Research Statement

Alex Warstadt

Computational Linguistics, NLP, Language acquisition What can the successes of neural networks for natural language processing teach us about human language acquisition? So far, linguists have learned relatively little from these models, but potentially transformative insights may be within reach. Linguists have long debated how much knowledge of grammar can be acquired through data-driven learning, but only in the last few years have genuine data-driven learners—specifically artificial neural networks—had success learning complex linguistic tasks like translation and summarization. **My research centers on what neural network-based language models (LMs) learn about grammar, and how we can use them as models of human learners.**

To probe LMs' understanding of English grammar, I have developed acceptability judgment benchmarks such as CoLA (The Corpus of Linguistic Acceptability; Warstadt et al., 2019) and BLiMP (The Benchmark of Linguistic Minimal Pairs for English; Warstadt et al., 2020a). Both datasets emphasize broad coverage of English phenomena widely studied in theoretical linguistics and have the scale needed for training and testing neural networks. With a combined citation count of nearly 500, they have enabled researchers in computational linguistics and NLP to show that LMs learn to reproduce human-like acceptability judgments across many English phenomena.

Some grammatical generalizations can be simulated based on surface features such as linear order. During language acquisition, humans always reject such generalizations in favor of rules referencing linguistic features like syntactic category and hierarchical structure. My work has developed the hypothesis that a linguistic inductive bias need not be innate to the learner, but can be acquired simply through exposure to linguistic data. I have found that large LMs like BERT learn to reject linear rules for subject-auxiliary inversion and reflexive licensing in favor of hierarchical rules (Warstadt and Bowman, 2020), and that LMs are more likely to reject surface generalizations as their exposure to training data increases (Warstadt et al., 2020b).

The grammatical abilities of huge models like BERT are interesting in their own right, but they do not give us strong evidence about human learning, because their learning environments are too unlike humans': One popular model, RoBERTa, is trained on a corpus that would require roughly thirty human lifetimes to accumulate. I have begun to address these discrepancies by training LMs on smaller corpora ranging from 1M to 1B words. Evaluating these models on BLiMP, we find that LMs are less data efficient than humans (Zhang et al., 2021). While they show the greatest improvement in grammatical performance at the scale of human learning (fewer than 100M words), improvements slow dramatically, and require billions of words to reach human levels.

To make learnability results from neural networks relevant to human language acquisition, we must make neural networks better models of human learners. My goal is to train models than learn more about grammar in an environment no richer than a typical human's. My lab will train LMs on spoken language of the kind and quantity available to children. We will also train model learners that take advantage of the many forms of non-linguistic input available to children. In an ongoing project, I test whether vision-and-language models can learn grammar more efficiently. The communicative function of language provides another untapped signal. Children, unlike LMs, are motivated to acquire grammar because it enables them to share information, make queries, and carry out directives. Language models, on the other hand, are incentivized only to reproduce their training distribution. Training language generation models in a multi-agent setting may provide the richer and more varied training signal needed to accelerate grammar learning.

Realistic model learners will enable us to establish robust causal links between the nature of the input and the acquired grammar. Studying children, we cannot ethically manipulate the learning environment in ways that could affect the course L1 acquisition, but there are no such limitations on model learners. In an ongoing study, I systematically remove direct evidence for a structural subject auxiliary inversion rule from the input to LMs to test whether they can use indirect evidence to learn a structural rule. More generally, I see great potential in such experiments to gain evidence about the necessary and sufficient inputs for human language acquisition.

Experimental Semantics, Pragmatics **My research also explores the communicative functions behind phenomena in semantics and pragmatics using deep learning, corpus analysis, and human experiments.** Humans have a remarkable ability to infer what is left unsaid, but the semantic analysis of presuppositions as definedness conditions leaves many questions unanswered. Why are presuppositions so abundant in discourse? Are they a device used by a speaker to establish common ground, or an inference made by a listener about a speaker's assumptions? Making these questions more puzzling, we do not understand why sentences like *Did John stop smoking last week?* only sometimes trigger a presupposition that John smoked prior to last week.

To address this gap, I was a lead author of NOPE (Parrish et al., 2021), a corpus of English presupposition triggering expressions in context and in their natural distribution, with gradient human judgments. Through statistical analysis of a wide array of triggers, we found large-scale evidence for challenges to semantic theories of presupposition, such as gradience of triggers and cancellation of presuppositions. In this and previous work (Jeretic et al., 2020), we found that some, but not all, of these properties can be learned from the distributional signal in text: LMs trained on semantic inference with few explicit examples of presuppositions nonetheless learn a more general notion of entailment that includes presuppositions and projective inferences.

Still, computational models and linguistic theories cannot fully capture the contextual cues that cause presuppositions to appear or disappear. Relevance is one cue that could help to close this gap, but previous attempts to formalize relevance in the question under discussion (QUD) framework are insufficiently expressive (Agha and Warstadt, 2020). In search of a gradient alternative, we have conducted human experiments testing information-theoretic measures of relevance from Bayesian pragmatics and the Rational Speech Acts framework (Warstadt and Agha, to appear). We find that an answer's relevance depends more on the degree to which it shifts one's beliefs (KL-divergence), than on how much certainty it provides (entropy reduction). On the other hand, the view of relevance in Bayesian pragmatics overlooks key factors, such as higher-order uncertainty (uncertainty about one's own distribution over the QUD), one's non-linguistic goals, and one's strategy for achieving those goals. A robust theory of relevance may be key to developing rational explanations for how and when we make inferences about the presuppositions of others.

Conclusion Machine learning and data science have led to many natural language processing applications in the last decade, but they have considerable untapped scientific potential in linguistics. In collaboration with computer scientists and acquisitionists, my lab will continue my work studying how variables in the input to LMs shape grammar; and in the long term, we will continue to improve the cognitive plausibility of model learners. We will work with semanticists and cognitive scientists to understand the communicative function of presuppositions, through analyzing the probabilistic inferences of humans and LMs in both experimentally controlled and naturalistic settings. **My lab's mission will be to take advantage of advances in natural language processing to unlock answers to long-standing questions about language acquisition and meaning.**

References

Agha, Omar and Alex Warstadt. 2020. Non-resolving responses to polar questions: A revision to the qud theory of relevance. In *Proceedings of Sinn und Bedeutung*, vol. 24, 17–34.

Jeretic, Paloma, 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*.

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

Warstadt, Alex and Omar Agha. to appear. Testing gradient measures of relevance. In *Proceedings of the Sinn und Bedeutung*, vol. 26.

Warstadt, Alex and Samuel R Bowman. 2020. Can neural networks acquire a structural bias from raw linguistic data? In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020a. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics* 8:377–392.

Warstadt, Alex, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics* 7:625–641.

Warstadt, Alex, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020b. 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)*.

Zhang, Yian, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? 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)*.