Developmental Negation Processing in Transformer Language Models

Antonio Laverghetta Jr. and **John Licato**
Advancing Machine and Human Reasoning (AMHR) Lab
Department of Computer Science and Engineering
University of South Florida
Tampa, FL, USA
{alaverghett,licato}@usf.edu

Abstract

Reasoning using negation is known to be difficult for transformer-based language models. While previous studies have used the tools of psycholinguistics to probe a transformer's ability to reason over negation, none have focused on the types of negation studied in developmental psychology. We explore how well transformers can process such categories of negation, by framing the problem as a natural language inference (NLI) task. We curate a set of diagnostic questions for our target categories from popular NLI datasets and evaluate how well a suite of models reason over them. We find that models perform consistently better only on certain categories, suggesting clear distinctions in how they are processed.[1]

1 Introduction

Negation is an important construct in language for reasoning over the truth of propositions (Heinemann, 2015), garnering interest from philosophy (Horn, 1989), psycholinguistics (Zwaan, 2012), and natural language processing (NLP) (Morante and Blanco, 2020). While transformer language models (TLMs) (Vaswani et al., 2017) have achieved impressive performance across many NLP tasks, a great deal of recent work has found that they do not process negation well, and often make predictions that would be trivially false in the eyes of a human (Rogers et al., 2020; Ettinger, 2020; Laverghetta Jr. et al., 2021).

In developmental psychology, there has likewise been a great deal of interest in how a child's ability to comprehend negation emerges in the early years of life (Nordmeyer and Frank, 2013, 2018b; Reuter et al., 2018; Grigoroglou et al., 2019). Unlike in NLP, which typically treats negation as representing a single monolithic competency, this research has long understood that there are many kinds of negation used in everyday interactions (Bloom, 1970; Pea, 1982). This ranges from using negation to express a child's rejection of something to clarifying a child's knowledge. These "developmental" categories of negation do not emerge simultaneously; children tend to start using certain kinds before others (Nordmeyer and Frank, 2018a).

Given that these categories represent some of the earliest uses of negation among humans, understanding how well TLMs can master them is important for building more human-like models of language processing. Understanding how well models perform on different categories will indicate whether they have mastery of some forms of negation, while also helping to identify failure points. Another interesting question is whether the proficiency of TLMs on these categories is at all related to competencies in human children (e.g., is the category which models consistently perform the best on the same that children most frequently employ?). However, to our knowledge, no prior work in NLP has focused on how well models perform on the forms of negation of interest to developmental psychology.

In this short paper, we investigate how well a suite of TLMs can process developmental negation,[2] by framing the problem as a natural language inference (NLI) task. We develop a rule-based parser to extract problems from existing NLI datasets, and evaluate our models on each category, in order to determine *(i)* whether certain categories are more solvable by our models than others, and *(ii)* what relationships exist among the categories. We find that models can consistently achieve stronger performance only on certain categories, and that training on combinations or sequences of these categories does not substantially improve a model's downstream performance.

[1] Code and data to reproduce our experiments can be found on Github: https://github.com/Advancing-Machine-Human-Reasoning-Lab/negation-processing-ACL-2022

[2] By which we mean the forms of negation studied in development psychology.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 545 - 551

2 Related Work

Negation is known to be frequently used in everyday conversation. While this includes its logical form, we primarily focus on negation's psycholinguistic forms, especially those that have been studied in the context of developmental psychology. Negation emerges early in child development, with 'no' sometimes being a child's first word (Schneider et al., 2015), and even infants appear to understand forms of negation (Piaget, 1980; Hochmann and Toro, 2021). Preschool children use at least three different kinds of negation (Bloom, 1970), but possibly as many as nine (Choi, 1988). As noted by Nordmeyer and Frank (2018a), one of the first categories children use is *rejection*, where a child rejects an object or activity. This is later followed by *existence*, where a child might express the lack of an object, and later still *denial*, which a child uses to deny the truth of a claim. Larger scale studies of child-directed speech have found that truth-functional kinds of negation tend to emerge later (Liu and Jasbi, 2021), but individual children do vary in their specific order of acquisition (Nordmeyer and Frank, 2018a). It is unknown whether this ordering reflects any deeper dependencies among the different categories, or whether the ordering is reflected in how artificial language models (LMs) learn negation.

In NLP, methods from psycholinguistics have been used to probe the reasoning capabilities of LMs. Results from some studies have indicated that TLMs are not human-like in their processing of negation (Ettinger, 2020; Kassner and Schütze, 2020). A similar line of work has used the NLI task to probe a model's ability to process negation and found that TLMs will often alter their predictions when negation is inserted or removed, even when the negation does not alter the entailment relationship (Hossain et al., 2020; Hartmann et al., 2021). As argued by Kruszewski et al. (2016), part of the challenge of modeling purely logical negation is that a predicate often occurs in very similar contexts regardless of whether it is being negated. They argue that we should view negation as being a "graded similarity function", and show that distributional models can predict human plausibility judgments quite well, even in the presence of negation. These works show that it is unclear how well distributional models, especially TLMs, are actually processing negation. We contribute to this literature from a new perspective, by studying how well models can reason over forms of negation common in developmental psychology.

Category	# Train	# Test
Possession (PO)	1053	520
Existence (EX)	5528	2723
Labeling (L)	2241	1104
Prohibition (PR)	814	400
Inability (I)	1384	682
Epistemic (EP)	1903	936
Rejection (R)	1737	856

Table 1: Summary statistics for the curated dataset.

3 The Developmental Negation Corpus

We use the NLI task to study the negation reasoning capabilities of our models. NLI problems consist of two sentences: a premise (p) and hypothesis (h), and solving such a problem involves assessing whether p textually entails h. The generic structure of the NLI task makes it suitable for studying a variety of underlying reasoning skills, including negation. We specifically use the SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) datasets.

To automatically identify questions that contain a specific kind of negation, we rely on the work by Liu and Jasbi (2021) which studied how frequently different kinds of developmental negation occur in child-directed speech, using the data from the CHILDES corpus (MacWhinney, 2014). To do this, they created a simple rule-based parser to automatically tag each sentence in CHILDES with the type of negation it contained (if any). We reimplement their parser, in some cases tweaking the rules slightly to better suit the structure of the NLI task. For each example across all the splits of both datasets, we first obtain a dependency parse of both p and h using the diaparser package (Wang et al., 2019), and check if either contains an explicit negation marker ("no", "not", or "n't"). If one span contains negation, we check if the syntactic structure obeys the rules of any of our categories. If the span falls into a category, we mark it as belonging to that category. We use these questions as the diagnostic set for our experiments, splitting out 1/3 of the questions in each category as a *diagnostic test* set, and leaving the remainder as a *diagnostic train* set (and we will refer to them as such). We place the remaining NLI questions containing no negation in a separate NLI_{train} set, giving us about 730,000 examples we use to finetune our models on the NLI task. We split out 9,000 questions from this train set at random to use as a NLI_{dev} set, bal-

Category	Premise	Hypothesis
PO	yeah you probably don't have the right temperatures...	You probably have ideal temperatures...
EX	This analysis pooled estimates...	The analysis proves that there is no link...
L	Not orders, no.	It is not orders.
PR	Two people are sitting against a building near shopping carts.	Run that way but don't run into the...
I	His manner was unfortunate, I observed thoughtfully.	I could not pick out what kind of manner he...
EP	yeah i don't know why	I know why
R	I lowered my voice...	I didn't want to be overheard.

Table 2: NLI examples extracted from each category, long examples have been trimmed to fit on one line.

anced for each label. In the following, we describe the precise rules used to determine which category a negated example should be assigned to:

Possession (PO) We require that the lemma of the root be *have*, *has*, or *had*, and that the root is directly modified by both the negation and the verb *do*.

Existence (EX) We require that *there* occur in the text and precede the negative marker and that the negative marker directly modifies a noun phrase, determiner, or an adverb.

Labeling (L) We require that the sentence begin with either *That* or *It*, and that the root of the sentence is a noun which is modified by *is* or *'s*.

Prohibition (PR) We require that the sentence not contain a subject and that the negation is immediately preceded by *do*. To not conflate this category with others, we filter out cases where the root contains one of the explicit markers of another category (e.g., *like* or *want* in the case of rejection).

Inability (I) We require that the negation directly modify the root of the sentence, and that the word immediately before the negation is either *can* or *could* (e.g., *can not do*). Prior literature has typically viewed inability from an egocentric perspective. However, we found that allowing only the first person severely restricted the number of examples extracted, and therefore chose to also allow the second and third person.

Epistemic (EP) We require that the root be *remember*, *know*, or *think*, and that the root be directly modified by the verb *do*.

Rejection (R) We require that the lemma of the root word be either *like* or *want*, and that the root is modified by the negative marker.

After performing extraction, categories L and PR contained fewer than 1000 examples, which we deemed was insufficient to split into separate train and test sets. To address this, we developed a simple data augmentation approach that utilized the Wordnet database (Miller, 1998). From the dependency parse of both p and h, we check if the root of either parse occurs in both spans. If it does, we obtain all synonyms of the word in Wordnet and replace the root in both spans with the synonym (doing this for every synonym). We found this simple approach increased the number of examples for both L and PR to at least 1500. Note that we performed no augmentation for the other categories, as our parser extracted at least 1500 examples for all other cases. Table 1 shows statistics for the dataset after augmentation.

Table 2 shows extracted examples, along with their category assignment. We generally found that the extracted examples matched up with the prototypical category quite well, although in some cases their semantics differed slightly. For instance, consider a PR example with p = *don't miss having a flick through the albums* and h = *The pictures of old Madeira show a more interesting city than now*, which is an MNLI example originally extracted from a travel guide. Although this technically counts as PR, it does not have quite the same semantics as an actual command. Unfortunately, these ambiguities are not easily resolved, given that negation takes on many forms and may occur at any location within a sentence. We, therefore, opted to focus on forms of negation that can be easily extracted, and leave improvements to our dataset creation protocol for future work.

4 Experiments

Using the curated dataset, we performed a series of exploratory experiments to help us understand how well TLMs process each of the negation categories. We use BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019), two popular transformer LMs that have demonstrated impressive results on a variety of language understanding tasks. We also examine MiniBERTa (Warstadt et al., 2020) and BabyBERTa (Huebner et al., 2021), which are both

based on the RoBERTa architecture but were pre-trained on a much smaller number of tokens (10 million and 5 million respectively), which is more realistic to the amount of language a child is exposed to in the first few years of life. We use the Huggingface implementation of all models (Wolf et al., 2020), and use both the *base* and *large* version of BERT and RoBERTa, which differ only in the number of trainable parameters.

Experiment 1: We began by investigating whether TLMs would master certain negation categories sooner than others over the course of training. We train our models on NLI_{train} for 10 epochs, using a learning rate of $1e-5$, a weight decay of 0.01, a batch size of 16, and a maximum sequence length 175.[3] We selected these hyperparameters to be similar to those which were previously reported to yield strong results when training on NLI datasets (Laverghetta Jr. et al., 2021). We additionally evaluated the models on NLI_{dev}, and found that they all achieved a Matthews Correlation of at least 0.6 (Matthews, 1975), and thus concluded that these hyperparameters were suitable. For every end of epoch checkpoint across all models, we obtained evaluation results on each diagnostic test set. Importantly, the models are not finetuned on any negated NLI questions for this experiment, meaning that all knowledge of negation comes from pre-training. Results are shown in Figure 1. We see that the categories have similar rankings in terms of accuracy. For example, L and PO are among the top two best-performing categories, while R is generally one of the worst-performing ones, indicating clear distinctions in how LMs process the categories. BabyBERTa, unlike other models, also shows stronger similarities to how children acquire negation. For instance, while R is thought to be one of the first categories children acquire, BabyBERTa is the only model where R is one of the highest-ranking categories in terms of accuracy.

Experiment 2: One might expect that children develop a more abstract understanding of negation as they are exposed to different categories. This was suggested by Pea (1978) who argued that more abstract forms of negation develop from less abstract ones, suggesting that mastering one form of negation can lead to positive transfer on others. In Experiment 2, we examined how much positive

[3]We set the maximum sequence length for BabyBERTa to 128, which is the longest that the model supports.

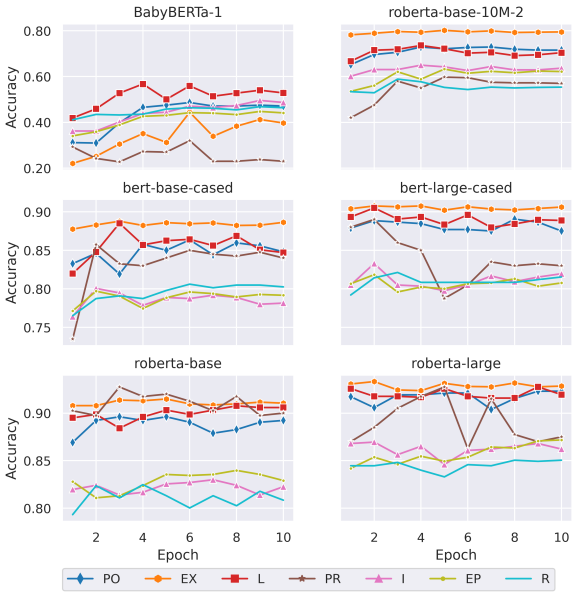

Figure 1: Performance of models finetuned on NLI_{train} for each diagnostic test set. We refer to MiniBERTa using its Huggingface model ID (*roberta-base-10M-2*).

transfer could be obtained from training on one of the negation categories, and then testing on the others. We adopt a similar methodology to Pruksachatkun et al. (2020), who explored the conditions that affect intermediate task transfer learning. Using the models trained in Experiment 1, we further finetune these models for 25 epochs on each diagnostic train set separately. We then evaluate the finetuned models on each diagnostic test set, which allows us to examine all possible pairwise interactions among categories. Figure 2 shows the results for all combinations of diagnostic categories for training and testing. Surprisingly, we find that positive transfer generally only occurs when a model is trained on the same category it is being tested on. Training on a different category has little to no effect on the target category. BabyBERTa is again an exception, as we do see positive transfer for most pairs, suggesting the model is generalizing across categories

Experiment 3: Building on Experiment 2, we examined how the performance of our models is affected when trained on all diagnostic categories in sequence. Assuming that no positive transfer exists among the categories, we would expect to see a model's performance on a particular category improve only after it has been trained on that same category, and even training on multiple other categories should not substantially improve perfor-

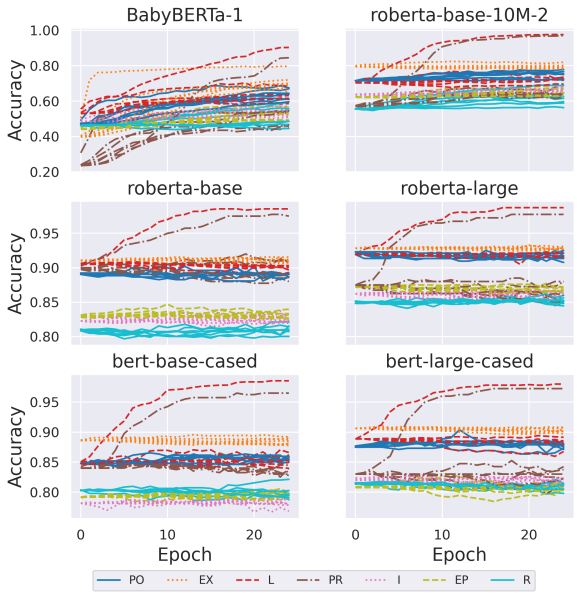

Figure 2: Accuracy of each model on every diagnostic test set, after being finetuned on every diagnostic train set. Plots are color-coded based on the target category.

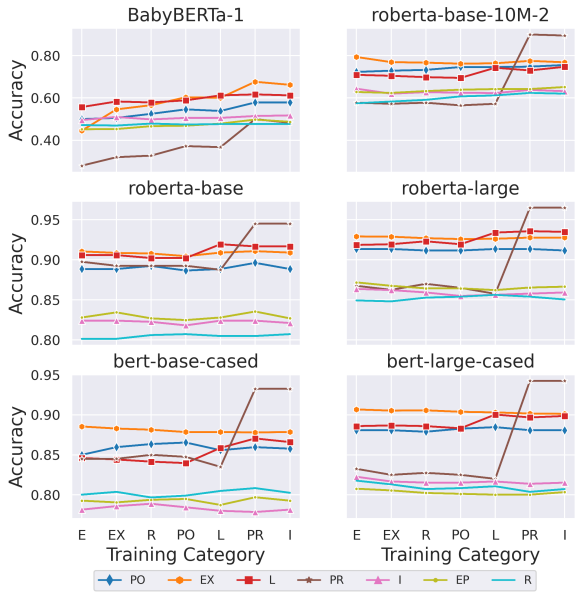

Figure 3: Results from Experiment 3. The x-axis shows the sequence of categories on which all models were trained, while the y-axis shows the accuracy obtained after being trained on a category.

mance on the target. Using the models from Experiment 1, we finetune each model for 10 epochs on every diagnostic train set, using the sequence of categories shown in the x-axis of Figure 3. Additionally, we under-sample all diagnostic train sets to have the same number of questions as PR, so that all categories contribute the same amount of data. Figure 3 shows the results. For some categories, such as L and PR, we see the expected trend. The largest accuracy gain for these categories occurs whenever the model is trained on the same category it is being tested on, and performance drops slightly after being trained on others. However, for categories such as R, the best performance gain is not always after being trained on the same category. We sometimes see the model continue to improve on R after being trained on R, and in some cases, training on R causes performance on R to *decrease*.

5 Discussion and Conclusion

In this paper, we have explored how well transformers process categories of developmental negation. We find that performance rankings across categories are generally consistent, but that the categories seem to test for orthogonal skills in the majority of LMs. In BabyBERTa, we see significant similarities with the order of negation acquisition in children. Two of the best performing categories are R and L, while two of the worst are EX and PR, which aligns quite well to the order observed by Liu and Jasbi (2021). It thus seems that TLMs do at least partially reflect the order of negation acquisition observed in children, although more experiments would be needed to understand the extent of this correlation. That we found category rankings to generally be consistent across LMs may have interesting implications, and understanding why LMs struggle with certain categories may help to improve the ability of LMs to process negation.

Future work can build on these experiments in several ways. In Experiments 2 and 3, we modeled interactions among the negation categories in either a pairwise or sequential fashion, which is unlikely to reflect how children are exposed to negation. More experiments, mixing all of the categories at once in various proportions, might yield a more realistic model of cognitive development. Our approach also requires that each category fits into a specific structure, which limits the amount of examples that can be extracted. Future work will need to expand our ruleset to include more variations in the negated utterances covered. Finally, while we primarily focus on finetuning, pre-training is likely to impact the proficiency of our models on the categories as well. Future work should precisely control the prevalence of each category in the pre-training corpus, to observe what effect this has on downstream performance.

References

Lois Bloom. 1970. Language development: Form and function in emerging grammars. mit research monograph, no 59.

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

Soonja Choi. 1988. The semantic development of negation: a cross-linguistic longitudinal study. *Journal of child language*, 15(3):517–531.

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.

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.

Myrto Grigoroglou, Sharon Chan, and Patricia A Ganea. 2019. Toddlers' understanding and use of verbal negation in inferential reasoning search tasks. *Journal of experimental child psychology*, 183:222–241.

Mareike Hartmann, Miryam de Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. 2021. A multilingual benchmark for probing negation-awareness with minimal pairs. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 244–257, Online. Association for Computational Linguistics.

F. H. Heinemann. 2015. VIII.—The Meaning of Negation. *Proceedings of the Aristotelian Society*, 44(1):127–152.

Jean-Rémy Hochmann and Juan M. Toro. 2021. Negative mental representations in infancy. *Cognition*, 213:104599. Special Issue in Honour of Jacques Mehler, Cognition's founding editor.

Laurence Horn. 1989. *A Natural History of Negation*. University of Chicago Press.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Germán Kruszewski, Denis Paperno, Raffaella Bernardi, and Marco Baroni. 2016. There is no logical negation here, but there are alternatives: Modeling conversational negation with distributional semantics. *Computational Linguistics*, 42(4):637–660.

Antonio Laverghetta Jr., Animesh Nighojkar, Jamshidbek Mirzakhalov, and John Licato. 2021. Can transformer language models predict psychometric properties? In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 12–25, Online. Association for Computational Linguistics.

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*.

Zoey Liu and Masoud Jasbi. 2021. English negative constructions and communicative functions in child language. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Brian MacWhinney. 2014. *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Roser Morante and Eduardo Blanco. 2020. Recent advances in processing negation. *Natural Language Engineering*, pages 1–10.

Ann Nordmeyer and Michael Frank. 2013. Measuring the comprehension of negation in 2-to 4-year-old children. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Ann Nordmeyer and Michael C Frank. 2018a. Individual variation in children's early production of negation. In *CogSci*.

Ann E Nordmeyer and Michael C Frank. 2018b. Early understanding of pragmatic principles in children's judgments of negative sentences. *Language Learning and Development*, 14(4):262–278.

Roy D Pea. 1978. *The development of negation in early child language*. Ph.D. thesis, University of Oxford.

Roy D Pea. 1982. Origins of verbal logic: Spontaneous denials by two-and three-year olds. *Journal of child language*, 9(3):597–626.

Jean Piaget. 1980. *Experiments in Contradiction*. University of Chicago Press.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Tracy Reuter, Roman Feiman, and Jesse Snedeker. 2018. Getting to no: Pragmatic and semantic factors in two-and three-year-olds' understanding of negation. *Child development*, 89(4):e364–e381.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Rose M Schneider, Daniel Yurovsky, and Mike Frank. 2015. Large-scale investigations of variability in children's first words. In *CogSci*, pages 2110–2115. Citeseer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xinyu Wang, Jingxian Huang, and Kewei Tu. 2019. Second-order semantic dependency parsing with end-to-end neural networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4618, Florence, Italy. Association for Computational Linguistics.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020. Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. 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, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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.

Rolf A Zwaan. 2012. The experiential view of language comprehension: How is negation represented. *Higher level language processes in the brain: Inference and comprehension processes*, page 255.