NOPE: A Corpus of Naturally-Occurring Presuppositions in English

Alicia Parrish* **Sebastian Schuster*** **Alex Warstadt***
Omar Agha **Soo-Hwan Lee** **Zhuoye Zhao** **Samuel R. Bowman** **Tal Linzen**
Department of Linguistics & Center for Data Science, New York University
{alicia.v.parrish,schuster,warstadt}@nyu.edu

Abstract

Understanding language requires grasping not only the overtly stated content, but also making inferences about things that were left unsaid. These inferences include *presuppositions*, a phenomenon by which a listener learns about new information through reasoning about what a speaker takes as given. Presuppositions require complex understanding of the lexical and syntactic properties that trigger them as well as the broader conversational context. In this work, we introduce the Naturally-Occurring Presuppositions in English (NOPE) Corpus to investigate the context-sensitivity of 10 different types of presupposition triggers and to evaluate machine learning models' ability to predict human inferences. We find that most of the triggers we investigate exhibit moderate variability. We further find that transformer-based models draw correct inferences in simple cases involving presuppositions, but they fail to capture the minority of exceptional cases in which human judgments reveal complex interactions between context and triggers.

1 Introduction

In every statement, certain facts are taken for granted by the speaker; such facts, while left unsaid, can often be inferred by the listener. A speaker who says *Chet finished law school*, for example, asserts that Chet reached the end of law school, and *presupposes* that he attended law school in the first place. While such presuppositions are usually not the main information that a speaker intends to convey, listeners can still learn new information from them by drawing inferences about what the speaker takes as a given.

One type of signal that helps human listeners draw such inferences is presupposition *triggers*, such as the verb *finish*. Triggers have long been known in linguistics to be associated with presuppositions, but the nature of this relationship is a matter of much debate (e.g., Heim, 1983; van der Sandt, 1992; Chemla, 2009; Abusch, 2010; Simons et al., 2010; Romoli, 2015; Abrusán, 2011; Schlenker, 2021). This is in part because presuppositions and triggers exhibit properties (summarized in Figure 1) that make the resulting inferences heterogeneous and make precise theory development challenging. How humans draw inferences involving a wide variety of triggers is therefore still poorly understood, and this lack of a comprehensive theory makes it difficult to construct test cases for machine learning models with as much diversity as found in naturalistic data.

We introduce a broad-coverage dataset of sentences with naturally-occurring presupposition triggers, called NOPE (**N**aturally-**O**ccurring **P**resuppositions in **E**nglish). We show that this corpus can be used to analyze the variability associated with different presupposition triggers, and we

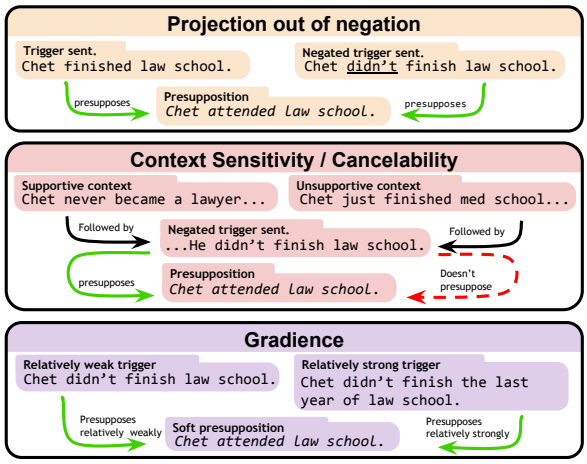

Figure 1: Inferences involving presuppositions can be challenging because presuppositions project out of negation (first box), they are context-sensitive and can be canceled (second box), and they can differ in how easily they can be cancelled and how likely a listener is to make the inference (third box).

* Equal contribution.

use this dataset to investigate how well current natural language understanding models (specifically natural language inference, or NLI, models; Dagan et al., 2006) are able to make similar inferences.

We find considerable variability in human ratings across items for two types of triggers, clause-embedding predicates (e.g, *know* and *think*) and implicatives (e.g., *manage to* and *fail to*), in line with previous work (de Marneffe et al., 2019; Ross and Pavlick, 2019), and we find that the other triggers exhibit lower contextual variability. Further, transformer-based models fine-tuned on NLI are largely able to draw correct inferences for simple cases involving presupposition triggers but they fail to fully capture the contextual variability and gradience in the human judgments. We release our dataset and code at https://github.com/nyu-mll/nope.

2 Background

Properties of presuppositions One important property of presuppositions is that they *project* out of environments like negation (Karttunen, 1973; Heim, 1983). While negating a sentence cancels its entailments, as the top box of Figure 1 illustrates, both the "trigger sentence" and its negation give rise to the same presupposition. Second, presuppositions, unlike entailments, can be canceled and disappear in certain contexts (Simons, 2001; Abusch, 2002; Simons et al., 2010).

Third, presupposition triggers exhibit gradience both in the sense that a listener may be more or less confident about inferring what a speaker presupposes upon hearing different triggers (Tonhauser et al., 2018; Tonhauser and Degen, 2020; Degen and Tonhauser, 2021; Mahler, 2020), and in the sense that the presuppositions of different triggers may be more or less easy to cancel (Abusch, 2002; de Marneffe et al., 2019). Specifically, some triggers are considered *hard* triggers, which require the presupposition to be satisfied for the statement to be well-formed. For example, *There are three apples and both of them are green* is not well-formed because it involves *both*, a hard trigger that always presupposes that there are two objects. *Soft* triggers such as the change of state verb *finished*, on the other hand, can easily be cancelled, as illustrated in the second box in Figure 1.

Existing resources Recent work in NLP has investigated the ability of neural networks trained for natural language understanding to pick up on subtle discourse cues (e.g, Upadhye et al., 2020; Schuster et al., 2020). Among them, some have introduced resources specifically targeting presuppositions (White et al., 2017; de Marneffe et al., 2019; Jeretič et al., 2020) and tested neural models on them (White et al., 2018; Jiang and de Marneffe, 2019; Jeretič et al., 2020; Ross and Pavlick, 2019). However, this previous work has either focused on a specific class of presupposition triggers (e.g., clause-embedding verbs or implicatives), or has made the simplifying assumption that context does not affect presuppositions and therefore did not include naturally-occurring contexts. Our NOPE corpus complements existing corpora by including a wider range of trigger types.

Kim et al. (2021) recently also demonstrated the importance of detecting and verifying presuppositions for natural language understanding tasks such as question answering. They found that verifying presupposed content in questions results in question-answering systems that generate more helpful responses to unanswerable questions.

3 Dataset Construction

3.1 Trigger selection

We identify 10 presupposition trigger types to focus on based mainly on Levinson's (1983) widely taught list (see Beaver, 1997; Potts, 2015, for other similar lists). These specific triggers were selected because they are common in English and systematic enough that we can extract them from a corpus. Table 1 contains a list of all trigger types in the present study, along with an example from the full dataset. Appendix A includes a more detailed discussion of each trigger type and the presuppositions they generally give rise to. In this work, we focus mainly on soft triggers since we expect them to exhibit more context-sensitivity. As a control for the judgment paradigm, we also include the hard triggers clefts and numeric determiners, for which we expect humans to endorse the purported presupposition highly and consistently.

3.2 Extraction of Examples

We extracted sentences with presupposition triggers from the Corpus of Contemporary American English (COCA, Davies, 2008). Following de Marneffe et al. (2019), in addition to the trigger-containing sentence, we also extracted the two immediately preceding sentences for context and, where applicable, the speaker of each sentence. Be-

Trigger	Affirmative Example	Negative Example	Presupposition
Change of state	A microsecond later, images from his exterior sensors snapped into focus.	A microsecond later, images from his exterior sensors **didn't** snap into focus.	Previously, images from his exterior sensors hadn't been in focus.
Clefts	But it is the horse racing that is just for children.	But it **isn't** the horse racing that is just for children.	There's something that is just for children
Comparatives	That is a bigger problem, than the chairman's claim.	That **isn't** a bigger problem, than the chairman's claim.	The chairman's claim is a problem.
Aspectual verbs	At the age of 55, I began preparing myself to die.	At the age of 55, I **didn't** begin preparing myself to die.	Before age 55, I was not yet preparing to die.
Embedded questions	I fail to see how you can rationalize rewarding illegality.	I **don't** fail to see how you can rationalize rewarding illegality.	You can rationalize rewarding illegality.
Clause-embed. verbs	In 20 years we'll realize that's a mistake.	In 20 years we **won't** realize that's a mistake.	[Pushing people towards pharmaceuticals] is a mistake.
Implicatives	The survivors managed to scramble out through the tiny gap in the rocks.	The survivors **didn't** manage to scramble out through the tiny gap in the rocks.	The survivors made an attempt to scramble out through the tiny gap in the rocks.
Numeric determiners	Both protagonists in the room defy a political force and receive aid from a higher authority.	Both protagonists in the room **do not** defy a political force and receive aid from a higher authority.	There are two protagonists in the room.
"Re-" prefixed verbs	Taoism reconnects aging to the great cycles of nature.	Taoism **doesn't** reconnect aging to the great cycles of nature.	Aging was once connected to the great cycles of nature.
Temporal adverbs	He took them to the NL Championship Series last year before being swept by the Atlanta Braves.	He **didn't** take them to the NL Championship Series last year before being swept by the Atlanta Braves.	Johnson was swept by the Atlanta Braves.

Table 1: Selected examples present in NOPE. Presupposition triggers are underlined.

cause some spans of text in COCA are redacted, we only extracted contexts that form a contiguous span with the trigger-containing sentence. We also detokenized text and removed HTML tags.

We identified trigger-containing sentences in two stages. We first automatically extracted sentences using syntactic features extracted from SpaCy dependency parses (Honnibal et al., 2020) or using lists of lexical items. For example, sentences with clefts such as *It is the president who has to sign the document* can be identified by their syntactic structure whereas open-class lexical triggers such as change of state predicates (e.g., *melt*) can only be identified using word lists (see Appendix B for the word lists used). Second, six of the authors, all of whom have graduate level training in formal linguistics, manually reviewed and annotated the extracted passages. All expert annotations were double checked by a native English speaker. We ensured both that the passage contains a genuine example of the trigger[1] and that the trigger sentence could be negated.[2] We also excluded examples with conditionals and examples in which the presupposition trigger was embedded under another predicate, since these examples could not be straightforwardly negated.

In some cases (12.9% of examples), we wrote an altered version of the trigger sentence by making small edits in order to make the subsequent annotations possible. For example, many sentences can be negated after removing an adverb: While the original *I know what Elissa's hobby is already* becomes much less natural when negated due to the presence of *already*, the slightly altered sentence without the adverb, *I know what Elissa's hobby is*, can easily be negated. We further systematically altered sentences for two types of triggers. For temporal adverbs, we rewrote sentences where the adverbial clause is preposed, in order to allow negation to scope over the adverbial (e.g. *Before it rained, we danced* becomes *We danced before it rained*). For numeric determiners, we added explicit domain restrictions in order to reduce vagueness in the spelled-out presupposition (e.g. *Both cats meowed* becomes *Both cats on the mat meowed*, with presupposition *There are two cats on the mat*).

In total, we extracted 2,482 passages, of which

[1] For example, we excluded from the dataset sentences with a referential *it* as in *It is a multi-purpose bread that should be part of any culinary repertoire*, which resemble clefts, and were occasionally labeled as such by the extraction pipeline.

[2] In some cases, negating the clause with the presupposition trigger led to contradictions or pragmatically odd sentences. For example, *The dog started digging and uncovered the bone* contains the trigger *start* and negating *start* would lead to the non-sensical sentence *The dog didn't start digging and uncovered the bone.*

1,279 (51.5%) met our criteria for inclusion. This resulted in more than 100 examples per trigger type for subsequent ratings.

Flipping the polarity of the trigger sentence To investigate to what extent the presupposition projects out of negation, we manually constructed a negated version of the trigger sentence for each example. We added sentential negation (e.g. *not*) to the main clause of the sentence or, in the case of clausal coordination, to the conjunct that contains the trigger. For the 102 corpus-extracted examples that already contained sentential negation, we removed the negation to create a non-negated version of the sentence.

Writing presuppositions To test to what extent humans and NLI models infer the content of the presupposition for a given trigger sentence, it is necessary to spell out the content of the purported presupposition. We therefore wrote a sentence expressing the presupposition for each pair of non-negated and negated trigger sentences. For instance, for sentences with change of state predicates and aspectual verbs, we wrote sentences that refer to the state before the event described by the trigger sentence (e.g., *Bill dropped the vase* presupposes *Bill had been holding the vase just before then*). See Appendix C for a description of our writing strategies for each trigger type. All sentences were checked by a second annotator and corrected if the second annotator discovered issues with the original phrasing.

3.3 Adversarial Examples

A potential bias in the dataset is that the majority of examples give rise to presuppositions and therefore both the non-negated and the negated trigger sentences entail the sentence spelling out the presupposition most of the time. This creates an issue for model evaluation, since high accuracy in predicting the entailment relation between the trigger sentences and the presupposition sentence may be caused by some heuristic that leads to frequent predictions of ENTAILMENT. This issue is further exacerbated by the fact that for many trigger types (e.g. clause-embedding verbs), the presupposition sentence tends to have very high lexical overlap with the trigger sentence and therefore a model that uses a heuristic to predict ENTAILMENT when lexical overlap between the trigger sentence and the presupposition sentence is high (which has been found to be the case for many models, e.g., McCoy et al., 2019), will likely achieve high accuracy despite not being able to draw the correct inferences.

To control for this issue, we randomly selected 200 examples (20 per trigger type) from the main corpus and constructed adversarial sentences that differ minimally from the presupposition sentences but are no longer entailed.[3] This adversarial dataset in combination with the main dataset thus constitutes a set of minimal pairs as advocated for by Gardner et al. (2020) and makes it possible to tease apart whether high accuracy on the main dataset is a result of a lexical overlap heuristic or more sophisticated inferences.

3.4 Crowdsourced Probability Judgments

We determined whether and how strongly naive participants infer the content of the expert-written presupposition after reading the passage containing the trigger for each example by crowdsourcing probability ratings. These ratings allow us to evaluate a) for which triggers the content of the purported presupposition projects out of negation, b) the context-sensitivity of a certain trigger type, and c) the level of gradience in inferences for different trigger types. Further, as we demonstrate in Section 5.1, we can use the ratings to evaluate how well NLI models mimic human inferences.

Task description For each example, we collected 5 probability judgments from participants via Amazon Mechanical Turk (MTurk). Each task consisted of 20-30 pairs of passages with presupposition triggers (the *premise*) and the sentence capturing the content of the purported presupposition (the *hypothesis*), presented one-by-one in randomized order. We further included 5 filler premise-hypothesis pairs from MNLI (Williams et al., 2018) as attention checks (see Appendix D for exclusion criteria). For each item, participants were presented with the following instructions: "Indicate how likely you think the statement is to be true, using the information in the text and your background knowledge about how the world works." Participants provided ratings from 0.0 to 100.0 by adjusting a non-linear slider that allowed greater precision at the slider's edges, capturing that the distinction between 99.0 and 99.5% probability is

[3] For example, for the trigger sentence *Women from both sides of town formed a mothers group*, we wrote the presupposition sentence *There are two sides of town*, which is entailed by the trigger sentence. In the adversarial corpus, we changed the presupposition sentence to *There are **three** sides of town*, which is no longer entailed by the trigger sentence.

more informative than the distinction between 60.0 and 60.5% probability (Tversky and Kahneman, 1981). Slider endpoints were labeled "impossible" (0) and "certain" (100).

We used a pre-screening task to find reliable participants (see Appendix D). Each task was advertised to qualified MTurk workers as taking 10 minutes to complete and paid USD 2.50 (USD 15/hr). Median completion time across all posted tasks was 9.6 minutes. Individual participants could provide a rating for at most 15% of the total dataset.

4 Analysis of Human Judgments

To investigate to what extent the different trigger types give rise to presuppositions, and to what extent they exhibit context-sensitivity and gradience, we analyzed the crowdsourced human judgements as described in the following sections.

4.1 Mapping to NLI Labels

To facilitate our analyses (and to allow us to use the widely-used NLI paradigm for evaluating models), we mapped the continuous ratings for each example to the three NLI labels CONTRADICTION, NEUTRAL, ENTAILMENT by inferring upper and lower thresholds for CONTRADICTION and ENTAILMENT, respectively. We determined these thresholds based on the participants' ratings of the filler items for which we have expected categorical judgments. Since different participants may use the scale differently, we computed individual thresholds for each participant (average for CONTRADICTION: 5.4%, average for ENTAILMENT: 93.5%; see Appendix E for details on the threshold inference procedure). We then assigned a label to each example based on the majority vote of the five participants. Following standard practice for NLI datasets (e.g., Bowman et al., 2015), we discarded any examples without a majority label (see Table 3).

4.2 Human Judgment Results

Figure 2 shows the distribution of labels for each type of trigger in the main corpus, for non-negated and negated premises. In cases in which the premise presupposes the hypothesis, we expect the label for the non-negated and the negated example to be ENTAILMENT (the entailment-canceling negation should not affect the presupposition); in cases in which the hypothesis is not presupposed, we expect that it is either not entailed by the negated premise or not entailed by both types of premises.

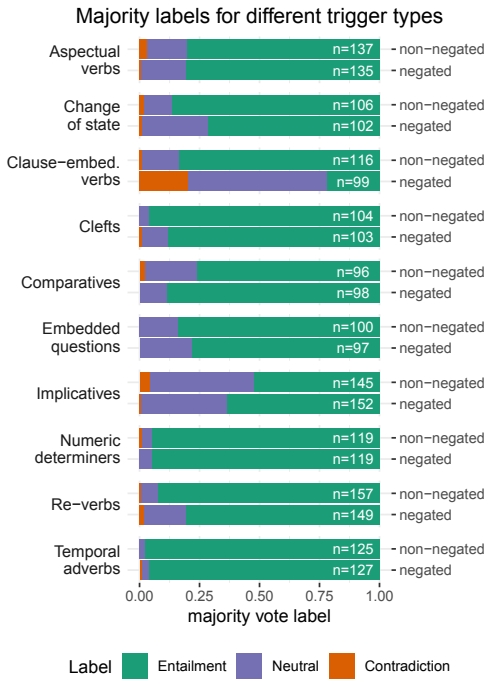

Figure 2: Distributions of mapped human NLI labels for examples in the main corpus, aggregated by trigger type. n indicates the number of examples.

Across all triggers, most examples were labeled as ENTAILMENT, with the remaining examples largely labeled as NEUTRAL.[4] This pattern suggests that participants judged most of our trigger sentences as presupposing the hypothesis, and their judgments were invariant to the negation of the premise. One exception to this finding was clause-embedding verbs: While the non-negated premises (e.g., *Airline officials, flight attendants and frequent travelers **say** Brown's case is hardly unique*) were often judged as entailing the hypothesis (underlined), this is no longer the case for the negated premises (e.g., *Airline officials, flight attendants and frequent travelers **don't say** Brown's case is hardly unique*). This result is expected, for two reasons. First, our clause-embedding verbs included non-factive verbs: although we generally do not expect a clause embedded under *think* or *say* to be presupposed, we included these verbs as there is no clear-cut distinction between factive and non-factive verbs (Tonhauser and Degen, 2020). Second, there can be significant variability across contexts, even for verbs like *know* that are commonly assumed to be factive (de Marneffe et al., 2019).

The second exception to the general pattern was

[4]Note that this pattern cannot be attributed to a response bias of generally assigning higher probabilities since we found that participants used the entire scale on the filler items.

Trigger	No change	E → {N,C}	{N,C} → E	Other
Aspectual verbs	74.8	11.5	13.0	0.8
Change of state	73.0	21.0	6.0	0.0
Clause-embed. verbs	31.3	62.7	3.0	3.0
Clefts	88.3	9.7	1.9	0.0
Comparatives	76.1	5.4	17.4	1.1
Embedded questions	82.1	12.6	5.3	0.0
Implicatives	70.0	7.1	21.4	1.4
Numeric determiners	95.0	2.5	1.7	0.8
Re-verbs	78.6	16.6	4.8	0.0
Temporal adverbs	96.0	2.4	1.6	0.0

Table 2: Percent of times the label assigned via majority vote of participants to a given example changed when the premise was negated. E=ENTAILMENT, N=NEUTRAL, and C=CONTRADICTION.

implicatives, which have a higher proportion of NEUTRAL labels than other triggers. This difference reflects the fact that, unlike what is generally assumed in the implicatives literature (e.g., Karttunen, 1971), statements such as *X took effort* which are supposedly always triggered by *managed to X* did not always receive full endorsement by participants and were therefore often mapped to neutral.

Context-sensitivity The high rates of ENTAILMENT labels for both non-negated and negated examples confirm the standard view that presuppositions project out of negation. More surprising are cases where one or both hypotheses are not entailed. In one sub-case, a presupposition of the non-negated sentence is canceled by negation. We observe this when a non-negated sentence is labeled ENTAILMENT, and the negated sentence either NEUTRAL or CONTRADICTION (E → {N,C} in Table 2). Generally, only presuppositions of soft triggers can be canceled under certain contextual conditions (Abusch, 2002; Simons et al., 2010).

As expected, presupposition cancellation is extremely uncommon for hard triggers (numeric determiners and clefts) which should always give rise to a presupposition, and high for clause-embedding verbs which do not all presuppose their clausal complement. There was also a non-negligible number of examples with other triggers such as change of state verbs and re-verbs whose purported presupposition does not survive negation, which indicates a moderate level of context-sensitivity.[5]

[5]For example, given the premise *Brother S (didn't) burst into tears*, participants judge that the hypothesis *Brother S wasn't crying before* holds, but judgments most often map on to NEUTRAL with the negated premise. This may be due to the use of *burst* with *into tears* since one can softly cry and then burst into tears, which makes the presupposition weaker in this context.

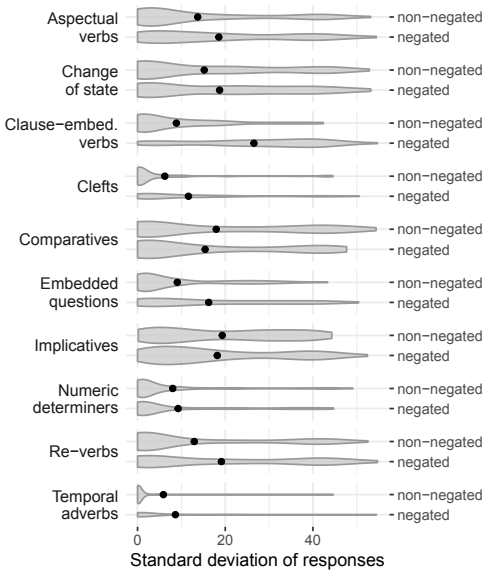

Figure 3: Distributions of standard deviations of responses for each example across all trigger types. Dots represent the mean standard deviation.

Examples with implicatives, comparatives, and aspectual verbs, however, sometimes exhibited the opposite behavior: the non-negated example was labeled NEUTRAL or CONTRADICTION whereas the negated example was labeled ENTAILMENT. Through manual inspection of these examples, we found that these cases involve a range of semantic and pragmatic inferences.[6]

Gradience To determine the extent of gradient judgments, i.e., judgments that don't lie at either end of the probability scale, we considered both the proportion of NEUTRAL labels (see Figure 2) as well as the variability in the continuous ratings for each example. A high proportion of NEUTRAL labels indicates that participants provided many ratings in the middle of the scale, and likewise, higher variability indicates not just that humans disagree more but also that they use a larger proportion of the scale as opposed to just the endpoints.

As we already mentioned, the proportion of NEUTRAL labels was particularly high for implicatives

[6]For example, for the non-negated version of the trigger sentence with a comparative *The PCA results suggest that ecosystem type is a stronger predictor of the index profiles than sequence quality*, participants judged that it does not entail that *sequence quality is a predictor*, presumably because participants concluded that a weaker predictor is potentially no predictor at all.

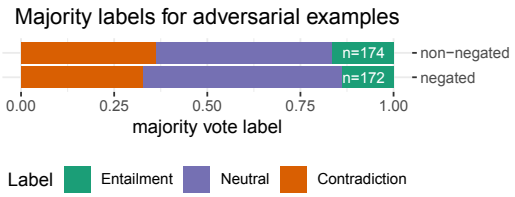

Figure 4: Distributions of mapped human NLI labels for examples in the adversarial corpus. n indicates the number of examples.

and clause-embedding verbs, and greater than 10% for all trigger types other than temporal adverbs and the hard triggers clefts and numeric determiners.

Figure 3 shows the distributions of the standard deviation across all five probability judgments for each example. These distributions confirm the findings from analyzing the proportion of NEUTRAL labels: Variability within examples was highest for clause-embedding verbs and factives; and lowest for temporal adverbs, clefts and numeric determiners. Taken together, these patterns provide evidence that there exists at least a moderate amount of gradience for all triggers except for the hard triggers and temporal adverbs.[7]

4.3 Adversarial Examples

Figure 4 shows the distribution of labels for the adversarial examples. As intended with these additional examples, the hypotheses in most of the adversarial examples were judged to be no longer entailed by the premise. The small proportion of examples labeled ENTAILMENT can largely be attributed to noise from the label-mapping procedure as well as participants being fooled by very subtle changes to the hypothesis (e.g., assigning a rating of 100 to an example in which *John Doe and Jim Miller* was replaced with *Jim Doe and John Miller*).

4.4 Judgment Consistency

To determine whether the human judgments were sufficiently consistent for evaluating models, we computed several agreement statistics commonly used for NLI datasets, and compared them to the statistics for existing NLI corpora. Table 3 shows how often the (mapped) labels of all participants agree, as well as how often individual (mapped) labels agree with the majority label. Despite the potentially noisy mapping from continuous ratings to categorical labels, the agreement numbers are comparable to the ANLI corpus which, like NOPE, contains many challenging examples.

[7]For certain triggers, some NEUTRAL ratings could reflect inconsistencies in stating the presupposition in English, rather than inherent gradience. A reasonable, semantically stable presupposition might exist in a logical metalanguage. This caveat applies to change of state predicates and implicatives where the presupposition statement cannot be mechanically derived from an expression in the trigger sentences.

Measure	SNLI	MNLI	ANLI	NOPE
Unanimous agreement	58.3	58.2	41.6	38.7
Individual lab. = majority lab.	89.0	88.7	81.7	81.5
No majority	2.0	1.8	13.0	5.1

Table 3: Validation statistics (%) reported for SNLI, MNLI, and ANLI compared with NOPE.

5 Machine Learning Experiments

5.1 Models

To evaluate the ability of NLI models to infer presuppositions, we evaluated two baseline models, a Bag-of-Words (BOW) model and InferSent (Conneau et al., 2017), as well as two pre-trained transformer models, RoBERTa-large (Liu et al., 2019) and DeBERTa-V2-XLarge (He et al., 2020), which both recently achieved state-of-the-art performance on the SuperGLUE natural language understanding benchmark (Wang et al., 2019).

The BOW model produces a sentence representation from the mean of FastText word vectors (Mikolov et al., 2018), and InferSent does so using a bidirectional LSTM. We trained baselines end-to-end on the combination of MNLI (Williams et al., 2018), SNLI (Bowman et al., 2015), ANLI (Nie et al., 2020), and FEVER (Thorne et al., 2018). For each baseline model type and training set, we trained 16 models, and evaluated the 5 with highest validation accuracy. We adapted Conneau et al.'s code to train and evaluate these baselines, and include this code in the NOPE codebase.

RoBERTa-large and DeBERTa-V2-XLarge are both transformer-based masked language models with 355M and 900M parameters, respectively. We fine-tuned these models on the above combination of NLI datasets using the HuggingFace Transformers library (Wolf et al., 2020) through an adapted version of the scripts by Nie et al. (2020). We fine-tuned RoBERTa five times with different random seeds. Due to the high computational cost of fine-tuning DeBERTa, we only fine-tuned it once.

5.2 Results and Discussion

As Figure 5 shows, the transformer models achieve high accuracy on the main dataset, and in particular for the trigger types clefts, numeric determin-

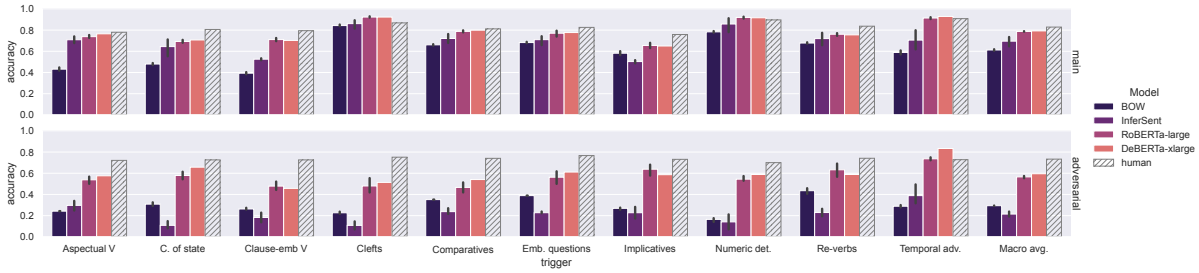

Figure 5: Accuracy on NOPE for both the main and adversarial versions of the corpus. Human scores correspond to the proportion of responses that agreed with the majority label.

Model	E ⟶ E		E ⟶ {N,C}		{N,C} ⟶ E	
	nonneg	neg	nonneg	neg	nonneg	neg
RoBERTa	89.9	89.2	80.6	32.7	35.1	68.4
DeBERTa	90.8	88.6	81.8	32.1	38.9	68.9

Table 4: Model accuracy on non-negated and negated examples for different subsets of the main corpus. "E → E" corresponds to examples in which the presupposition projects out negation and "E → {N,C}" and "{N,C} → E" correspond to examples in which negation affects whether the hypothesis is entailed.

ers, and temporal adverbs, performance is almost at ceiling. The BOW and the InferSent baselines achieve much lower overall accuracy, highlighting that NOPE is a challenging evaluation dataset.

Lexical overlap bias All models achieve lower accuracy on the adversarial data than on the main data, suggesting that all models exhibit a bias for predicting ENTAILMENT when lexical overlap is high. This is particularly true for the two baseline models, which suggests that a large amount of their success on the main dataset could be driven by the lexical overlap between the spelled-out presupposition and the trigger sentence rather than proper inferences involving presuppositions. We thus exclude the baseline models from the subsequent analyses. The performance gap between the main corpus and the adversarial corpus for transformer models, however, is much smaller, which suggests that these models do not primarily rely on a lexical overlap heuristic when making predictions.

Projection To investigate whether models draw correct inferences in cases where the presupposition projects out of negation, i.e., whether they predict ENTAILMENT for both the non-negated and the negated version of the trigger sentence, we computed model accuracy on the subset of the data for which both the non-negated and negated versions of the same example were labeled ENTAILMENT ("E → E" in Table 4). Accuracy was high (>88% for both models) for both non-negated and negated sentences, suggesting that models often draw the correct inferences for sentences in which the presupposition projects out of negation.

Context-sensitivity To investigate the models' behavior on examples in which the trigger sentence does not give rise to a presupposition, we also computed the accuracy on the subset of the main corpus in which the mapped human label differed between the negated and non-negated trigger sentence (columns "E → {N,C}" and "{N,C} → E" in Table 4). These correspond to sentences without actual presupposition triggers (e.g., non-factives such as *think*) and sentences in which the context cancels the presupposition. Accuracy on this subset of the data is much lower, especially for sentences in which the hypothesis is not entailed, suggesting that the models do not exhibit the same sensitivity to negation in the trigger sentences as humans. This could either be because the models are not sensitive to negation at all, as found for other models (Ettinger, 2020), or fail to draw the correct inferences when the context cancels a presupposition.

Gradience Finally, to investigate the models' ability to capture the gradience in human judgements, we computed the model accuracy on the subset of examples that were assigned a NEUTRAL label. Accuracy on these examples is again much lower (39.2% for RoBERTa, and 39.1% for DeBERTa) than accuracy on the overall corpus, which suggests that the transformer models do not draw the correct inferences for trigger sentences for which participants neither fully endorsed nor rejected the hypothesis.

6 Conclusion

We presented NOPE, a dataset of naturally occurring presuppositions involving a diverse set of presupposition triggers. The human judgments in this dataset suggest that the examples we identified give rise to presuppositions in most cases, and that presupposition cancellation and variability exist in a small but non-negligible minority of cases across contexts for presuppositions other than those triggered by clause-embedding verbs and implicatives. We used this dataset to evaluate the pretrained transformer-based models RoBERTa and DeBERTa. We found that the models were able to draw inferences involving presuppositions in the simple cases but failed to fully capture human-level context-sensitivity and gradience.

Acknowledgements

We thank the three anonymous reviewers and the meta reviewer for their thoughtful feedback. This project has benefited from financial support to SB by Eric and Wendy Schmidt (made by recommendation of the Schmidt Futures program), Samsung Research (under the project *Improving Deep Learning using Latent Structure*), Apple, and Intuit, and from in-kind support by the NYU High-Performance Computing Center. SS was in part supported by the Stanford HAI Hoffman-Yee Project *Towards Grounded, Adaptive Communication Agents*. This material is based upon work supported by the National Science Foundation under Grant No. 1850208, as well as under Grant No. 2030859 to the Computing Research Association for the CIFellows Project. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation nor the Computing Research Association.

References

Márta Abrusán. 2011. Predicting the presuppositions of soft triggers. *Linguistics and philosophy*, 34(6):491–535.

Dorit Abusch. 2002. Lexical alternatives as a source of pragmatic presuppositions. In *Semantics and linguistic theory*, volume 12, pages 1–19.

Dorit Abusch. 2010. Presupposition triggering from alternatives. *Journal of Semantics*, 27(1):37–80.

Dorit Abusch and Mats Rooth. 2004. Empty-domain effects for presuppositional and nonpresuppositional determiners. *Context dependency in the analysis of linguistic meaning*.

David I. Beaver. 1997. Presupposition. In *Handbook of logic and language*, pages 939–1008. Elsevier.

David I. Beaver. 2001. *Presupposition and assertion in dynamic semantics*, volume 29. CSLI publications Stanford.

David I. Beaver and Cleo Condoravdi. 2003. A uniform analysis of 'before' and 'after'. In *Semantics and Linguistic Theory*, volume 13, pages 37–54.

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

Emmanuel Chemla. 2009. Similarity: towards a unified account of scalar implicatures, free choice permission and presupposition projection. Ms. LSCP & MIT. Also similar in content to Chapter 5 of my dissertation.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. 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.

Chris Cummins, Patrícia Amaral, and Napoleon Katsos. 2012. Experimental investigations of the typology of presupposition triggers. *HUMANA. MENTE Journal of Philosophical Studies*, 5(23):1–15.

Chris Cummins, Patricia Amaral, and Napoleon Katsos. 2013. Backgrounding and accomodation of presuppositions: An experimental approach. In *Proceedings of Sinn und Bedeutung*, volume 17, pages 201–218.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In J. Quiñonero-Candela, I. Dagan, B. Magnini, and F. D'Alché-Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190. Springer.

Mark Davies. 2008. The corpus of contemporary american english: 450 million words, 1990-present.

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

Judith Degen and Judith Tonhauser. 2021. Prior beliefs modulate projection. *Open Mind*, pages 1–12.

Judy Delin. 1992. Properties of it-cleft presupposition. *Journal of semantics*, 9(4):289–306.

Judy Delin. 1995. Presupposition and shared knowledge in it-clefts. *Language and Cognitive Processes*, 10(2):97–120.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1307–1323.

Jeroen Groenendijk and Martin Stokhof. 1991. Dynamic predicate logic. *Linguistics and philosophy*, pages 39–100.

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 projection problem for presuppositions. *Formal semantics–the essential readings*, pages 249–260.

Irene Heim. 1992. Presupposition projection and the semantics of attitude verbs. *Journal of semantics*, 9(3):183–221.

Orvokki Tellervo Heinamaki. 1974. *Semantics of English temporal connectives*. The University of Texas at Austin.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

Paloma Jeretič, 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.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Evaluating BERT for natural language inference: A case study on the CommitmentBank. 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 6086–6091, Hong Kong, China. Association for Computational Linguistics.

Lauri Karttunen. 1971. Implicative verbs. *Language*, pages 340–358.

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

Lauri Karttunen. 1974. Presupposition and linguistic context.

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

Lauri Karttunen and Stanley Peters. 1979. Conventional implicature. In *Presupposition*, pages 1–56. Brill.

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which linguist invented the lightbulb? presupposition verification for question-answering. 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 3932–3945. Association for Computational Linguistics.

Shalom Lappin and Tanya Reinhart. 1988. Presuppositional effects of strong determiners: a processing account. *Linguistics*, 26(6):1021–1038.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Stephen Levinson. 1983. *Pragmatics*. Cambridge University Press.

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*.

Taylor Mahler. 2020. The social component of the projection behavior of clausal complement contents. In *Proceedings of the Linguistic Society of America*, volume 5, pages 777–791.

Alec Marantz. 2009. Roots, re-, and affected agents: can roots pull the agent under little v. In *workshop Roots: Word Formation from the Perspective of "Core Lexical Elements", Stuttgart*.

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.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901.

Christopher Potts. 2015. Presupposition and implicature. *The handbook of contemporary semantic theory*, 2:168–202.

Ellen F. Prince. 1986. On the syntactic marking of presupposed open propositions. In *Proceedings of the 22nd Annual Meeting of the Chicago Linguistic Society*.

Jacopo Romoli. 2015. The presuppositions of soft triggers are obligatory scalar implicatures. *Journal of semantics*, 32(2):173–219.

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.

Philippe Schlenker. 2021. Triggering presuppositions. *Glossa: a journal of general linguistics*, 6(1).

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5387–5403, Online. Association for Computational Linguistics.

Mandy Simons. 2001. On the conversational basis of some presuppositions. In *Semantics and Linguistic Theory*, volume 11, pages 431–448.

Mandy Simons, Judith Tonhauser, David I. Beaver, and Craige Roberts. 2010. What projects and why. In *Semantics and linguistic theory*, volume 20, pages 309–327.

Scott Soames. 1982. How presuppositions are inherited: A solution to the projection problem. *Linguistic inquiry*, 13(3):483–545.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for Fact Extraction and VERification. 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 809–819.

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 and Judith Degen. 2020. Which predicates are factive? an empirical investigation.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

Wataru Uegaki. 2019. The semantics of question-embedding predicates. *Language and Linguistics Compass*, 13(1):e12308.

Shiva Upadhye, Leon Bergen, and Andrew Kehler. 2020. Predicting reference: What do language models learn about discourse models? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 977–982, Online. Association for Computational Linguistics.

Rob A. van der Sandt. 1992. Presupposition projection as anaphora resolution. *Journal of Semantics*, 9(4):333–377.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Stephen Wechsler. 1989. Accomplishments and the prefix rei-. In *North East Linguistics Society*, volume 19, page 29.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2018. Lexicosyntactic inference in neural models. In *Proceedings of the*

2018 Conference on Empirical Methods in Natural Language Processing, pages 4717–4724, Brussels, Belgium. Association for Computational Linguistics.

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. 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.

A Trigger Types

Change of state verbs Change of state verbs such as *appear* and *snap* presuppose that the entity that was affected by the described event was in a different state just before the event happened. For instance, the sentence *Cats appeared on the street* presupposes that *cats had not been on the street right before then*.

Clefts Cleft constructions such as *It was my cat that made a noise* take the form *It was X who/that did Y* This example presupposes that *something made a noise*, i.e., *someone/something did Y* (see Delin 1992, 1995; Prince 1986; Soames 1982). Note that this example also carries the additional inference that *my cat is the **only** thing that made a noise*, but to keep matters simple, we consider only one presupposition per trigger type and therefore ignore this second inference for present purposes.

Comparatives Comparative constructions such as *Sandy is a bigger cat than Holly* take the form *X is a W-er Y than Z*. The example provided above presupposes that *Holly is a cat* (i.e., *Z is a Y*) (see Levinson 1983; Cummins et al. 2012, 2013).

Aspectual verbs Aspectual verbs such as *start* and *stop* presuppose whether the event that is embedded under these verbs had previously been happening or not (see Abrusán 2011; Abusch 2002). The sentence *Lisa stopped petting Tom's cat* presupposes that *Lisa had previously been petting Tom's cat*. Aspectual verbs are frequently subsumed under change of state verbs but unlike the verbs we consider change of state verbs, aspectual verbs take a non-finite verb phrase (e.g., *petting Tom's cat* or *to eat the kibble*) as a complement.

Embedded questions Embedded questions are realized when a clause is embedded under a wh-word such as *why*, *how*, *where*, or *when* as in *Julia knows why Lisa likes Tom's cat* (see Karttunen 1977; Groenendijk and Stokhof 1991; Uegaki 2019). These triggers presuppose the truth of the embedded content; the example provided above presupposes that *Lisa likes Tom's cat*.

Clause-embedding predicates Clause-embedding predicates frequently presuppose the truth of the finite clause that they embed. This set includes verbs such as *realize*, *know*, and *regret* (Abusch and Rooth, 2004; Beaver, 2001; Heim, 1992; Karttunen, 1973, 1974). The sentence *Julia regrets having forgotten to feed Holly* presupposes that *it is true that Julia forgot to feed Holly*.

Importantly, not all clause-embedding predicates presuppose their complement. For example, predicates such as *think* or *say* do not necessarily entail their clausal complement. However, unlike originally assumed there exists no clear-cut distinction between clause-embedding predicates that do and ones that do not presuppose their complement (e.g., de Marneffe et al., 2019; Tonhauser and Degen, 2020), and therefore we include all clause-embedding predicates that appeared in COCA in the dataset presented in this paper.

Implicative predicates Implicative verbs such as *manage to* and *fail to* presuppose some property of the action in the clause that they embed (see Karttunen 1971; Karttunen and Peters 1979). For example, the sentence *Holly failed to escape her pet taxi* presupposes that *Holly attempted to escape her pet taxi*, since *fail to* implies an attempt at the action. Likewise, *Holly managed to X* presupposes that *X would take effort for Holly*.

Numeric determiners Numeric determiners such as *both* or *all X*, where X is numeric expression such as *three* presuppose that there is a precise number of modified entities in the context (see Lappin and Reinhart 1988). The sentence *All three cat owners that Julia spoke to want another cat* presupposes that *there are three cat owners that Julia spoke to*. We extracted sentences that contain *both (of the) N*, *all NUM (of the) N*, or *all (of the) NUM N*, where *N* is a noun with optional modifiers, and *NUM* is a numeric expression.

***Re*-prefixed verbs** Verbs with the prefix *re-* presuppose that the action of the verb (attaching to *re-*) had taken place in the past (see Beaver 1997; Marantz 2009; Wechsler 1989). The sentence *Holly re-entered the room* presupposes that *Holly had entered the room before*. Hence, *re-V* presupposes that *V had been carried out before*.

Temporal adverbs Adverbial embedded clauses headed by prepositions such as *before*, *after*, *since*, and *while* presuppose the content of the clause they embed (see Beaver and Condoravdi 2003; Heinamaki 1974). The sentence *Lisa petted Tom's cat after she washed her hands* presupposes that *Lisa washed her hands*. We extracted sentences with adverbial clauses headed by *after*, *since*, *before*, *because*, and *while*.

B Word Lists

Many of the presupposition triggers included in NOPE depend on certain lexical items. We used the following word lists for extracting examples with lexical triggers.

Change of state verbs We compiled a list of common change of state verbs by manually extracting unambiguous change of state verbs from the 250 most frequent verbs in COCA, resulting in the following list of 43 verbs: *appear*, *arrive*, *ascend*, *break*, *burst*, *clean*, *close*, *collapse*, *crack*, *crash*, *curl*, *descend*, *die*, *drop*, *enter*, *erupt*, *escape*, *explode*, *expose*, *fall*, *fill*, *fire*, *fix*, *freeze*, *graduate*, *hide*, *hire*, *leave*, *lose*, *melt*, *open*, *pop*, *remain*, *return*, *rise*, *shut*, *sink*, *snap*, *split*, *stay*, *tear*, *wake*, *win*.

Aspectual verbs We used the verbs from the Levin verb classes (Levin, 1993) 55.1 ("begin" verbs) and 55.2 ("complete" verbs), resulting in the following list: *begin*, *cease*, *commence*, *complete*, *continue*, *discontinue*, *end*, *finish*, *halt*, *initiate*, *keep*, *proceed*, *quit*, *repeat*, *resume*, *start*, *stop*, *terminate*.

Embedded questions We used the following wh-words (in combination with syntactic patterns) to extract sentences with embedded questions: *why*, *how*, *where*, *when*, *who*, *what*, *which*.

Implicative predicates We used the following list of implicative predicates, adapted from the list by Karttunen (1971): *avoid*, *bother*, *care to*, *condescend to*, *dare*, *decline*, *fail*, *forget to*, *happen to*, *have the misfortune*, *manage to*, *neglect*, *refrain*, *remember to*, *resist*, *see fit*, *take the time*, *take the trouble*, *venture*.

C Writing Guidelines

In constructing the presupposition statement (the hypothesis) for each example, we adhered to the following guidelines:

- If a first person pronoun is used in the premise, use a first person in the hypothesis
- If the premise uses a third person pronoun, replace it with the full antecedent from context, even if the full antecedent is in a sentence from the prior context rather than the sentence with the trigger.
- If the presupposition trigger is inside a quotation, it can be included only if it is part of a simple speech tag. In this case, use the same pronoun information from the premise in the hypothesis
- If the presupposition trigger is already embedded under negation, create a non-negated minimal pair premise.
- Make small edits to the premise if it makes an otherwise unviable example able to be included. If edits are made, make them to both the negated and non-negated premises. Examples of common edits include:
  - Removing an entire conjunct
  - Changing *but* or *or* to *and* (and vice versa)
  - Removing additional words like *even*, *so*, or *also* that can be unnatural with negation
  - Removing NPIs like *yet* that would not be licensed when the negation is removed
- When there are multiple conjuncts that contain negateable verbs, only negate the conjunct that contains the trigger.

Trigger-specific guidelines Since the presupposition that arises from different triggers can be very specific to that trigger, we also followed some trigger-specific guidelines:

- Aspectual verbs: The reference time is important for judging the presupposition, so adjust the tense of the presupposition to match a more natural-sounding phrase rather than keeping the exact tense from the trigger sentence.
- Change of state predicates: Adjust tense or aspect as needed when referring to the past to make sure the sentence sounds natural, while the presupposition still holds. Can adjust tense/aspect in ways that do not affect truth conditions.
- Clause-embedding verbs: No trigger specific guidelines.
- Clefts: No trigger specific guidelines.
- Comparatives: No trigger specific guidelines.
- Embedded questions: No trigger specific guidelines.
- Implicative predicates: For *manage*, make sure the wording is something like *it* ***would*** *take effort to X* instead of *it* ***took/takes*** *effort to X* to allow the presupposition to hold under negation. For example, if the premise is *They managed to make it to shore*, then the presupposition *it took effort to make it to shore* would

no longer hold with the negated premise *They didn't manage to make it to shore*, but this is not due to the presupposition failing to project, as a presupposition phrased as *it would take effort to make it to shore* holds in both the negated and non-negated premises.

- Numeric determiners: Edit the original (and negated) premise to include a specific domain restriction, and reference that domain restriction in the hypothesis. For example, *both men talked* can become *both men in the room talked* to allow the presupposition to be *there are exactly two men in the room* as opposed to the less natural *there are exactly two men*. We found that statements like *there are exactly two men* were rated as very improbable by participants, even though the statement holds of the event in the premise, but adding a domain restriction that did not otherwise affect the truth conditions of the sentence allowed for a more natural presupposition. When the domain restriction is unstated but highly salient (e.g., *both hands*), it does not need to be added to the premise.
- Re-verbs: Adjust tense/aspect as needed to keep the presupposition sounding natural, so long as the changes do not affect truth conditions. Where possible, use the passive form of the presupposition, as the presuppositions with *re-* hold for the event but not necessarily the event agent.
- Temporal adverbs: If the adverb phrase begins the sentence (e.g., *after Jody left, I called Bill*), reverse the order of clauses so that when the matrix clause is negated, the negation precedes the temporal adverb.

Metadata recorded We additionally recorded many properties of the example sentences that may have an effect on the strength of the presupposition. Though full analysis of these components is beyond the scope of this paper, the full corpus metadata includes expert annotations indicating if any of the following hold:

- *Small edits*: We indicate any cases where we made small edits (e.g., changing 'but' to 'and') to the original sentence that was pulled from COCA.
- *Conjunction*: We indicate whether the premise contains conjunction, regardless of whether the conjoined phrases are DPs, VPs, or CPs.
- *Original is negated*: We indicate if, in the original premise, the presupposition trigger is already embedded under negation, including if the matrix verb is negated, there is a negative existential, and if there's a negative quantifier.
- *Innocent embedding*: We indicate cases where the trigger is embedded, but in a way that is not complicated or does not affect the interpretation of the presupposition (e.g. in a quotation embedded under a speech tag).

D Crowdsourcing Experiment

As mentioned in the main text, we used a prescreener to find high quality participants, and we excluded items from participants who performed poorly on filler items, as described in the following paragraphs.

Pre-screener The pre-screener was open to participants in the US with approval ratings greater than 98% and more than 10,000 previously-completed HITs. It included the same task as the main rating task with 10 hand-selected items to test basic numerical reasoning and the use of the full rating scale. To ensure that participants understood how the slider worked, they completed three practice trials that provided feedback on their use of the slider before they continued to the 10 screening trials.

For each item in the screener, we determined the acceptable range of ratings in advance (e.g., for a premise such as *Jody loves dogs so she went out and adopted a cute little poodle*, the range of acceptable judgments for the hypothesis *Jody adopted a dog* would be anywhere from 95 to 100). To qualify for the main task, participants had to achieve an accuracy of 70% on the pre-sceener. A total of 150 participants completed the screener, of which 107 scored above the accuracy threshold and were given the qualification to complete the main task.

Exclusions In the main task, we used the filler items to filter participants who did not seem to pay attention. Considering that we took the items from the existing MNLI dataset, we knew the entailment relation between the premise and hypothesis. We could therefore compare whether participants provided ratings consistent with the entailment relation or not and use these comparisons as attention checks. Specifically, we counted ratings on fillers with an ENTAILMENT label as correct if a participant provided a rating above 90; and we counted

ratings on fillers with a CONTRADICTION label as correct if a participant provided a rating below 10.[8] We excluded data from in total 15 participants whose accuracy on fillers was below 70%, and collected additional ratings such that we obtained 5 ratings from high-accuracy participants for each item.

E NLI Label Mapping Procedure

As mentioned in the main text, for each participant, we inferred an upper threshold for mapping a probability rating to the NLI label CONTRADICTION and a lower threshold for ENTAILMENT. Probability ratings that fall in between these two thresholds were mapped to NEUTRAL. To infer the lower threshold for ENTAILMENT, we considered the filler examples whose labels in the MNLI dataset are either NEUTRAL or ENTAILMENT. Further, since we expected the threshold to lie above 50%, we only considered examples with a probability rating greater than 50% to filter out noise. We created participant-specific datasets for each participant by upweighting the participant's ratings by a factor of 50. This combination of upweighted participant-specific ratings and global ratings was intended to infer thresholds that better reflect the participant's use of the scale while at the same time avoiding extreme thresholds through still considering the ratings by all other participants. We then optimized the threshold such that the accuracy of the mapping from probability ratings to NLI labels is maximized for the weighted examples under consideration.

To infer the upper threshold for CONTRADICTION, we applied the same procedure to the filler examples whose labels are either CONTRADICTION or NEUTRAL and which received a rating below 50%.

F Comparison to ImpPres

Jeretič et al. (2020) automatically generated a dataset (ImpPres) using templates for evaluating NLI models' abilities to draw inferences involving different presupposition triggers. This work made the simplifying assumptions that there exists no contextual variability and therefore that triggers always give rise to presuppositions.

To compare to what extent the model behavior differs between naturalistic and automatically generated examples, we also evaluated all models on the ImpPres dataset and compared accuracy for the trigger types that are present in both ImpPres and NOPE.

As Figure 6 shows, accuracy on clefts and embedded question triggers are almost at ceiling on the ImpPres dataset, whereas the same models achieve lower accuracy on the examples in NOPE. In the case of clefts, the difference in accuracy between ImpPres and NOPE is very small and the lower accuracy on clefts on the NOPE dataset might be a result of noise in the human annotations.

However, the larger gap in the results for embedded questions, for which we found that variability exists in naturally occurring examples, suggests that for this trigger, the models are only able to draw correct inferences when the presupposition is not canceled by the context (which is the case for all examples in ImpPres).

Accuracy on examples with numeric determiners is extremely low for the transformer models evaluated on ImpPres. This seems to be a result of a difference in how the presupposition was spelled out. In ImpPres, all spelled-out presuppositions triggered by numeric determiners include the modifier *exactly* (e.g, *There are **exactly** three cats on the mat*), whereas we omitted *exactly* from the spelled-out presupposition in NOPE since the presence of *exactly* might trigger additional inferences in humans independent of the presupposition. The low accuracy on ImpPres is caused by the model generally predicting NEUTRAL when the hypothesis contains *exactly*, which may or may not differ from how humans would judge these examples.

Finally, we also observe differences in accuracy on examples with change of state verbs. Manual inspection of the examples in ImpPres revealed that some of the examples with this trigger have been incorrectly generated (e.g., *Gary did get a job* allegedly entails that *Gary is unemployed*), which explains the lower accuracy on these items in ImpPres and also highlights the importance to verify inferences with human participants.

G Model performance on SNLI, MNLI, and ANLI

We evaluated our NLI models on common NLI validation sets. The results are shown in Table 5. Whenever possible, we also include previously published results from similar models. Note, however, that with the exception of RoBERTa, these

[8] These cutoffs are loosely based on the average ratings for ENTAILMENT and CONTRADICTION on the portion of SNLI that Chen et al. (2020) re-annotated with probability ratings.

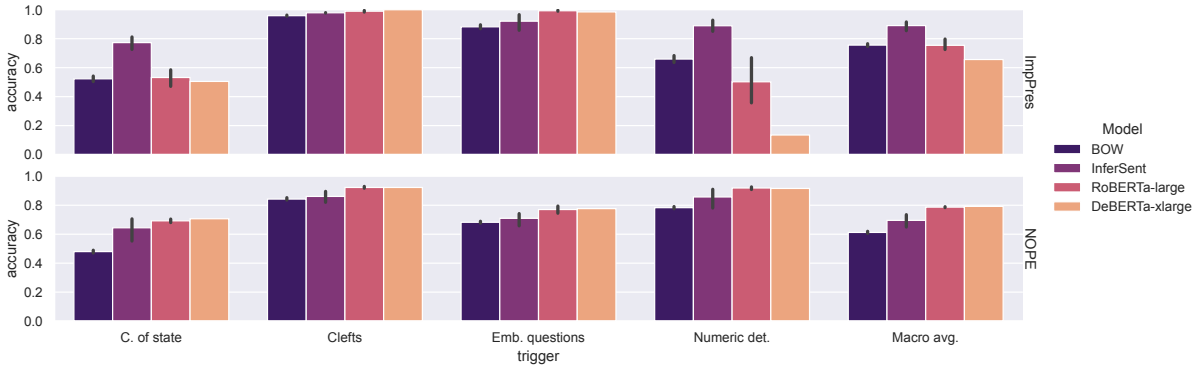

Figure 6: Comparison of model accuracries on ImpPres (Jeretič et al., 2020) and NOPE for all trigger types that are present in both datasets.

results are not entirely comparable, since our models are trained on a combined set of ANLI, MNLI, SNLI, and FEVER, while other models are generally trained or fine-tuned only on a single dataset. We also report an average over several models, while others report best performance. We generally observe similar performance between our models and previously published ones. The main exception is InferSent, which performs over 10 percentage points worse on MNLI than the model evaluated by (Wang et al., 2018). We found that the best of 16 InferSent models trained using our code on MNLI alone achieved 60.9% accuracy on both MNLI validation sets.

	ANLI R1	**ANLI R2**	**ANLI R3**	**MNLI-m**	**MNLI-mm**	**SNLI**	**Macro avg**
BOW (ours)	33.9 (0.81)	35.6 (0.19)	33.4 (0.37)	52.8 (0.33)	54.2 (0.35)	59.9 (0.70)	45.0 (0.30)
InferSent (ours)	34.2 (0.88)	39.3 (1.37)	35.7 (1.16)	54.8 (1.27)	55.5 (1.03)	69.8 (1.67)	48.2 (0.96)
RoBERTa-L (ours)	73.6 (1.07)	52.3 (0.98)	49.2 (0.91)	90.2 (0.17)	90.0 (0.16)	92.9 (0.24)	74.71 (0.16)
DeBERTa-XL (ours)	80 (–)	62.3 (–)	60.5 (–)	91.7 (–)	91.4 (–)	93.9 (–)	80.0 (–)
BOW (Wang et al.)	–	–	–	56.0	56.4	–	–
InferSent (Wang et al.)	–	–	–	66.1	65.7	–	–
RoBERTa (Nie et al.)	73.8	48.9	44.4	91.0	90.6	92.6	73.6
DeBERTa (He et al.)	–	–	–	91.7	91.6	–	–

Table 5: Validation set results for our trained NLI models on ANLI, MNLI, and SNLI. Results given are the average percent accuracy (standard deviation) over the top five models (except for DeBERTa). ANLI-Rn is the validation set from the n^{th} annotation round. MNLI-m and MNLI-mm are the matched and mismatched subsets of MNLI, respectively.