Can neural networks acquire a structural bias from raw linguistic data?

Alex Warstadt (warstadt@nyu.edu)
Department of Linguistics, New York University
New York, NY 10003 USA

Samuel R. Bowman (bowman@nyu.edu)
Department of Linguistics & Center for Data Science & Department of Computer Science, New York University
New York, NY 10003 USA

Abstract

We evaluate whether BERT, a widely used neural network for sentence processing, acquires an inductive bias towards forming structural generalizations through pretraining on raw data. We conduct four experiments testing its preference for structural vs. linear generalizations in different structure-dependent phenomena. We find that BERT makes a structural generalization in 3 out of 4 empirical domains—subject-auxiliary inversion, reflexive binding, and verb tense detection in embedded clauses—but makes a linear generalization when tested on NPI licensing. We argue that these results are the strongest evidence so far from artificial learners supporting the proposition that a structural bias can be acquired from raw data. If this conclusion is correct, it is tentative evidence that some linguistic universals can be acquired by learners without innate biases. However, the precise implications for human language acquisition are unclear, as humans learn language from significantly less data than BERT.

Keywords: inductive bias; structure dependence; BERT; learnability of grammar; poverty of the stimulus; neural network; self-supervised learning

Introduction

Humans appear to use structural biases to guide language acquisition. A classic example is the rule for subject-auxiliary inversion: Native English speakers uniformly acquire a rule like the structural generalization in Figure 1 that makes reference to hierarchical syntactic structures, despite the fact that the raw linguistic input often supports linear generalizations which are intuitively just as simple (Chomsky, 1965). Humans are not alone in possessing this inductive bias: Prior investigations have identified some artificial learners with a structural bias by virtue of having a significantly restricted the hypothesis space (Perfors, Tenenbaum, & Regier, 2011) or a hierarchically structured architecture that learns from preparsed data (McCoy, Frank, & Linzen, 2020).

However, these results cannot tell us whether a learner starting with very weak biases can *acquire* a structural bias merely from exposure to raw linguistic data. While inductive biases are often understood to be unchangeable properties of a learner, this need not be the case. For instance, in one dominant paradigm in natural language processing, *pretraining* on raw data is used to create a general purpose sentence processing model like BERT (Bidirectional Encoder Representations from Transformers; Devlin, Chang, Lee, & Toutanova, 2019), which can subsequently be fine-tuned to perform a downstream task. The model's inductive biases with respect to the downstream task may be substantially influenced by the prior knowledge acquired during pretraining.

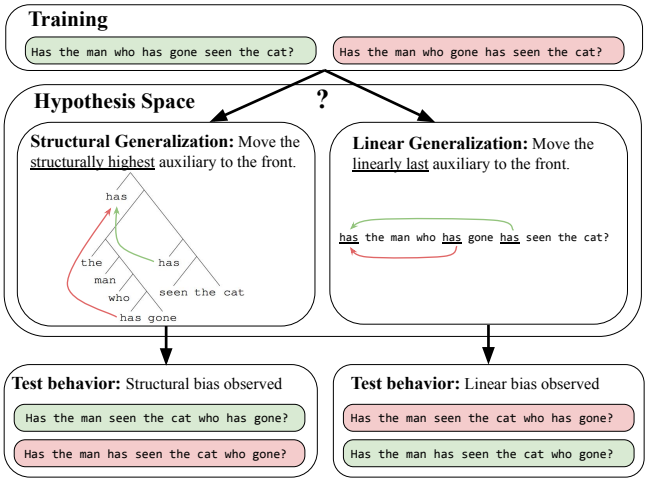

Figure 1: Illustration of the poverty of the stimulus design experiment for subject-auxiliary inversion. Colors correspond to the binary classes for sentences.

In this work, we present new experimental evidence that BERT may acquire an inductive bias towards structural generalizations from exposure to raw data alone. We conduct four experiments inspired by McCoy, Frank, and Linzen (2018, 2020) to evaluate whether BERT has a structural or linear bias when generalizing about structure-dependent English phenomena. We follow the *poverty of the stimulus* design (Wilson, 2006), outlined in Figure 1. First, we fine-tune BERT to perform a classification task using data intentionally ambiguous between structural and linear generalizations. Then, we probe the inductive biases of the learner by observing the behavior of the learner on held-out examples that disambiguate between the generalizations. The experimental datasets illustrate three structure dependent rules of English grammar regarding subject-auxiliary inversion, reflexive-antecedent agreement, and negative polarity item (NPI) licensing. A fourth dataset classifies sentences based on an arbitrary rule: whether a verb in an embedded clause is in the past tense. The data is generated from templates using the generation tools and lexicon created by Warstadt, Cao, et al. (2019) and Warstadt, Parrish, et al. (2019).

©2020 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

The results of these experiments suggest that BERT likely acquires a inductive bias towards structural rules from self-supervised pretraining. BERT generalizes in a way consistent with a structural bias in 3 out of 4 experiments: those involving subject-auxiliary inversion, reflexive binding, and embedded verb tense detection. While these experiments leave open several alternative explanations for this generalization behavior, they add to mounting evidence that significant syntactic knowledge, including a structural biases, can be acquired from self-supervised learning on raw data.

Background & Related Work

Self-supervised Learning and BERT

Recent advances in machine learning for natural language processing give new reason to believe that low-bias learners can acquire significant grammatical knowledge from raw data. The Transformer neural network architecture (Vaswani et al., 2017) used in models like BERT has very weak biases: It is a universal approximator of the class of sequence transduction functions (Yun, Bhojanapalli, Rawat, Reddi, & Kumar, 2019), and it has been applied effectively in non-linguistic domains such computer vision (Parmar et al., 2018) and protein sequence modeling (Rives et al., 2019). However, rather than training a low bias model from scratch to perform a particular linguistic task, recent results show that it is far more more effective to pretrain a general purpose model on raw data and subsequently fine-tune it on a downstream task (e.g. Howard & Ruder, 2018). The implication is that they acquire helpful biases from pretraining.

Crucially, these models are usually pretrained with only raw data using the technique of *self-supervised learning*, which circumvents the need for labeled data by using the data itself as the labels. The most common self-supervised tasks for pretraining are the language modeling task, where the objective is predict the next word in a string (e.g. Peters et al., 2018; Radford et al., 2019), and, in the case of BERT, the cloze task where the objective is predict the identity of a masked token anywhere in a string.

Despite containing no explicit information about grammatical concepts, self-supervised tasks appear to teach neural models significant knowledge of grammar and hierarchical syntax. These models can perform human-like acceptability judgments, which are understood in linguistics as a probe on linguistics competence (Schütze, 1996). When fine-tuned to perform acceptability judgments, BERT approaches human performance on the Corpus of Linguistic Acceptability (CoLA; Warstadt, Singh, & Bowman, 2019), a dataset of over 10k example sentences from linguistics publications with Boolean acceptability judgments. Language models can also correctly discriminate minimal pairs for subject-verb agreement (Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2019), wh-dependencies (Wilcox, Levy, Morita, & Futrell, 2018), and numerous other linguistic phenomena in English (Warstadt, Parrish, et al., 2019) without any supervised training on acceptability. BERT's internal representations appear to attend to linguistic features such as syntactic category (Clark, Khandelwal, Levy, & Manning, 2019) and contain sufficient information from which to recover a dependency parse for an inputted sentence (Hewitt & Manning, 2019). However, it is not known whether BERT is biased towards forming generalizations based on structural features when fine-tuned on structure-dependent phenomena. Indeed, it is possible that BERT could acquire knowledge of hierarchical syntax but still preferentially use surface features to generalize. Our experiments are designed to address this question.

Structure Dependence & the Innateness Hypothesis

The learnability of structural bias has played a large role in debates about human language acquisition. Chomsky (1965, 1971) proposes that humans have an innate bias towards learning structural grammatical rules. For example, children must learn a general rule for subject-auxiliary inversion primarily from input like (1). With this input, a learner could form a structural generalization (front the highest auxiliary in the corresponding declarative) or a linear one (e.g. front the first or last auxiliary), but human learners always choose the former. That is, no human learner of English acquires a linear rule that produces the form in (2a) over (2b). From such examples, Chomsky (1971) concludes that humans have an innate preference for structure-dependent rules. Otherwise it would be difficult to explain how we so consistently avoid deeply un-language-like hypotheses in lieu of significant disconfirming evidence.

(1) a. The cat has gone.
    b. Has the cat gone?

(2) a. *Is the man has seen the cat who going?
    b. Has the man seen the cat who is going?

These examples play a key role in Chomsky's influential argument from the poverty of the stimulus in support of this hypothesis. A version of the argument is given below:[1]

> **Premise 1** Humans form grammatical hypotheses about their native language using either innate biases or data-driven learning.
>
> **Premise 2** Human language learners preferentially form structural hypotheses over equally simple linear hypotheses.
>
> **Premise 3** The raw linguistic input during language acquisition favors neither the structural nor the linear hypothesis.
>
> **Conclusion** Humans' preference for structural generalization is not learned from the raw linguistic input (or any other part of the learner's environment), i.e., it is innate.

This conclusion has spurred much fruitful research into the nature of linguistic universals (see e.g., Chomsky, 1981) and the argument has been fleshed out with evidence child language acquisition (Crain & Nakayama, 1987; Yang, 2000). Nonetheless the argument's validity has been at times been questioned (Pullum & Scholz, 2002; Reali & Christiansen,

[1] See Laurence and Margolis (2001) and Pullum and Scholz (2002) for a more detailed exposition of this argument.

2005; Perfors et al., 2011). If the premises are granted, the reasoning to the conclusion is sound. Premise 1 is a tautology given a suitable definition of data-driven learning. Premise 2 is an robust empirical result of generative linguistics.

Therefore, those arguing against this result have generally taken issue with Premise 3 of the argument, which I refer to henceforth as the *impoverishment assumption*. For instance, Pullum and Scholz (2002) conduct a corpus search to show that such sentences that would provide evidence for the structural rule such as (2b) are attested in naturalistic speech. However, Legate and Yang (2002) counter that such examples are insufficiently frequent to meaningfully impact learning. More importantly, whether or not crucial evidence is lacking for the acquisition of subject-auxiliary inversion, there are certainly numerous other structural rules which have been proposed for natural languages for which crucial examples are vanishingly rare.

Reali and Christiansen (2005) articulate a broader criticism of the impoverishment assumption. It is not only sentences of the form in (2b) that count as evidence for the structural hypothesis. Rather, data from all domains of language provide indirect evidence that militates in favor of a structural understanding of grammar. Statistical regularities in language, they suggest, may be sufficient for "bootstrapping syntax". Indeed, Perfors et al. (2011) find that Bayesian grammar induction system given a choice between several grammar types will preferentially hypothesize phrase-structural rules for English over a flat grammar when presented with child directed speech. This result suggests that the raw data seen be children does favor the acquisition of structural rules in general, counter to the impoverishment assumption, at least if the hypothesis space is strictly limited. However, their models are not low-bias learners, nor do they discover syntactic representations on their own. Rather, they are presented with several hand-crafted candidate grammars of various types including a generally adequate phrase-structure grammar, as well as the syntactic categories of the input data.

Testing the Biases of Neural Networks

There have been several prior efforts to test neural networks for a structural bias in the domain of subject-auxiliary inversion (Lewis & Elman, 2001; Frank & Mathis, 2007; McCoy et al., 2018, 2020). These studies all adopt some form of the *poverty of the stimulus design* (Wilson, 2006), an experimental paradigm that probes the inductive biases of learners by training them on data that is ambiguous between several hypotheses, and evaluating them on examples that disambiguate between these hypotheses. For instance, in a series of papers McCoy et al. (2018, 2020) train neural networks to generate a polar question from the corresponding declarative, using training and test data similar to our subject-auxiliary inversion paradigm shown in Table 1. They find that, while tree-structured models trained using parsed data make a structural generalization, low-bias models never do so consistently.

However, there is good reason to revisit this question with BERT. McCoy et al. do not evaluate BERT, but rather LSTMs that are pretrained on an auto-encoding task in which the model must reproduce the input sentence verbatim. Furthermore, the pretraining data is not naturally occurring, but generated from a restricted lexicon and small context-free grammar. Conneau, Kruszewski, Lample, Barrault, and Baroni (2018) have already been shown that auto-encoders are much weaker at learning syntactic features than surface features even with naturalistic training data. Thus, there is little reason to expect that BERT would perform similarly at this task.

Materials & Methods

We apply the poverty of the stimulus design to test whether unsupervised pretraining gives BERT a structural bias. We conduct four experiments using semi-automatically generated data illustrating different structural generalizations, including the subject-auxiliary inversion paradigm investigated by McCoy et al. (2018, 2020). Each experiment consists of a training phase in which BERT is fine-tuned to classify sentences from an impoverished paradigm consistent with both a structural hypothesis and a linear hypothesis. Then, the classifier is evaluated on the full paradigm which disambiguates between the two hypotheses. If the classifier makes the structural generalization, this is evidence that BERT has learned a structural bias from pretraining on raw data.

Data and Tasks

Examples from the four experimental datasets are shown in Table 1. Each dataset is associated with a binary classification task. In the subject-auxiliary inversion, reflexive, and NPI datasets, the classes correspond to grammatical acceptability of the sentence. In the tense dataset, the classes correspond to whether or not the embedded verb appears in the past tense.

The tense detection task enables us to draw additional conclusions not possible with the acceptability task alone. Since BERT appears to acquire some knowledge of acceptability from pretraining, it might converge on the structural generalization on the acceptability task not by forming a new structural generalization from ambiguous training data, but by accessing an implicit structural rule acquired during pretraining. In the tense detection task there is no reason to expect BERT has acquired the structural generalization during pretraining, and the structural and linear hypotheses are equally arbitrary. Thus, this paradigm tests whether BERT has a structural bias when forming completely novel generalizations.

Subject-Auxiliary Inversion In the subject-auxiliary inversion dataset, the structural and linear hypotheses are defined in terms of where the auxiliary at the front of the sentence has moved from. Each sentence contains two clauses—a main clause and an embedded relative clause—each with an auxiliary verb. In the training examples the embedded auxiliary precedes the main auxiliary because the relative clause modifies the subject. In the test examples the embedded auxiliary follows the main auxiliary because the relative clause modifies the object. Therefore, a linear generalization which

Experiment	Set	Acceptability	Unacceptable
S-Aux-Inv	**Training**	Has the man who is going seen the cat?	Is the man who going has seen the cat?
	Test	Has the man seen the cat who is going?	Is the man has seen the cat who going?
Reflexive	**Training**	The boy that loves himself talks to ladies.	The boy that loves themselves talks to ladies?
	Test	The boy that loves ladies talks to himself.	The boy that loves ladies talks to themselves.
NPI	**Training**	Kids who saw the cats won't get any dogs.	Kids who saw any cats won't get the dogs.
	Test	Kids who won't see any cats get the dogs.	Kids who won't see the cats get any dogs.
		Embedded Past	**Embedded Present**
Tense	**Training**	The critic who sang arias praised a lady.	The critic who sings arias praised a lady.
	Test	The critic praised a lady who sang arias.	The critic praised a lady who sings arias.

Table 1: Data from the subject-auxiliary inversion experiment. According to the relevant linear generalizations, sentences shaded in gray will belong to the positive class, and sentences in white belong to the negative class.

targets the last auxiliary will give the same result for the training as a structural generalization that targets the main auxiliary, but the two generalizations diverge for the test examples.

Reflexive Binding In the reflexive dataset, adapted from Marvin and Linzen (2018), the structural and linear hypotheses depend on the relation between the reflexive pronoun (e.g. *himself*) and a *binder* noun phrase that agrees with it in person and number (e.g. *the boy*). To a first approximation, the structural c-command relation must hold between the binder and reflexive in order for the sentence to be acceptable (Chomsky, 1981). However, in each training examples, the binder also precedes the reflexive only in the acceptable sentences. These generalizations diverge in the test examples, in which there is an unacceptable sentence where the binder precedes, but does not c-command, the reflexive.

NPI Licensing In the NPI dataset, also adapted from Marvin and Linzen (2018), the structural and linear hypotheses depend on the relation between a negative polarity item (NPI; e.g., *any*) and negation. Loosely speaking, negation must c-command the NPI in order for the sentence to be acceptable. The hypotheses are the same (*mutatis mutandis*) as those for the reflexive dataset.

Tense In the tense dataset, each sentence has a main verb and an embedded verb, and the classes correspond to whether the embedded verb has past tense inflection. A related objective is used to evaluate structural knowledge in pretrained models by Shi, Padhi, and Knight (2016) and Conneau et al. (2018). As with the subject-auxiliary inversion dataset, the embedded verb precedes the main verb in the training examples, and follows it in the test examples. Therefore, structural and linear generalizations about the position of the verb learned from the training data will diverge for the test data.

Data Generation The data for the experiments are automatically generated using a method similar to that adopted by Ettinger, Elgohary, and Resnik (2016) and Marvin and Linzen (2018). Sentences are generated from templates describable by a simple context-free grammar. Lexical items are sampled from a hand-crafted vocabulary of over 1000 items labeled with over 30 features required for morphological, syntactic, and semantic well-formedness. Each generated dataset consists of training, development, and test sets each with 10k examples. Examples with similar lexical content are generated in sets of 4, as in Table 1. Within a set of 4, the training items form a minimal pair, as do the test items.

Methods

In each experiment, we train 20 random restarts of classifiers on top of BERT using the labeled training items. Following Devlin et al. (2019), we fine-tune BERT itself in addition to the classifier layer during training. We use Huggingface's (Wolf et al., 2019) implementation of BERT-Large in PyTorch (Paszke et al., 2017), and carry out fine-tuning in `jiant` (Wang et al., 2019). All 20 random restarts for each experiment use identical hyperparameters. The only difference between them is the random seed used to initialize the classifier weights. The models use a learning rate of 2e-5, a dropout rate of 0.2, and a batch size of 16. Training is carried out for 4 epochs, or until 5 evaluations occur without any improvement in development accuracy. During training, the model is evaluated on the development set after 10 batches. These hyperparameters are selected based on an exploratory grid search using the recommended hyperparamer ranges suggested by (Devlin et al., 2019). The hyperparameters selected for the experiments consistently led to the best or near-best development accuracy for each of the datasets.

Results

The results for the experiments in Figure 2 show that BERT has an overwhelming tendency to generalize in a way consistent with structural generalization in the subject-auxiliary inversion, reflexive, and tense settings, but not in the NPI setting. This plot shows the proportion of test minimal *pairs* classified correctly, rather than the proportion of test items

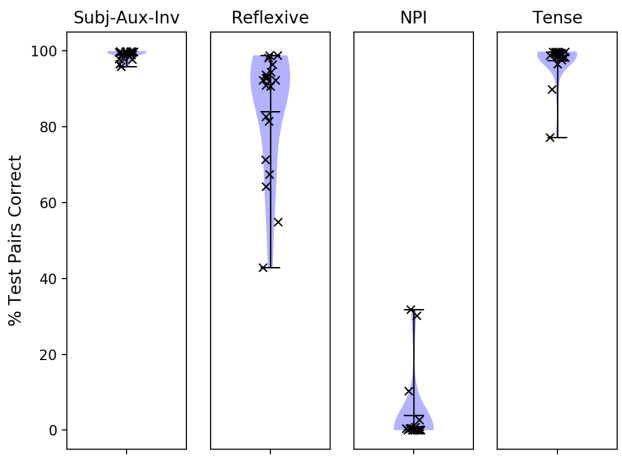

Figure 2: Test results for 20 random restarts of the 4 experiments. "% Test Pairs Correct" is the percentage of minimal pairs from the test templates correctly classified. Individual runs are plotted as x's with random horizontal jitter.

classified correctly. Because there are four ways to classify a pair of items, a totally random classifier would have an expected accuracy of 25% on this metric. In each experiment, minimal pair accuracy on held-out examples from the training templates is over 90%, indicating that the models robustly learned a generalization consistent with the training data.

On the subject-auxiliary inversion task, BERT classifiers appear to make structural generalization with very high consistency and accuracy. The median minimal pair accuracy is over 99%, and the minimum is over 95%.

The reflexive classifiers also appear to make the structural generalization in most cases, but are less consistent. The median minimal pair accuracy is over 92% and the maximum exceeds 99%. However, there is a portion of classifiers that classify the test minimal pairs correctly only about half the time, despite achieving high performance on the training examples. The other half of the time, they tend to classify the test examples in a way consistent with the linear hypothesis. However none of these classifiers *systematically* makes predictions consistent with the linear generalization.

The tense classifiers make a structural generalization very consistently, with a median minimal pair accuracy over 99%. There are two outliers where accuracy is between 75%-90%.

Finally, the results from the NPI experiment are the exceptional case where we do not observe behavior consistent with a structural generalization. The median minimal pair accuracy is effectively 0%, and the maximum is barely above chance at 32%. While every model classifier is able to classify the training pairs perfectly, only 6/20 consistently classify the test pairs in the same way. All of these classifiers learn a generalization that is different from the hypothesized linear generalization shown in Table 1. We find that their performance is consistent with grouping together all and only sentences with an NPI towards the end of the sentence.

Discussion

These results suggest that it is likely that BERT does acquire some form of a structural inductive bias from self-supervised pretraining, at least outside of the NPI domain. They point more strongly in this direction than earlier results by McCoy et al. (2018, 2020). If this interpretation is correct, it would cast some doubt on the impoverishment assumption from Chomsky's (1965) argument from the poverty of the stimulus by showing that raw data does contain overwhelming evidence that language is hierarchical. If some learner does not require innate bias to discover the utility of preferring structural rules over linear ones, it stands to reason humans may not either. On the other hand, our results are consistent with other interpretations, and so we caution against leaping to this strong conclusion, at least without further evidence.

The NPI Results First, in the NPI domain, BERT does not show a structural bias. However, it does not immediately follow that BERT does not have a structural bias at all. By virtue of the paradigm's design, the classes that the model converged on are consistent with both a linear and a structural position. As mentioned above, 6/20 classifiers classified the test items systematically, though not in the way predicted in Table 1. Instead, they grouped the top left and bottom right sentences in one class, and the top right and bottom left sentences in the other. The sentences in the first class might be characterized as all sentences with an NPI at the end of the sentence (a linear generalization), or all sentences with an NPI in the main clause (a structural generalization). Additional experiments are needed to determine which of these outcomes has occurred.

Furthermore, humans do not necessarily show a structural bias in processing similar examples. The unacceptable test items in the NPI paradigm, where the negation precedes the NPI, are known as *NPI illusions*, because they can spuriously appear to be acceptable to humans, and pattern with grammatical sentences in self-paced reading and ERP experiments (Xiang, Dillon, & Phillips, 2009). Thus, NPIs may in retrospect not be the clearest example of humans' structural bias.

Limitations of the Poverty of the Stimulus Design Some doubts remain even in the domains where BERT appears to show a structural generalization. Some surface features could accidentally give the same predictions as the structural generalization on the test data. For instance, in the subject-auxiliary inversion data in Table 1, a classifier could coincidentally identify the acceptable examples by learning to identify a string with a relativizer adjacent to an auxiliary (e.g. *who is*). In fact, we control for this particular confound by generating acceptable examples where a finite verb follows the relativizer (e.g. *Has the man who went seen the cat?*).

However, there are likely other surface generalizations that are consistent with the results. This problem can be addressed

by training and evaluating on data that contradict these generalizations, but alternative convergent hypotheses cannot be eliminated entirely. This is a fundamental limitation of the poverty of the stimulus design: It is not possible to determine that BERT is adopting any particular generalization.

That said, if we continue to find convergent from multiple unrelated domains, Occam's Razor tells us that we should conclude BERT has a structural bias. Given the large number of conceivable surface generalizations, let us assume that an arbitrary generalization is equally likely to support any of the four classification behaviors for the test minimal pair. It follows that if the classifier does make some surface generalization, there is a 1 in 4 chance for each experiment that it would accidentally align with the structural generalization. Then the probability that this chance alignment would occur in at least 3 out of 4 domains is about 5%.

This worry could be alleviated further if it could be shown that baseline models without unsupervised pretraining tend to make the linear generalization on these datasets. These experiments will have to be included in future work. However, at present, the results of McCoy et al.'s (2018) experiments can be used as a proxy. As described in Section (2), these experiments test the ability of sentence encoders without substantial unsupervised pretraining to generalize from a paradigm resembling the polar question data in my experiments. In 5 out of 6 of the model architectures they tested, the linear generalization was preferred. While the task in their experiment is different the acceptability judgment task in the present work, based on this finding it seems that sequence models without substantial unsupervised pretraining are likely to prefer the linear generalization in the polar question domain.

Conclusion This work presents new evidence that highlights the possibility that language learners could *acquire* a structural inductive bias from statistical regularities in raw linguistic data. In particular, we find the most comprehensive evidence to date (to our knowledge) of a low-bias learner demonstrating a structural bias acquired through unsupervised learning on raw data. However, evidence from other empirical domains is needed to fully evaluate this conclusion.

Future work should draw more direct connections between neural networks and human language learners. BERT is trained on data from domains far outside the input to human learners, and in much greater quantities. Indeed, the quantity and quality of data are two other pillars of Chomsky's (1965) argument from the poverty of the stimulus. Therefore, it is essential to replicate these experiments with models trained on less data, which bears greater resemblance to the input to a child. Furthermore, (Chomsky, 1965) observes that human languages generally lack linear rules. If neural networks can acquire human-like biases, they should also struggle to form certain kinds linear generalizations. As techniques for machine language learning and self-supervised pretraining continue to advance, we expect to learn more about which linguistic universals are and are not learnable from data.

Acknowledgments

We thank Chris Barker, Chris Collins, Stephanie Harves, Brenden Lake, Tal Linzen, Alec Marantz, Tom McCoy, and the audience at NYU's Syntax Brown Bag for helpful feedback. This material is based on work supported by the National Science Foundation under Grant No. 1850208. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This project has also benefited from support to SB by Eric and Wendy Schmidt (made by recommendation of the Schmidt Futures program), by Samsung Research (under the project Improving Deep Learning using Latent Structure), by Intuit, Inc., and by NVIDIA Corporation (with the donation of a Titan V GPU).

References

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Chomsky, N. (1971). Problems of knowledge and freedom: The Russell lectures.

Chomsky, N. (1981). *Lectures on government and binding*.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single &!#* vector: Probing sentence embeddings for linguistic properties. In *Acl 2018-56th annual meeting of the association for computational linguistics* (Vol. 1, pp. 2126–2136).

Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language*, 522–543.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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)* (pp. 4171–4186).

Ettinger, A., Elgohary, A., & Resnik, P. (2016). Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st workshop on evaluating vector-space representations for nlp* (pp. 134–139).

Frank, R., & Mathis, D. (2007). Transformational networks. *Models of Human Language Acquisition*, 22.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2019). Colorless green recurrent networks dream hierarchically. *Proceedings of the Society for Computation in Linguistics*, *2*(1), 363–364.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. 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)* (pp. 4129–4138).

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 328–339).

Laurence, S., & Margolis, E. (2001). The poverty of the stimulus argument. *The British Journal for the Philosophy of Science*, *52*(2), 217–276.

Legate, J. A., & Yang, C. D. (2002). Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review*, *18*(1-2), 151–162.

Lewis, J. D., & Elman, J. L. (2001). Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In *Proceedings of the 26th annual boston university conference on language development* (Vol. 1, pp. 359–370).

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1192–1202).

McCoy, R. T., Frank, R., & Linzen, T. (2018). Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. In *Proceedings of the 40th annual conference of the cognitive science society.*

McCoy, R. T., Frank, R., & Linzen, T. (2020). Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, *8*, 125-140. Retrieved from https://doi.org/10.1162/tacl_a_00304 doi: 10.1162/tacl_a_00304

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., & Tran, D. (2018). Image transformer. In *International conference on machine learning* (pp. 4055–4064).

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., ... Lerer, A. (2017). Automatic differentiation in pytorch.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (Vol. 1, pp. 2227–2237).

Pullum, G. K., & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, *18*(1-2), 9–50.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8).

Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, *29*(6), 1007–1028.

Rives, A., Goyal, S., Meier, J., Guo, D., Ott, M., Zitnick, C. L., ... Fergus, R. (2019). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*, 622803.

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. University of Chicago Press.

Shi, X., Padhi, I., & Knight, K. (2016). Does string-based neural MT learn source syntax? In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1526–1534).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30* (pp. 5998–6008). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Wang, A., Tenney, I. F., Pruksachatkun, Y., Yu, K., Hula, J., Xia, P., ... Bowman, S. R. (2019). *jiant 1.2: A software toolkit for research on general-purpose text understanding models.* http://jiant.info/.

Warstadt, A., Cao, Y., Grosu, I., Peng, W., Blix, H., Nie, Y., ... Bowman, S. R. (2019). Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of emnlp-ijcnlp* (pp. 2870–2880).

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2019). BLiMP: A benchmark of linguistic minimal pairs for English. *arXiv preprint arXiv:1912.00582*.

Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, *7*, 625–641.

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp* (pp. 211–221).

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive science*, *30*(5), 945–982.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, *abs/1910.03771*.

Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: Erp evidence. *Brain and Language*, *108*(1), 40–55.

Yang, C. D. (2000). *Knowledge and learning in natural language*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S. J., & Kumar, S. (2019). Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*.