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

Daiki Asami
University of Delaware
daiasami@udel.edu

Saku Sugawara
National Institute of Informatics
saku@nii.ac.jp

Abstract

What makes a presupposition of an utterance—information taken for granted by its speaker—different from other pragmatic inferences such as an entailment is projectivity (e.g., the negative sentence *the boy did not stop shedding tears* presupposes *the boy had shed tears before*). The projectivity may vary depending on the combination of presupposition triggers and environments. However, prior natural language understanding studies fail to take it into account as they either use no human baseline or include only negation as an entailment-canceling environment to evaluate models' performance. The current study attempts to reconcile these issues. We introduce a new dataset, projectivity of presupposition (PROPRES), which includes 12k premise–hypothesis pairs crossing six triggers involving some lexical variety with five environments. Our human evaluation reveals that humans exhibit variable projectivity in some cases. However, the model evaluation shows that the best-performed model, DeBERTa, does not fully capture it. Our findings suggest that probing studies on pragmatic inferences should take extra care of the human judgment variability and the combination of linguistic items.

1 Introduction

It is an open question as to whether language models can learn a human-like pragmatic inference (Pavlick, 2022). A speaker does not always explicitly say everything in an utterance, but a hearer can infer what is implicit in it. One notable case concerns a *presupposition* that refers to information taken for granted by a speaker of an utterance (Stalnaker, 1974; Beaver, 1997). Presuppositions are prevalent in our everyday communication; hence, a comprehensive investigation of whether models can understand them in the same way as humans can contribute to the development of a better language understanding system.

Presupposition triggers introduce presuppositions (e.g., *again* in Figure 1 (a)). A presuppo-
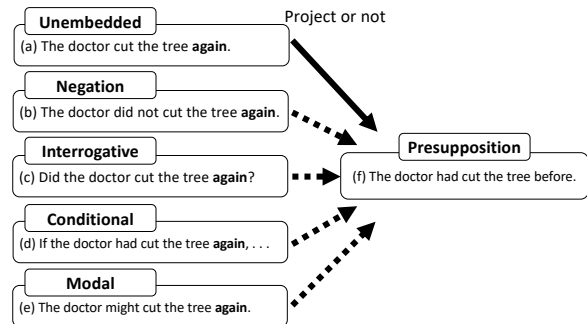

Figure 1: Projectivity of presupposition. A presupposition can project out of entailment-canceling environments. The dashed arrows indicate that the projectivity varies depending on the combination of triggers and environments.

sition of (a) is *the doctor had cut the tree before* (f). What makes the presupposition different from an entailment (in this case, *the doctor cut the tree one more time*) is projectivity: the presupposition projects out of entailment-canceling environments (e.g., negative (b), interrogative (c), conditional (d), and modal (e) sentences) while the entailment does not.[1] In other words, the presupposition (f) holds in the environments (b–e), but the entailment (*the doctor cut the tree one more time*) does not.

Crucially, linguistic studies suggest that the projectivity can vary depending on many factors (Karttunen, 1971; Simons, 2001; Sevegnani et al., 2021; Tonhauser et al., 2018, 2019; Degen and Tonhauser, 2021b). Previous probing studies in natural language processing examine models' performance on presuppositions in the natural language inference (NLI) task (Jeretic et al., 2020; Parrish et al., 2021). However, they do not fully take into account the variable aspect of the projectivity. For instance, Jeretic et al. (2020) obtain no human baseline, which makes models' performance hard to

---

[1] Formal semantic and pragmatic literature generally uses the term, operators, rather than environments but we use the latter for the sake of readability.

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

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

interpret. Parrish et al. (2021) collect human data but use only one entailment-canceling environment, negation. Hence, it remains unclear about the projectivity out of other environments.

This study attempts to reconcile these issues. We first evaluate recent pretrained language models against a presupposition portion of IMPPRES (Jeretic et al., 2020). Specifically, we conduct a human evaluation on its subset (900 pairs), each of which ends up receiving 9.4 labels on average, and then evaluate RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020). We find that humans exhibit relatively weak projectivity in some examples but the best-performed model, DeBERTa, does not perform in a human-like way.

IMPPRES is imperfect in terms of comprehensiveness: the nine triggers that it uses are not exhaustive (cf. Levinson (1983) and Potts (2015) list a total of 27 triggers) and are lexically limited. Thus, using six new triggers with some lexical variety (Table 1) and five environments (Table 2), we construct an extensive evaluation dataset: projectivity of presupposition (PROPRES), which consists 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) Each pair has more than 50 human labels on average. This second evaluation reveals that human data exhibit variable projectivity not only in previously attested cases such as manner adverbs in interrogative and negative environments (Stevens

et al., 2017; Tonhauser et al., 2018) but also in unattested cases such as those in conditional and modal environments. Additionally, we find some within-trigger-type variation. However, the best-performed model, DeBERTa, shows poor performance on controls and does not fully capture the variable projectivity patterns, indicating that it does not learn the pragmatic knowledge necessary to understand presuppositions. These findings suggest that the combination of the various linguistic items in PROPRES and the human evaluation allow us to probe the model's behavior more adequately.

The results from our two evaluations suggest that studies evaluating language understanding systems and creating datasets targeting pragmatic inferences should take extra care of the human judgment variability and the combination of linguistic items. In conclusion, this study makes the following contributions:[2]

- We introduce PROPRES using six novel presupposition triggers embedded under five environments, which enables a comprehensive investigation of the projectivity of presupposition.

- Our human evaluation provides evidence for the variable projectivity depending on the combination of triggers and environments.

- Our model evaluation against human results re-

veals that the models and humans behave differently in the understanding of presuppositions.

2 Background

2.1 Presupposition in Linguistics

Linguistic items or constructions introducing a presupposition are referred to as presupposition triggers (e.g., *again* in Figure 1; Stalnaker, 1974; Beaver, 1997). One property that makes presuppositions distinct from other pragmatic inferences such as an entailment is projectivity: presuppositions survive in entailment-canceling environments such as negation (Karttunen, 1973; Heim, 1983). For instance, a presupposition of the affirmative sentence with the presupposition trigger *again* ((f) given (a)) holds when embedded under negation (b). In contrast, the same environment cancels an entailment (here, *the doctor cut the tree one more time*).

Importantly, previous linguistic studies show that the projectivity of presupposition can vary depending on factors such as context, lexical items, prior beliefs, a speaker's social identity, and prosodic focus (Karttunen, 1971; Simons, 2001; Stevens et al., 2017; Tonhauser et al., 2018, 2019; Degen and Tonhauser, 2021b). This variability is in line with the observation that humans make unsystematic judgments about projectivity on both natural (Ross and Pavlick, 2019; de Marneffe et al., 2019) and controlled (White and Rawlins, 2018) sentences. One remaining question here is whether the variable projectivity has to do with the interaction of triggers and environments (e.g., is a presupposition triggered by *again* more likely to project over the negation (b) than the conditional (d)?). To tackle this question comprehensively, this study collects human judgments on presuppositions using a wide range of triggers and environments.

2.2 Presupposition in NLI

Previous studies introduce NLI datasets to evaluate model performance on presuppositions (Jeretic et al., 2020; Parrish et al., 2021). One example is a template-based dataset: IMPPRES (Jeretic et al., 2020). Using this dataset, Jeretic et al. (2020) conclude that models (e.g., BERT (Devlin et al., 2019)) learn the projectivity of presuppositions triggered by *only*, cleft existence, possessive existence, and question. However, there is one problem with them, that is, no human evaluation. As discussed in Section 2.1, it is possible that projectivity varies depending on the combination of triggers and environments. Thus, it is unknown whether the results of the model evaluation reported by Jeretic et al. (2020) align with human data. To solve this issue, following Parrish et al. (2021), we conduct human evaluation on a subset of IMPPRES as well as our dataset, PROPRES.

Another dataset relevant to our study is NOPE (Parrish et al., 2021), which consists of naturally-occurring sentences with presupposition triggers. With this dataset, Parrish et al. (2021) evaluate transformer-based models against human performance, finding that models behave similarly to humans. One limitation of NOPE is that it includes only negation as an entailment-canceling environment. As a result, the generalizability of the findings by Parrish et al. (2021) is unclear beyond negation. To draw a more general conclusion, 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 no human evaluation, which leaves it open whether models capture any variable projectivity exhibited by humans. To overcome it, we collect human labels on a subset of IMPPRES, testing the performance of the two models, RoBERTa and DeBERTa, against the human results.

3.1 Setup

Human Evaluation Our human evaluation targets a subset of IMPPRES, which uses nine triggers (*all N*, *both*, change of state verbs (CoS), cleft existence, *only*, possessive definites, possessive uniqueness, and question). Specifically, we focus on conditions where triggers occur in one of the five environments (the affirmative sentence (unembedded), negative sentence (negation), conditional antecedent (conditional), modal sentence (modal), and interrogative)[3] and where a hypothesis is either an affirmative or negative sentence. We randomly extract ten items from each condition (a total of 900 sentences).

Using Amazon Mechanical Turk,[4] we conduct

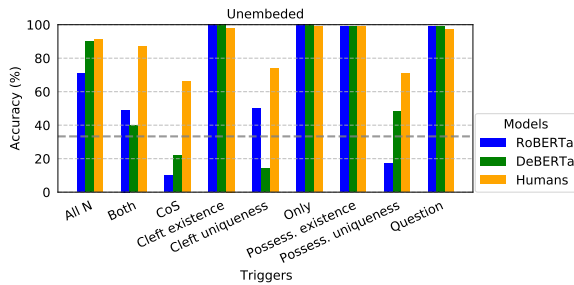[3]Examples of triggers and environments in IMPPRES appear in Appendix D.

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

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

Figure 3: An example prompt in the human evaluation.

the human evaluation run on PCIbex.[5] Figure 3 shows an example prompt that we use in the human evaluation. We adopt and modify the instruction for the human evaluation from Parrish et al. (2021). As a result of the human evaluation, each of the extracted items receives 9.4 labels on average.[6]

Model Evaluation We evaluate Huggingface's (Wolf et al., 2020) pretrained RoBERTa-base (Liu et al., 2019) and DeBERTa-v3-large (He et al., 2020) fine-tuned on MNLI (Williams et al., 2018). We do not evaluate a bag-of-words (BOW) model and an InferSent model (Conneau et al., 2017) because Jeretic et al. (2020) show that their accuracy for control conditions is below chance (33.3%).

3.2 Results and Discussion

Unembedded Triggers We use 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), a presupposition equals

an entailment (e.g., *Bob only ran* presupposes and entails *Bob ran*) (Jeretic et al., 2020). If humans show low accuracy for any unembedded triggers, we manually analyze the relevant triggers to identify their cause. We interpret models' low accuracy as lack of knowledge of relevant triggers if humans show high accuracy for the same triggers.

The results of the human evaluation (Figure 2) show lower accuracy for CoS (66.3%), cleft uniqueness (74.1%), and possessed uniqueness (71.9%), examples of which are provided below, compared to the other triggers (acc. > 87.3%).[7]

(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 reason that the low accuracy for CoS is due to lexical ambiguity. For instance, people might label the pair (1) as neutral or contradiction because Ben was not necessarily exposed before being hidden. Regarding the other two conditions, we do not understand the exact source of the low accuracy at this point. In linguistics, results from human judgment experiments sometimes contradict generalizations made by theoreticians (Gibson and Fedorenko, 2013). Additionally, NLI research reports disagreements in human labels (Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Zhang and de Marneffe, 2021; Jiang and de Marneffe, 2022). Thus, the current results suggest that judgments on presuppositions of cleft and possessive uniqueness are not as robust as Jeretic et al. (2020) might assume. Consequently, we remove CoS, cleft uniqueness, and possessed uniqueness from the following analysis as they might confound the results.

The results of the model evaluation reveal that both RoBERTa and DeBERTa achieve high accuracy for most triggers (acc. > 89.5%). 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*
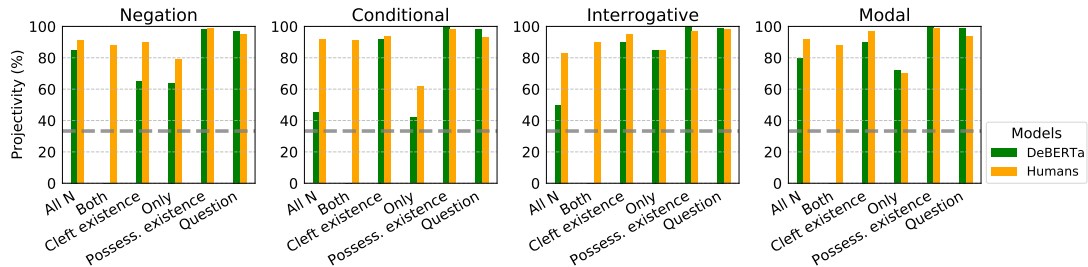
Figure 4: Results on entailment-canceling environments in IMPPRES. DeBERTa's results on *both* are not presented.

guys ran), neither DeBERTa nor RoBERTa performs well (39.0% and 49.0%, respectively). Otherwise, the two models are roughly comparable in performance. Thus, we analyze only DeBERTa.

Based on the human and model results, our analysis of entailment-canceling environments includes the five triggers: *all N*, cleft existence, *only*, possessive existence, and question.[8]

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 gold labels for sentence pairs involving presupposition. We calculate projectivity based on whether presupposition holds when embedded under an entailment-canceling environment. For instance, if one classifies the pair, *did Tom only terrify Ken?* and *Tom terrified Ken*, as entailment, we consider it as projective. Taking another example, if one judges the hypothesis *Tom did not terrify Ken* as contradiction given the same premise, it counts as projective. Otherwise, we take these two examples as non-projective.

Figure 4 presents results on the four environments: negation, conditional, interrogative, and modal. Overall, DeBERTa and humans behave similarly. For instance, they show relatively low projectivity in *only* in conditional (e.g., *if Mary only testifies, ... → Mary testifies*) and modal (e.g., *Mary might only testify → Mary testifies*) (61.8% and 69.8% for humans and 41.5% and 72.0% for DeBERTa, respectively).

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 complained → someone complained*) as projective (89.7%) while DeBERTa predicts it as less projective (65.0%). 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 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 except *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 differences in projectivity.

4   Experiment 2: PROPRES

An investigation of the projectivity of presupposition with IMPPRES is far from comprehensive because we can find more triggers in the literature (e.g., 27 triggers in Levinson (1983) and Potts (2015) in total) and none of the six triggers which we analyze in IMPPRES has lexical variation. Using six additional triggers with some lexical variety, we create a new dataset, PROPRES, which allows us to investigate the variable projectivity and models' behavior more comprehensively.

4.1   Data Generation

Triggers and Environments   PROPRES has six types of presupposition triggers: (1) the 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. Crucially, these triggers allow us to use different lexical items (e.g.,

---

[8]We report all results including excluded triggers in Appendix E.

126

we use seven verbs and nine adverbs for aspectual verbs and manner adverbs, respectively). One exception is *again*, but it is a standard presupposition trigger investigated by theoretical linguistic (von Stechow, 1995; Bale, 2007) and natural language processing (Cianflone et al., 2018) research. Thus, it is worth including this trigger in the dataset.

PROPRES uses 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. We include the unembedded environment to test whether humans and models can identify presupposition as entailment when triggers occur in affirmative sentences. The counterfactual conditional antecedent is not a typical entailment-canceling environment, but we include it to ensure that conditional controls have clear gold labels (entailment or contradiction) as we discuss in the following paragraph. We generate affirmative and negative hypotheses for each premise sentence. 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.[9]

We make a control condition corresponding to each target condition where a hypothesis is either an affirmative or negative version of its premise, as shown in Table 2. The control conditions serve as a sanity check in a human evaluation. They are also important to test whether the models rely on lexical overlap (McCoy et al., 2019) or negation (Gururangan et al., 2018) heuristics. For instance, models should 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 with *not* as contradiction if they use the negation heuristic. Only if models predict correctly in the control conditions, we can say that their predictions about the corresponding target conditions indicate projectivity rather than heuristics. Creating 100 pairs for each control condition results in 6,000 pairs. In total, PROPRES comprises 12,000 sentence pairs.

Templates We make templates and generate sentences with them using the codebase developed by Yanaka and Mineshima (2021).[10] Following are examples of templates and sentences.[11]

(4) The N did not VP again.
     (The girl did not hurt others again.)
     → (↛) 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 check whether models rely on the lexical overlap heuristic in the control conditions.

The use of templates has three advantages. First, it allows us to systematically test whether models rely on the lexical overlap (McCoy et al., 2019) and negation (Gururangan et al., 2018) heuristics. In addition, it enables us to conduct a targeted evaluation with a large number of sentences including presupposition triggers embedded under particular environments. Preparing the same number of data might be impossible if we use corpora. Finally, we can rule out the effect of plausibility. Previous linguistic 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 might stem from our world knowledge: it is more plausible for one to go to a restaurant than the moon. As the plausibility effect is not the focus of this study, we use templates to control it.

4.2 Setup

Human Evaluation We randomly select ten out of 100 pairs from each target condition and two pairs from each control condition, extracting 600 and 120 pairs in total, respectively. The human evaluation procedure is identical to the one reported in Section 3.1: using Amazon Mechanical Turk, we conduct the evaluation run on PCIbex. As a result, each of the extracted pairs has 56.7 labels on average. Due to some revision of PROPRES during the dataset creation, we collect judgments on the

---

[9]We provide examples for each condition in Appendix B.

[10]https://github.com/verypluming/JaNLI
[11]A full list of the templates and their example sentences appears in Appendix B.
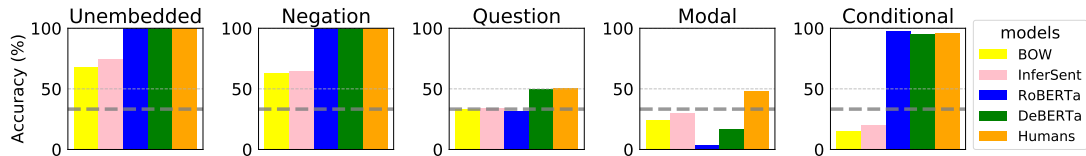
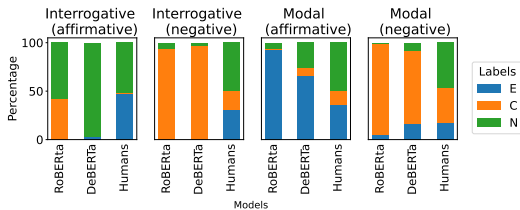Figure 5: Results on control conditions in PROPRES.

Figure 6: Distributions of labels in the interrogative and modal with an affirmative or negative hypothesis.

modal environment and comparative trigger in Experiment 1 (200 pairs in total). As a consequence, they receive 9.4 labels 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[12] 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 5 shows results on control conditions in which a hypothesis is either an affirmative or negative version of its premise. The performance of InferSent and BOW models is poor, which makes their performance on target conditions hard to analyze. Thus, we exclude them from our analysis below. Similar to humans, RoBERTa and DeBERTa perform well on the unembedded, negation, and conditional (e.g., P_1–P_3 in (5)), indicating that they do not rely on 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?

---

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

P_5: The boy might cut the tree again.
$H_{1(2)}$: The boy (did not) cut the tree again.

RoBERTa, DeBERTa, and humans perform poorly on the interrogative and modal (e.g., P_4 and P_5 in (5)) in which the correct label is supposed to be neutral (Jeretic et al., 2020) (31.8%, 50.0%, and 51.1% for interrogative and 3.5%, 16.7%, and 48.1% for modal, respectively). Distributions of labels in these conditions (Figure 6) show that the majority of labels in humans are neutral, which is consistent with the view that a yes/no question does not have a truth value and thus one cannot decide whether its affirmative or negative version is true or not (Groenendijk and Stokhof, 1984; Roberts, 2012). One exception is the interrogative with an affirmative hypothesis (e.g., P_4 and H_1 in (5)): distributions of entailment and neutral are comparable (46.5% and 52.4%, respectively). We suspect that some people interpret this condition as a confirmation question in which the affirmative counterpart of the interrogative (in this case, H_1) is presupposed, resulting in a 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 (e.g., P_4 and H_2), RoBERTa and DeBERTa assign contradiction to the hypothesis the majority of the time (93.7% and 97.1%, respectively), indicating the negation heuristic: models are likely to label a given hypothesis as contradiction if it includes *not* (Gururangan et al., 2018).

The two models do not mirror humans in performance on the modal, either. Their majority labels in the modal with affirmative and negative hypotheses (e.g., P_5 with H_1 and H_2) are entailment and contradiction, respectively. These results suggest that in the modal, they rely on the lexical overlap heuristic if a hypothesis is affirmative but they adopt a negation heuristic if it is negative, overriding the lexical overlap heuristic. Specifically, they label a
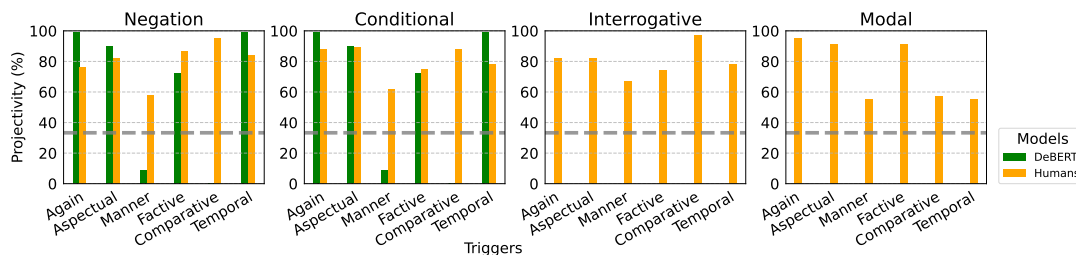
Figure 7: Results on entailment-canceling environments in PROPRES. DeBERTa's results on the interrogative and modal environments and the comparative trigger are not shown due to its unstable performance on their control counterparts.
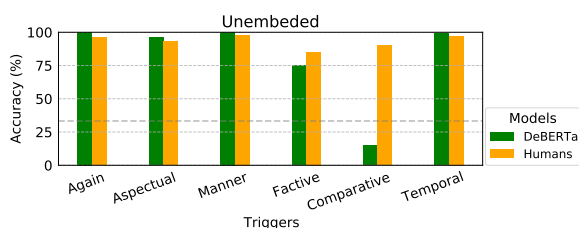

Figure 8: Results on the unembedded condition in PRO-PRES for DeBERTa and humans.

hypothesis as entailment if it is affirmative whereas if *not* is present in it, they label it as contradiction.

These variable results for DeBERTa and RoBERTa are inconsistent with Jeretic et al. (2020), who find that BERT achieves high accuracy for the interrogative and modal controls by correctly assigning them the neutral label. The discrepancy between our results and Jeretic et al. (2020)'s indicates that the combination of the two environments with new triggers in PROPRES makes a more thorough model evaluation possible.

Overall, the performance of RoBERTa and De-BERTa is interpretable regarding the three environments: unembedded, negation, and conditional; hence, we omit model results on the interrogative and modal below.[13] Additionally, since the two models are comparable in accuracy, we only report DeBERTa's performance in what follows.

Unembedded Triggers   Figure 8 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 limited knowledge of this trigger. Hence, we exclude DeBERTa's predictions about the comparative when we report results on

[13]We report all results including excluded conditions in Appendix E.

entailment-canceling environments.

Entailment-Canceling Environments   Figure 7 shows results on the entailment-canceling environments. Our human results provide evidence for variable projectivity (range 55.1–99.8%).

First, the human results indicate that the iterative *again* weakly projects over the negation (75.8%) compared to the other three environments (86.3% on average). We provide the example sentence pairs for *again* embedded under negation below.

(6) P: The man did not shed tears again.
       $H_{1(2)}$: The man had (not) shed tears before.

We reason that this apparent low projectivity is attributable to the fact that the negative sentence with *again* is ambiguous as to whether *again* takes scope over the proposition with negation or without negation (Bale, 2007). In the first reading, the presupposition is that the man had shed tears before; in the second reading, it is that the man had not shed tears before. If humans infer the second presupposition, they should label the hypotheses such as H_1 and H_2 as entailment and contradiction, respectively, giving rise to the seemingly low projectivity rates. Since this ambiguity itself has nothing to do with the projectivity, we leave it open whether the observed rate (75.8%) truly reflects the projectivity or not. Contrary to humans, the DeBERTa judges the same condition as projective (95%), indicating that it virtually always predicts the second presupposition (e.g., the man had shed tears before).

Next, manner adverbs exhibit relatively weak projectivity over the negation (e.g., P_1 in (7)) and interrogative (e.g., P_2) (58.3% and 66.6%, respectively).

(7) P_1: The man did not hurt others seriously.
       P_2: Did the man hurt others seriously?

129

P_3: If the man had hurt others seriously, ...
P_4: The man might hurt others seriously.
$H_{1(2)}$: The man (did not) hurt others.

According to Stevens et al. (2017) and Tonhauser et al. (2019), a focalized element in the utterance affects the projectivity of the presupposition introduced by manner adverbs in interrogatives and negation. For instance, the presupposition (H_1) is more likely to project when the focus falls into the manner adverb (*did the man hurt others SERIOUSLY?*) than when it falls into the subject (*did the MAN hurt others seriously?*). Since our human evaluation provides no prosodic information signaling focus, humans might find these conditions ambiguous, yielding weak projectivity. Furthermore, our item-by-item analysis with human data reveals that in the manner adverbs embedded under negation, the projectivity ranges between 43.3% (for *angrily*) and 66.6% (for *easily*), indicating the within-trigger-type variability.

Adding to Stevens et al. (2017) and Tonhauser et al. (2019), we find that the manner adverbs are weakly projective in the conditional (e.g., P_3) and modal (e.g., P_4) (62.0% and 55.1%, respectively). This suggests that information structural cues such as prosodic focus play a role in the projectivity of presupposition introduced by the manner adverbs embedded under the conditional and modal.

Third, in the modal, temporal adverbs (e.g., P_1 in (8)) and comparatives (e.g., P_2) have weaker projectivity (54.7% and 57.4%, respectively) than the other three triggers excluding the manner adverbs (92.5% on average). These two triggers are projective in the other three environments (79.7% and 93.4% on average for the temporal adverbs and comparatives, respectively). This indicates that the projectivity of presuppositions of these triggers varies depending on the environment.

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

DeBERTa's performance does not mirror humans' in some cases. It predicts that the manner adverbs in the negation and conditional (P_1 and P_3 in (7), 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?*).

In summary, the human evaluation in Experiment 2 shows variable projectivity in six out of the 24 new conditions, contrary to the first one, in which we observe it in two out of 24 conditions. This contrast highlights that the combination of various triggers and environments can lead to the discovery of new cases of variable projectivity. In addition, we find that DeBERTa does not capture variable projectivity in some cases, suggesting that DeBERTa's ability to process presupposition is not necessarily human-like.

5   Conclusion

Our experiments reveal that humans exhibit the variable projectivity of presupposition in some conditions (two out of 24 and six 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 generalize pragmatic inferences for presuppositions.

In our experiments, quite a few conditions are excluded from the analysis for various reasons such as lexical ambiguity in some items, disagreements in human judgments, and the models' lack of knowledge. To tease apart these factors carries us well beyond the scope of this study. However, this fact suggests that we need to be careful with dataset creation so that we can train or evaluate models on well-designed datasets targeting pragmatic inferences.

References

Alan Clinton Bale. 2007. Quantifiers and verb phrases: An exploration of propositional complexity. *Natural Language & Linguistic Theory*, 25:447–483.

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.

Andre Cianflone, Yulan Feng, Jad Kabbara, and Jackie Chi Kit Cheung. 2018. Let's do it "again": A first computational approach to detecting adverbial presupposition triggers. In *Proceedings of the 56th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 2747–2755, Melbourne, Australia. 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.

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.

Jeroen Groenendijk and Martin Stokhof. 1984. *Studies on the semantics of questions and the pragmatics of answers*. Ph.D. thesis, University of Amsterdam dissertation.

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.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

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.

Stephen 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.

Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, 5:6–1.

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.

Arnim von Stechow. 1995. Lexical decomposition in syntax. In Egli Urs, E. Pause Peter, Schwarze Christoph, von Stechow Arnim, and Wienold Götz,

editors, *The Lexicon in the Organization of Language*, pages 81–118. John Benjamins, Amsterdam/Philedelphia.

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   Limitations

One of the limitations of our study is that not all data have human labels. However, it is not feasible to get many judgments for all the data in IMPPRES and PROPRES in terms of cost. Extending this study, we hope to conduct a targeted human evaluation with some of the triggers that exhibit the variable projectivity (e.g., manner adverbs).

The second limitation has to do with humans' low accuracy in control modal and question conditions. We attribute this to the procedure of our

evaluation. The participants are asked to judge whether the hypothesis contradicts, entails, or is neutral to the question or modal premise. Since it is hard to imagine the situation in which the modal and question sentences are true or false, people might be confused with the instruction. We hope to collect more valid data using a better instruction in our future study.

The third limitation is that we do not conduct the thorough analyses of between-item variability and between-participant variability in data from the two human evaluations. It is likely that the projectivity of the presupposition depends on lexical items and participants. We take these into consideration in the future study.

The final limitation is that this study investigates presuppositions without any context. Taking *John did not stop cutting trees* as an example, whether the presupposition *John had cut trees before* projects over negation depends on a context. For instance, the presupposition does not project over negation if we associate the sentence with the appropriate context. Consider the following example: *Mary liked cutting trees but never smoked. In contrast, John never cut trees but liked smoking. One day Mary and John stopped cutting trees and smoking, respectively. Later Bob said to Ken "John stopped cutting trees." Then Ken responded "wait, John didn't stop cutting trees but he stopped smoking"*. In this example, the sentence *John did not stop cutting trees* does not presuppose *John had cut trees before*. It remains to be seen how the contextual information affects each trigger embedded under different environments.

B Templates

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

C Crowdsourcing Human Evaluation

Before the experiment, each participant is asked to read a written instruction about the NLI task carefully. All data are collected anonymously except workers' ID.

Evaluation 1 Using Amazon Mechanical Turk, 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
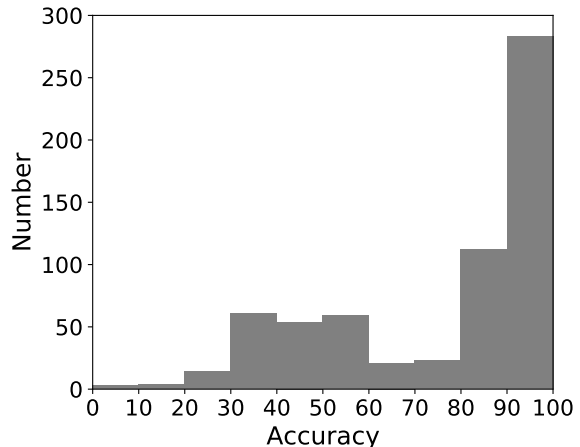

Figure 9: Distributions of accuracy in the control conditions in PROPRES.

paid at least $12.0 USD per hour. Among them, we exclude the responses of 46 participants from the analysis because their accuracy rates for a sanity check are below 80.0%. We analyze the data of the remaining 71 participants.

Evaluation 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 9. 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.

D Triggers and Environments in IMPPRES

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

E Results without Exclusion

Figures 10 and 11 present results without exclusion of triggers and environments in IMPPRES and PROPRES, respectively.

133

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 man might shed tears before spreading the rumor. H_1: The man spread the rumor. H_2: The man 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 sentences.

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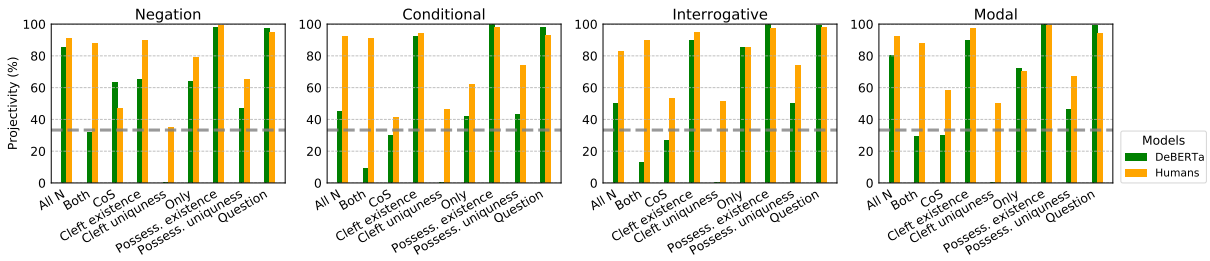
Figure 10: Results on triggers embedded under the negation, conditional, interrogative, and modal in IMPPRES.
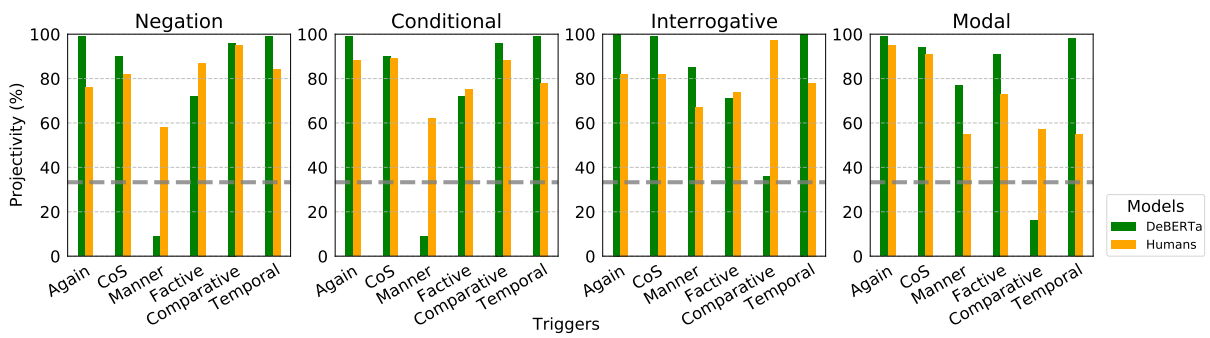

Figure 11: Results on triggers embedded under the negation, conditional, interrogative, and modal in PROPRES.