al. (2020), the entailment-canceling environments in ProPres include not only negation but also an interrogative, conditional, and modal.

3 Experiment 1: Reevaluating IMPPRES

One limitation in Jeretic et al. (2020) is that they evaluate language models without human evaluation, leaving it open whether models capture any variable projectivity in IMPPRES. We thus collect human judgments on a subset of IMPPRES, then evaluating whether the performance of two transformer-based models, RoBERTa and DeBERTa, aligns with human results.[2]

3.1 Experimental Setup

Human Evaluation We collect human judgments on a subset of IMPPRES (900 sentence pairs). It uses nine triggers (*all N*, *both*, change of state verbs (CoS), cleft existence, *only*, possessive definites, possessive uniqueness, and question). We focus on conditions in which triggers are embedded under five environments, namely the affirmative sentence (unembedded), negative sentence (negation), conditional antecedent (conditional), modal sentence (modal), and interrogative,[3] and in which a hypothesis is affirmative or negated. Each of the extracted sentence pairs is judged by 9.4 people on average.

Model Evaluation We evaluate RoBERTa-base (Liu et al., 2019), and DeBERTa-v3-large (He et al., 2020). We use Huggingface's (Wolf et al., 2020) pretrained RoBERTa-base and DeBERTa-v3-large fine-tuned on MNLI. We do not evaluate models such as a bag-of-words (BOW) model and an InferSent model (Conneau et al., 2017) because their

[2]Details of our experiment (e.g., qualification, instructions, exclusion criteria) is reported in Appendix B.

[3]Examples of triggers and environments in IMPPRES are provided in Appendix C.

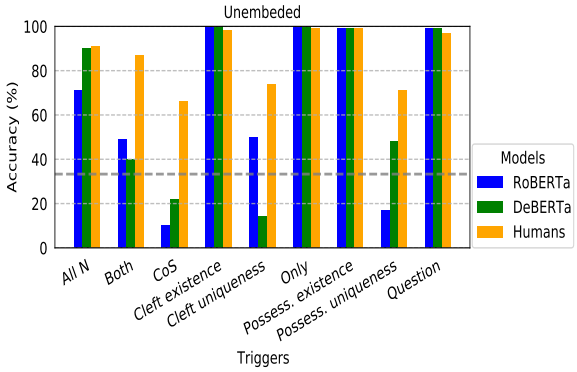

Figure 2: Results on the unembedded triggers in IMPPRES. The dashed lines indicate chance performance (33.3%).

performance is not interpretable due to their variable performance on controls (Jeretic et al., 2020).

3.2 Results and Discussion

Unembedded We use the accuracy for the unembedded triggers as criteria to exclude triggers from the analysis of entailment-canceling environments. When a trigger occurs in an affirmative sentence (unembedded), presupposition equals entailment (e.g., *Bob only ran* presupposes and entails *Bob ran*) (Jeretic et al., 2020). Hence, the low accuracy of humans for any unembedded triggers can be taken as an indication of dataset artifacts. In conjunction with human results, we interpret models' low accuracy as the dataset artifacts or their lack of knowledge of the basic meanings of the triggers.

Compared to the other triggers (acc. $<$ 90.0% on average), humans show low accuracy for CoS (66.3%), cleft uniqueness (74.1%), and possessed uniqueness (71.9%), as exemplified below:[4]

(1) CoS: Omar is hiding Ben.
→ Ben was out in the open.

(2) Cleft uniqueness: It is that doctor who left.
↛ More than one person left.

(3) Possessive uniqueness: Tom's car that broke bored this committee.
→ Tom has exactly one car that broke.

We suspect that the low accuracy for CoS is due to semantic ambiguity. For instance, people might label the pair in (1) as neutral or contradiction because Ben was not necessarily exposed before being hidden. Regarding the other two conditions, we do not understand the source of the low accuracy at this point. In linguistics, results from judgment experiments sometimes do not support generalizations made by theoreticians (Gibson and Fedorenko, 2013). Additionally, NLI research reports disagreements about natural language inferences (Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Zhang and de Marneffe, 2021). The current results then suggest that judgments on presuppositions of cleft and possessive uniqueness are not as robust as Jeretic et al. (2020) may assume. We do not address these three triggers in the following analysis as they might confound the results.

Both RoBERTa and DeBERTa show high accuracy for most triggers (acc. $<$ 90.0%). Two exceptions are *all N* and *both*. RoBERTa shows lower accuracy for *all N* (71.0%) than DeBERTa (89.5%) (e.g., *all four men that departed telephoned* → *exactly four men departed*). With respect to *both* (e.g., *both guys who ran jumped* → *exactly two guys ran*), neither DeBERTa nor RoBERTa performs well (39.0% and 49.0%, respectively). Since the two models are roughly comparable in performance, we analyze only DeBERTa below.

Based on these results, the following analysis includes the five triggers, *all N*, cleft existence, *only*, possessive existence, and question.[5]

Entailment-Canceling Environments To analyze results on entailment-canceling environments, we use the term, projectivity, instead of accuracy. Since human judgments on projectivity can vary, as discussed in Section 2.1, we should not define correct labels for sentence pairs involving presupposition. Projectivity is calculated based on whether presupposition projects. For instance, if one classifies the pair (*P*: *did Tom only terrify Ken?* and *H*: *Tom terrified Ken*) as entailment, it is considered projective. Taking another example, if the hypothesis *Tom did not terrify Ken* is judged as contradiction given the same premise, it counts as projective. Otherwise, these two examples are taken as non-projective.

Figure 3 presents results on the four environments: negation, conditional, interrogative, and modal. Overall, DeBERTa and humans are similar in projectivity. One notable similarity between them is that *only* in conditional (e.g., *if Mary only testifies, ...* → *Mary testifies*) and modal (e.g., *Mary might only testify* → *Mary testifies*) has relatively

[4]Throughout the paper, the examples from the dataset are slightly simplified (e.g., changing *Thomas* to *Tom*) for the space reason.

[5]We report all results including excluded triggers in Appendix D.

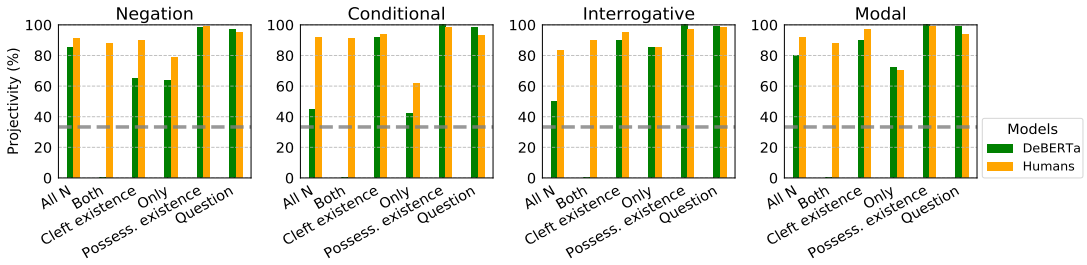

Figure 3: Results on entailment-canceling environments in IMPPRES. DeBERTa's results on *both* are not presented.

low projectivity (61.8% and 69.8% for humans and 41.5% and 72.0% for DeBERTa, respectively). We confirm this by evaluating DeBERTa only with the human-judged sentence pairs (35.0% and 65.0% for conditional and modal, respectively).[6]

A closer look at the results reveals that DeBERTa takes some conditions less projective than humans. Humans take cleft existence in negation (e.g., *it isn't that guest who ran complained* → *someone complained*) as projective (89.7%) while DeBERTa predicts it as less projective (65.0%). We also see a difference in *only* in negation (e.g., *Katy didn't only testify, ...* → *Katy testified*) (78.6% and 64.0% for humans and models, respectively), but our model evaluation with the human-judged pairs does not confirm it (80.0%). In addition, humans judge *all N* in conditional (e.g., *if all nine actors that left slept, ...* → *exactly nine actors left*) and in interrogative (e.g., *did all nine actors that left sleep?* → *exactly nine actors left*) as projective (91.8% and 82.6%, respectively) but DeBERTa takes them as less projective (45.0% and 49.5%, respectively). These results collectively indicate DeBERTa's lack of knowledge of cleft existence in negation and *all N* in conditional and interrogative.

In summary, humans take most presupposition cases as projective with some variability in *only* embedded under conditional and modal. This finding adds to the previous research on variable projectivity in other cases (Stevens-Guille et al., 2020; Tonhauser et al., 2018, 2019; Degen and Tonhauser, 2021a,b). Additionally, DeBERTa and humans show not only similarities but also quite a few differences in projectivity. This leads us to conclude that DeBERTa does not learn how to process presuppositions in a human-like way. These results cannot be obtained without human judgments since there is no predetermined correct label.

[6]In the following analysis and Experiment 2, we report results of the model evaluation with human-judged data only if they do not confirm the similarity or difference between models and humans based on all data.

4 Experiment 2: ProPres

The triggers in IMPPRES are not exhaustive as we can find more triggers in the literature (e.g., 27 triggers in Levinson (1983) and Potts (2015) in total). To investigate variable projectivity and models' behavior more comprehensively, we conduct the second experiment with ProPres using new triggers embedded under various environments.

4.1 Data Generation

Triggers and Environments ProPres has six types of presupposition triggers: (1) an iterative *again*, (2) aspectual verbs, (3) manner adverbs, (4) factive verbs, (5) comparatives, and (6) temporal adverbs, as presented in Table 1. We select these triggers from Levinson (1983) and Potts (2015) because they are not included in IMPPRES and can be easily incorporated into templates.

ProPres has five environments: (1) affirmative sentences (unembedded), (2) negative sentences (negation), (3) polar questions (interrogative), (4) counterfactual conditional antecedents (conditional), and (5) modal sentences (modal), as exemplified in Table 2. The unembedded is used to test whether humans and models can identify presupposition as entailment when triggers are unembedded, as discussed in Section 3.2. The counterfactual conditional antecedent is not usually used as an entailment-canceling environment, but we include it to ensure that conditional controls have clear correct labels (entailment or contradiction), as discussed in the following paragraph. We generate affirmative and negative premises for each condition. Combining six trigger types, five environment types, and two hypothesis polarity types results in 60 conditions. Generating 100 premise–hypothesis pairs for each condition yields 6,000 pairs.[7]

[7]We provide examples for each condition in Appendix A.

We make a control condition corresponding to each target condition where a hypothesis is an affirmative or negative version of its premise, as shown in Table 2. The control conditions are used as a sanity check in a human experiment. They are also important to investigate whether the models rely on lexical overlap (McCoy et al., 2019) or *negation* heuristics (Gururangan et al., 2018). For instance, models are expected to label the affirmative hypothesis in Table 2 as entailment if they rely on the lexical overlap heuristic because of the high lexical overlap between the premise and hypothesis. Additionally, they should label the negative hypothesis as contradiction if they use the *negation* heuristic due to the presence of *not*. Only if models predict correctly in the control conditions, we can say that their predictions in the corresponding target conditions reflect projectivity rather than heuristics. Creating 100 pairs for each control condition results in 6,000 pairs. In total, ProPres comprises 12,000 pairs.

Templates We generate sentence pairs with templates on the basis of the codebase developed by Yanaka and Mineshima (2021).[8] The examples are given below:[9]

(4) The N did not VP again.
(The girl did not hurt others again.)
$\rightarrow$ ($\not\rightarrow$) The N had (not) VP before.
(The girl had (not) hurt others before.)

In VP, we use verbs having the same form in past tense and past participle forms (e.g., *hurt*) to make the morphological difference between a premise and hypothesis as small as possible. This is crucial to test whether models rely on the lexical overlap heuristic in the control conditions.

The use of templates has two advantages. First, it allows us to test whether models rely on the lexical overlap heuristic and *negation* heuristic. In addition, we can control the effect of plausibility. Previous work shows that the projectivity of presupposition varies depending on its content (Karttunen, 1971; Simons, 2001; Tonhauser et al., 2018). For instance, the sentence *John didn't stop going to the restaurant* leads to the inference *John had been going to the restaurant before*. In contrast, the sentence *John didn't stop going to the moon* is less likely to yield the inference *John had been going to the moon before*. This difference can be attributed to our world knowledge: it is more plausible for one to go to the restaurant than the moon. As the plausibility effect is not the focus of this study, we use templates to control it.

[8] https://github.com/verypluming/JaNLI

[9] A full list of the templates and their example sentences is provided in Appendix A.

4.2 Experimental Setup

Human Evaluation We randomly select 10 out of 100 pairs from each target condition and two pairs from each control condition, extracting 600 and 120 pairs in total, respectively. Due to some revision of ProPres during the dataset creation, judgments on the modal environment and comparative trigger are collected in Experiment 1 (200 pairs in total). As a result, each of the extracted pairs is judged by 56.7 people on average (9.4 people for the modal and comparative on average).

Model Evaluation We evaluate four models: BOW, InferSent (Conneau et al., 2017), RoBERTa-base (Liu et al., 2019), and DeBERTa-v3-large (He et al., 2020). For the first two models, we follow Parrish et al. (2021)'s implementation[10] and use MNLI (Williams et al., 2018) to fine-tune the parameters. We use the GloVe embeddings for the word-level representations (Pennington et al., 2014). For the two transformer-based models, we use RoBERTa-base and DeBERTa-v3-large fine-tuned on MNLI as in Experiment 1.

4.3 Results and Discussion

Control Conditions Figure 4 shows results on control conditions. The performance of InferSent and BOW models is variable; hence, we do not analyze them below. Similar to humans, RoBERTa and DeBERTa perform well on the unembedded, negation, and conditional (P_1–P_3 in (5)), indicating that they do not use the lexical overlap heuristic or *negation* heuristic in these cases.

(5) P_1: The boy cut the tree again.
P_2: The boy did not cut the tree again.
P_3: If the boy had cut the tree again, ...
P_4: Did the boy cut the tree again?
P_5: The boy might cut the tree again.
$H_{I(2)}$: The boy (did not) cut the tree again.

RoBERTa, DeBERTa, and humans perform poorly on the interrogative and modal (P_4 and P_5 in (5)) in which the correct label is neutral (Jeretic et al., 2020) (31.8%, 50.0%, and 51.1% for interrogative and 3.5%, 16.7%, and 48.1% for modal,

[10] https://github.com/nyu-mll/nope

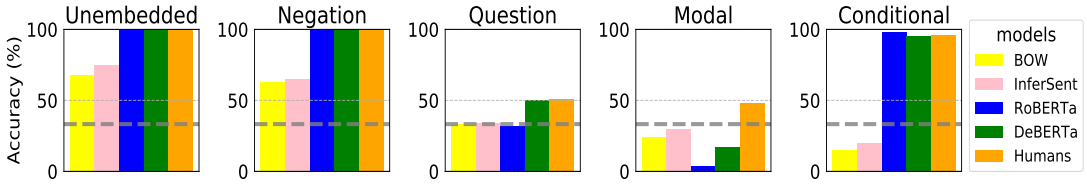

Figure 4: Results on control conditions in ProPres.

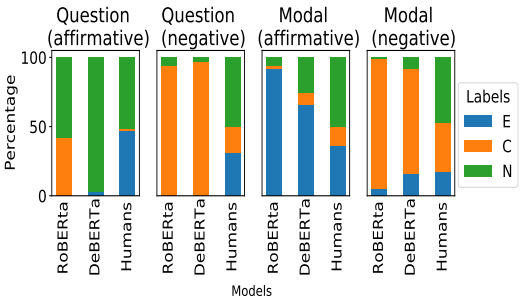

Figure 5: Distributions of labels in the interrogative and modal with an affirmative or negative hypothesis.

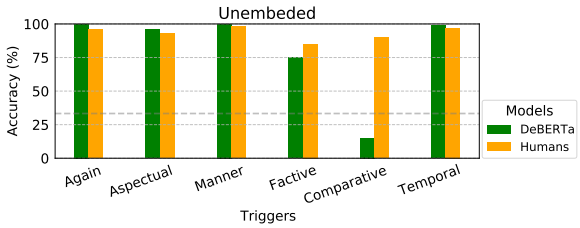

Figure 6: Results on the unembedded condition in ProPres for DeBERTa and humans.

respectively). Distributions of labels in these conditions (Figure 5) show that the majority of labels in humans are neutral. One exception is the interrogative with an affirmative hypothesis (P_4 and H_1 in (5)): distributions of entailment and neutral are comparable (46.5% and 52.4%, respectively). We suspect that humans understood this condition as a confirmation question in which the affirmative form of the interrogative (in this case, H_1) is presupposed, resulting in the high percentage of entailment.

In the same condition, the label distributions of DeBERTa and RoBERTa do not mirror those of humans. RoBERTa shows a relatively high percentage of contradiction (57.5%) whereas DeBERTa shows a very high percentage of neutral (97.1%). In the interrogative with the negative hypothesis (P_4 and H_2), RoBERTa and DeBERTa assign contradiction the majority of the time (93.7% and 97.1%, respectively), indicating the *negation* heuristic.

The two models do not mirror humans in performance on the modal. Their majority labels in the modal with affirmative and negative hypotheses (P_5 with H_1 and H_2) are entailment and contradiction, respectively, suggesting that they use the lexical overlap and *negation* heuristics in the modal.

These variable results for DeBERTa and RoBERTa are inconsistent with Jeretic et al. (2020), in which BERT achieves high accuracy for the interrogative and modal controls by assigning the neutral label. The reason might be that the combination of the two environments with new triggers in ProPres perturbs the models.

Overall, the performance of RoBERTa and DeBERTa is interpretable in the case of three environments: unembedded, negation, and conditional. Hence, we do not report model results on the interrogative and modal.[11] In addition, since the two models are comparable in accuracy, we only analyze DeBERTa's performance below.

Unembedded Figure 6 shows results on the unembedded triggers. Overall, DeBERTa and humans achieve high accuracy for all triggers. One exception is DeBERTa's poor performance on the comparative (e.g., *the girl read the letter better than the boy* → *the boy read the letter*) (14.5%), indicating its lack of the basic knowledge of this trigger. Hence, we do not report DeBERTa's predictions about the comparative below.

Entailment-Canceling Environments Figure 7 shows results on the entailment-canceling environments. Humans provide evidence for variable projectivity (range 55.1–99.8%). Manner adverbs show relatively weak projectivity over the negation (P_1 in (6)) and interrogative (P_2) (58.3% and 66.6%, respectively).

(6) P_1: The man did not hurt others seriously.
P_2: Did the man hurt others seriously?
P_3: If the man had hurt others seriously, ...
P_4: The man might hurt others seriously.

[11] We report all results including excluded conditions in Appendix D.

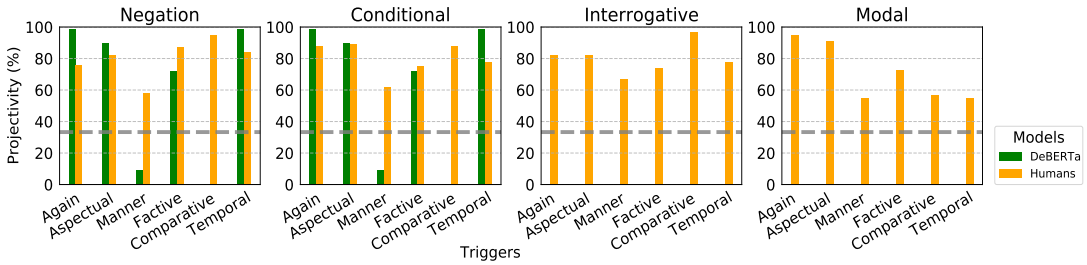

Figure 7: Results on entailment-canceling environments in ProPres. DeBERTa's results on the interrogative and modal environments and the comparative trigger are not shown.

$H_{1(2)}$: The man (did not) hurt others.

Stevens et al. (2017) and Tonhauser et al. (2019) show that the projectivity of the presupposition of manner adverbs is sensitive to what is focalized in the utterance. For instance, the presupposition (H_1) is more likely to project when the manner adverb is focused (*did the man hurt others SERIOUSLY?*) than when the subject is focused (*did the MAN hurt others seriously?*). Since our dataset has no prosodic information signaling focus, humans might find these conditions ambiguous, yielding the weak projectivity. Crucially, we also discover that the manner adverbs are weakly projective in the conditional (P_3) and modal (P_4) (62.0% and 55.1%, respectively). This suggests that information structural cues such as prosodic focus might also play a role in the projectivity of presupposition triggered by the manner adverbs embedded under the conditional and modal.

In the modal, temporal adverbs (P_1 in (7)) and comparatives (P_2) have weaker projectivity (54.7% and 57.4%, respectively) than the other three triggers excluding the manner adverbs (range 73.2–95.2%). These two triggers are projective in the other three environments (ranges 77.7–83.6% and 87.9–97.5% for the temporal adverbs and comparatives, respectively). This suggests that the projectivity of presuppositions of these triggers varies depending on the environment.

(7) P_1: Tom might sing after reading.
P_2: The lady might sing better than Tom.
$H_{1(2)}$: Tom (did not) read.

DeBERTa mirrors humans in projectivity to some extent but it is different from them. It predicts that the manner adverbs in the negation and conditional (P_1 and P_3 in (6), respectively) are not projective (8.5% and 14%, respectively), contrary to humans (58.3% and 62.0%, respectively). This indicates that either DeBERTa lacks the knowledge of these two cases or processes them as if the subject is focalized (e.g., *did the MAN hurt others seriously?*). DeBERTa takes the other six conditions as projective (range 71.5–99.5%), similar to humans.

In summary, Experiment 2 shows variable projectivity in 6 out of the new 24 conditions, contrary to Experiment 1, in which we observe it in two out of 24 conditions. This contrast highlights that we need to combine various triggers and environments to investigate variable projectivity. In addition, we discover that DeBERTa does not capture the variable projectivity, suggesting that DeBERTa's ability to process presupposition is not human-like.

5 Conclusion

In Experiment 1, we conclude that humans and models are similar but different in making a pragmatic inference: presupposition. Experiment 2 then makes this conclusion stronger by using new presupposition triggers. Overall, human results provide evidence for variable projectivity in some conditions (2 out of 24 and 6 out of 24 conditions in Experiments 1 and 2, respectively) but the best-performed model, DeBERTa, does not capture it most of the time, indicating that it does not learn generalizations consistent with the human intuition.

In our experiments, quite a few conditions are excluded from the analysis due to the dataset artifacts, disagreements in judgments, or the models' lack of knowledge. This indicates that we need to be careful with dataset creation and that we may need to train models with data targeting presuppositions so that models can learn their basic meanings.

This study might be limited in terms of social impacts, but it demonstrates the importance of the annotation process and the combination of various items, which can be applied to other research directly relevant to real-life applications such as machine translation.

References

David I. Beaver. 1997. Presupposition. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of logic and language*, pages 939–1008. MIT Press and North-Holland, Cambridge, MA and Amsterdam.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. Evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. In *Sinn und Bedeutung 23*, volume 2, pages 107–124.

Judith Degen and Judith Tonhauser. 2021a. Are there factive predicates? an empirical investigation. *LingBuzz*.

Judith Degen and Judith Tonhauser. 2021b. Prior beliefs modulate projection. *Open Mind*, 5:59–70.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Edward Gibson and Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1-2):88–124.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Irene Heim. 1983. On the conversational basis of some presuppositions. In *Proceedings of the 2nd West Coast Conference on Formal Linguistics*, pages 114–125, Stanford, CA. Stanford Linguistics Association.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Lauri Karttunen. 1971. Some observations on factivity. *Papers in Linguistics*, 4:55–69.

Lauri Karttunen. 1973. Presuppositions of compound sentences. *Linguistic inquiry*, 4(2):169–193.

Steven C. Levinson. 1983. *Pragmatics*. Cambridge University Press, Cambridge, UK.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A corpus of naturally-occurring presuppositions in English. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 349–366, Online. Association for Computational Linguistics.

Ellie Pavlick. 2022. Semantic structure in deep learning. *Annual Review of Linguistics*, 8.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word

representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Christopher Potts. 2015. Presupposition and implicature. In Shalom Lappin and Chris Fox, editors, *The handbook of contemporary semantic theory*, volume 2, pages 168–202. Wiley-Blackwell, Oxford, UK.

Alexis Ross and 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)*, pages 2230–2240, Hong Kong, China. Association for Computational Linguistics.

Karin Sevegnani, David M. Howcroft, Ioannis Konstas, and Verena Rieser. 2021. OTTers: One-turn topic transitions for open-domain dialogue. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2492–2504, Online. Association for Computational Linguistics.

Mandy Simons. 2001. On the conversational basis of some presuppositions. In *Proceedings of Semantics and Linguistics Theory XI*, pages 431–448, Ithaca, NY. CLC Publications.

Robert Stalnaker. 1974. Pragmatic presuppositions. In Milton K. Munitz and Peter K. Unger, editors, *Semantics and Philosophy*, pages 135–148. New York University Press, New York.

Jon Stevens, Marie-Catherine de Marneffe, Shari R Speer, and Judith Tonhauser. 2017. Rational use of prosody predicts projection in manner adverb utterances. In *39th Annual Meeting of the Cognitive Science Society*, pages 1144–1149.

Symon Stevens-Guille, Aleksandre Maskharashvili, Amy Isard, Xintong Li, and Michael White. 2020. Neural NLG for methodius: From RST meaning representations to texts. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 306–315, Dublin, Ireland. Association for Computational Linguistics.

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

Judith Tonhauser, Marie-Catherine de Marneffe, Shari R Speer, and Jon Stevens. 2019. On the information structure sensitivity of projective content. In *Proceedings of Sinn und Bedeutung*, volume 23, pages 363–390.

Aaron S White and Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. In *Proceedings of the 48th Annual Meeting of the North East Linguistic Society*, pages 221–234.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hitomi Yanaka and Koji Mineshima. 2021. Assessing the generalization capacity of pre-trained language models through Japanese adversarial natural language inference. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 337–349, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. Identifying inherent disagreement in natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.

A Templates

Tables 3–7 contain templates of premises and hypotheses for six triggers crossed with five environments in ProPres.

B Crowdsourcing Experiments

Before the experiment, each participant is asked to read a written instruction about the NLI task carefully. During the experiment, the following instruction is presented on a screen: 'Select the response based on how likely you think the second statement is to be true, using the information in the first statement and your background knowledge about how the world works. If you think that the second statement is true, click Entailment. If you think that it is false, select Contradiction. If you are not sure, select Neutral.' This instruction is adopted from Parrish et al. (2021) and modified

Trigger	Template	Premise and Hypothesis
Again	P: The N VP again. H_1: The N had VP before. H_2: The N had not VP before.	P: The doctor shed tears again. H_1: The doctor had cut the tree before. H_2: The doctor had not shed tears before.
Manner adverbs	P: The N VP MADV. H_1: The N VP. H_2: The N did not VP.	P: The doctor shed tears slowly. H_1: The doctor shed tears. H_2: The doctor did not shed tears.
Comparatives	P: The N_1 VP ADVer than N_2. H_1: The N_2 VP. H_2: The N_2 did not VP.	P: The doctor shed tears better than the singer. H_1: The singer shed tears. H_2: The singer did not shed tears.
Temporal adverbs	P: The N VP_1 TADV VP_2ing. H_1: The N VP_2. H_2: The N did not VP_2.	P: The doctor shed tears before hurting others. H_1: The doctor hurt others. H_2: The doctor did not hurt others
Aspectual verbs	P: The N ASP VPing. H_1: The N had been VPing. H_2: The N had not been VPing.	P: The doctor stopped shedding tears. H_1: The doctor had been shedding tears. H_2: The doctor had not been shedding tears.
Factive verbs	P: The N Factive VPing. H_1: The N VP. H_2: The N did not VP.	P: The doctor regretted shedding tears. H_1: The doctor shed tears. H_2: The doctor shed tears.

Table 3: Templates for affirmative sentences.

Trigger	Template	Premise and Hypothesis
Again	P: The N did not VP again. H_1: The N had VP before. H_2: The N had not VP before.	P: The doctor did not shed tears again. H_1: The doctor had shed tears before. H_2: The doctor had not shed tears before.
Manner adverbs	P: The N did not VP MADV. H_1: The N VP. H_2: The N did not VP.	P: The doctor did not shed tears slowly. H_1: The doctor shed tears. H_2: The doctor did not shed tears.
Comparatives	P: The N_1 did not VP ADVer than N_2. H_1: The N_2 VP. H_2: The N_2 did not VP.	P: The doctor did not shed tears better than the singer. H_1: The singer shed tears. H_2: The singer did not shed tears.
Temporal adverbs	P: The N did not VP_1 TADV VP_2ing. H_1: The N VP_2. H_2: The N did not VP_2.	P: The doctor did not shed tears before hurting others. H_1: The doctor hurt others. H_2: The doctor did not hurt others.
Aspectual verbs	P: The N did not ASP VPing. H_1: The N had been VPing. H_2: The N had not been VPing.	P: The doctor did not stop shedding tears. H_1: The doctor had been shedding tears. H_2: The doctor had not been shedding tears.
Factive verbs	P: The N did not Factive VPing. H_1: The N VP. H_2: The N did not VP.	P: The doctor did not regret shedding tears. H_1: The doctor shed tears. H_2: The doctor did not shed tears.

Table 4: Templates for negative sentences.

Trigger	Template	Premise and Hypothesis
Again	P: Did the N VP again? H_1: The N had VP before. H_2: The N had not VP before.	P: Did the doctor shed tears again? H_1: The doctor had shed tears before. H_2: The doctor had not shed tears before.
Manner adverbs	P: Did the N VP MADV? H_1: The N VP. H_2: The N did not VP.	P: Did the doctor shed tears slowly? H_1: The doctor shed tear. H_2: The doctor did not shed tears.
Comparatives	P: Did the N_1 VP ADVer than N_2? H_1: The N_2 VP. H_2: The N_2 did not VP.	P: Did the doctor shed tears better than the singer? H_1: The doctor shed tears. H_2: The doctor did not shed tears.
Temporal adverbs	P: Did the N VP_1 TADV VP_2ing? H_1: The N VP_2. H_2: The N did not VP_2.	P: Did the doctor shed tears before spreading the rumor? H_1: The doctor spread the rumor. H_2: The doctor did not spread the rumor.
Aspectual verbs	P: Did the N ASP VPing? H_1: The N had been VPing. H_2: The N had not been VPing.	P: Did the doctor stop shedding tears? H_1: The doctor had been shedding tears. H_2: The doctor had not been shedding tears.
Factive verbs	P: Did the N Factive VPing? H_1: The N VP. H_2: The N did not VP.	P: Did the doctor regret shedding tears? H_1: The doctor shed tears. H_2: The doctor did not shed tears.

Table 5: Templates for interrogatives.

Trigger	Template	Examples
Again	P: If the N_1 had VP again, the N_2 would have VP_2. H_1: The N_1 had VP_1 before. H_2: The N_1 had not VP_1 before.	P: If the doctor had shed tears again, the singer could have spread the news. H_1: The doctor had shed tears before. H_2: The doctor had not shed tears before.
Manner adverbs	P: If the N_1 VP_1 MADV, the N_2 would have VP_2. H_1: The N_1 VP_1. H_2: The N_1 did not VP_1.	P: If the doctor shed tears slowly, the singer could have spread the news. H_1: The doctor shed tears. H_2: The doctor did not shed tears.
Comparatives	P: If the N_1 had VP_1 ADVer than N_3, the N_2 would have VP_2. H_1: The N_1 VP_1. H_2: The N_1 did not VP_1.	P: If the doctor had shed tears better than the singer, the artist could have spread the news. H_1: The singer shed tears. H_2: The singer did not shed tears.
Temporal adverbs	P: If the N_1 had VP_1 TADV VP_2ing, the N_2 would have VP_3. H_1: The N_1 VP_2. H_2: The N_1 did not VP_2.	P: If the doctor had shed tears before spreading the rumor, the singer could have burst into the room. H_1: The doctor spread the rumor. H_2: The doctor did not spread the rumor.
Aspectual verbs	P: If the N_1 ASP VP_1ing, the N_2 would have VP_2. H_1: The N_1 had been VP_1ing. H_2: The N_1 had not been VP_1ing.	P: If the doctor had stopped shedding tears, the singer could have spread the rumor. H_1: The doctor had been shedding tears. H_2: The doctor had not been shedding tears.
Factive verbs	P: If the N_1 Factive VP_1ing, the N_2 would have VP_2. H_1: The N_1 VP_1. H_2: The N_1 did not VP_1.	P: If the doctor had regretted shedding tears, the singer could have spread the rumor. H_1: The doctor shed tears. H_2: The doctor did not shed tears.

Table 6: Templates for counterfactual conditionals.

Trigger	Template	Premise and Hypothesis
Again	P: The N Modal VP again. H_1: The N had VP before. H_2: The N had not VP before.	P: The doctor might shed tears again. H_1: The doctor had shed tears before. H_2: The doctor had not shed tears before.
Manner adverbs	P: The N Modal VP MADV. H_1: The N VP. H_2: The N did not VP.	P: The doctor might shed tears slowly. H_1: The doctor shed tears. H_2: The doctor did not shed tears.
Comparatives	P: The N_1 Modal VP ADVer than N_2. H_1: The N_2 VP. H_2: The N_2 did not VP.	P: The doctor might shed tears better than the singer. H_1: The singer shed tears. H_2: The singer did not shed tears.
Temporal adverbs	P: The N Modal VP_1 TADV VP_2ing. H_1: The N VP_2. H_2: The N did not VP_2.	P: The doctor might shed tears before spreading the rumor. H_1: The doctor spread the rumor. H_2: The doctor did not spread the rumor.
Aspectual verbs	P: The N Modal ASP VPing. H_1: The N had been VPing. H_2: The N had not been VPing.	P: The doctor might stop shedding tears. H_1: The doctor had been shedding tears. H_2: The doctor had not been shedding tears.
Factive verbs	P: The N Modal Factive VPing. H_1: The N VP. H_2: The N did not VP.	P: The doctor might regret shedding tears. H_1: The doctor shed tears. H_2: The doctor did not shed tears.

Table 7: Templates for modal sentenses.

according to our experiment. All data are collected anonymously except workers' ID.

Experiment 1 We randomly select 10 out of 100 premise–hypothesis pairs from each condition in IMPPRES, extracting 900 pairs in total. These sentence pairs are divided into eight lists.

Using Amazon Mechanical Turk,[12] we recruit 116 people with the requirements of having an approval rating of 99.0% or higher, having at least 5,000 approved tasks, being located in the US, the UK, or Canada, and having passed a qualification task. We make sure that the workers are paid at least $12.0 USD per hour. Among them, we exclude the responses of 46 participants from the analysis because their accuracy for a sanity check is below 80.0%. We analyze the data of the remaining 71 participants.

Experiment 2 Using Amazon Mechanical Turk, we recruit 635 people with the requirements of having an approval rating of 99.0% or higher, having at least 5,000 approved tasks, and being located in the US, the UK, or Canada. We make sure that the workers are paid at least $12.0 USD per hour. Among them, we exclude the responses of 352 participants whose accuracy for the control conditions is less than 90% based on the distributions of accuracy in Figure 8. The control results include results for unembedded, negation, and conditional conditions. The interrogative control condition is not included in the mean calculation, because its mean accuracy is around chance (36.0% over the chance level 33.3%). As a result, we analyze the data of the remaining 283 participants.

[12] https://www.mturk.com

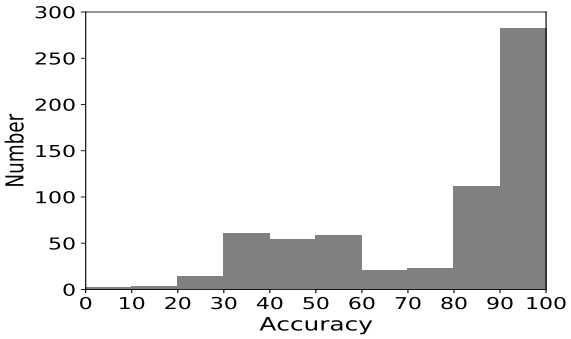

Figure 8: Distributions of accuracy in the control conditions in ProPres.

C Triggers and Environments in IMPPRES

Table 8 and 9 present triggers and environments used in IMPPRES, respectively.

Trigger	Example	Presupposition
All N	All four waiters that bothered Paul telephoned.	Exactly four waiters telephoned.
Both	Both people that hoped to move have married.	Exactly two people have married.
Change of state verb	Marie was leaving.	Marie was here.
Cleft existence	It is Margaret that forgot Dan.	Someone forgot Dan.
Cleft uniqueness	It is Donna who studied.	Exactly one person studied.
Only	The pasta only annoys Roger.	The pasta annoys Roger.
Possessive definites	The boy's rugs did look like these prints.	The boy has rugs.
Possessive uniqueness	Maria's apple that ripened annoys the boy.	Maria has exactly one apple that ripened.
Question	Bob learns how Rachel approaches Melanie.	Rachel approaches Melanie.

Table 8: Examples of triggers in IMPPRES.

Environment	Example
Unembedded	All four waiters that bothered Paul telephoned.
Negation	All four waiters that bothered Paul did not telephone.
Interrogative	Did all four waiters that bothered Paul telephone?
Conditional	If all four waiters that bothered Paul telephoned, it's okay.
Modal	All four waiters that bothered Paul might telephone.

Table 9: Environments used in IMPPRES.

D Results without Exclusion

Figures 9 and 10 present results without exclusion of triggers and environments in IMPPRES and ProPres, respectively.

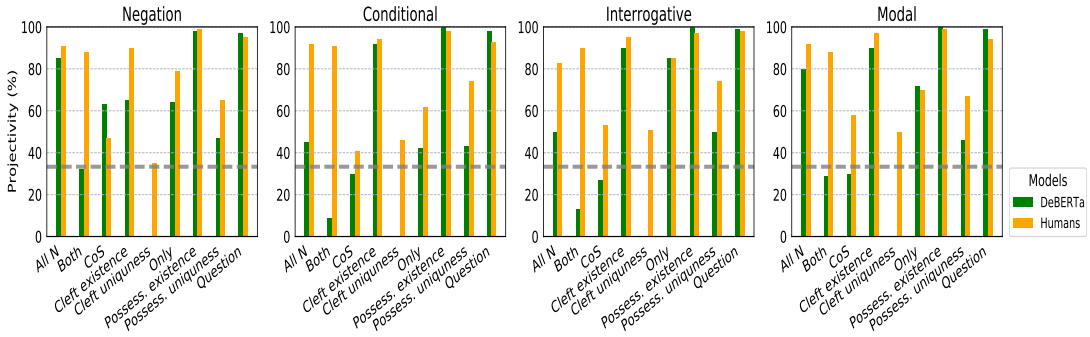

Figure 9: Results on triggers embedded under the negation, conditional, interrogative, and modal in IMPPRES.

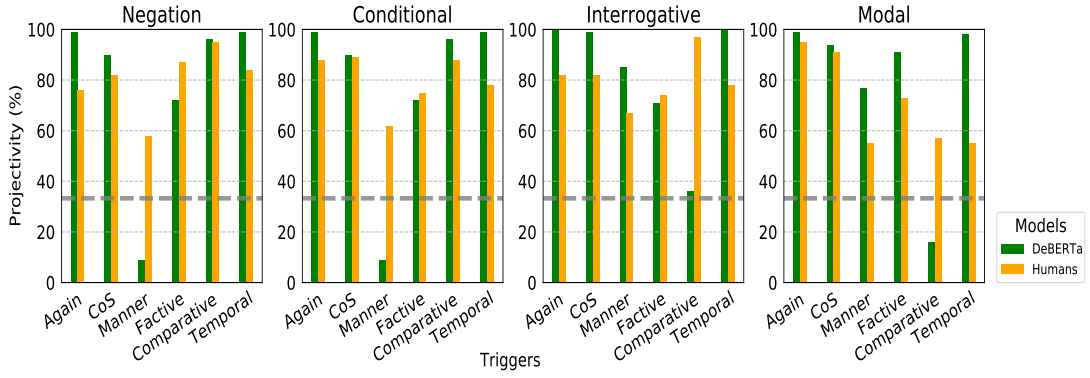

Figure 10: Results on triggers embedded under the negation, conditional, interrogative, and modal in ProPres.