Evaluating Neural Language Models as Cognitive Models of Language Acquisition

Héctor Javier Vázquez Martínez[1] **Annika Heuser**[1] **Charles Yang**[1,2] **Jordan Kodner**[3]
[1]University of Pennsylvania, Deptartment of Linguistics
[2]University of Pennsylvania, Deptartment of Computer and Information Science
[3]Stony Brook University, Dept. of Linguistics & Inst. for Advanced Computational Science
hjvm@sas.upenn.edu aheuser@sas.upenn.edu
charles.yang@ling.upenn.edu jordan.kodner@stonybrook.edu

Abstract

The success of neural language models (LMs) on many technological tasks has brought about their potential relevance as scientific theories of language despite some clear differences between LM training and child language acquisition. In this paper we argue that some of the most prominent benchmarks for evaluating the syntactic capacities of LMs may not be sufficiently rigorous. In particular, we show that the template-based benchmarks lack the structural diversity commonly found in the theoretical and psychological studies of language. When trained on small-scale data modeling child language acquisition, the LMs can be readily matched by simple baseline models. We advocate for the use of the readily available, carefully curated datasets that have been evaluated for gradient acceptability by large pools of native speakers and are designed to probe the structural basis of grammar specifically. On one such dataset, the LI-Adger dataset, LMs evaluate sentences in a way inconsistent with human language users. We conclude with suggestions for better connecting LMs with the empirical study of child language acquisition.

1 Introduction

The growth of neural language models (LMs) for NLP over the past decade has been followed by a growth in research on the potential of these models to provide insights into the cognitive aspects of human language acquisition, representation, and processing (Linzen and Baroni, 2021). Good, even human-like, performance on NLP tasks does not necessarily imply that LMs solve these in human-like ways, so computational linguists have designed a wide variety of experimental paradigms to probe specific properties of the models' linguistic knowledge (Linzen et al., 2016a; Chowdhury and Zamparelli, 2018; Gulordava et al., 2018; Wilcox et al., 2018; McCoy et al., 2020; Hu et al., 2020; Warstadt et al., 2020; Papadimitriou et al., 2021; Huebner et al., 2021) These range from ways of classifying or extracting structures from internal representations (e.g., Hewitt and Manning, 2019; Tenney et al., 2019; Tucker et al., 2021; Papadimitriou et al., 2021), to building tasks inspired by psycholinguistic processing studies and classic acceptability rating task that theoretical linguists use to infer grammatical knowledge (e.g., Linzen et al., 2016a; Warstadt et al., 2020; Huebner et al., 2021; Sinclair et al., 2022).

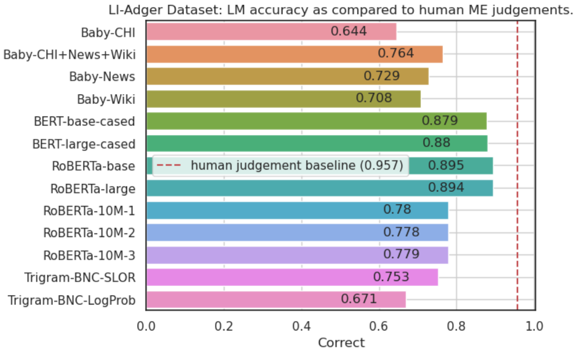


Figure 1: LM performance on the LI-Adger dataset. Human performance is marked by the vertical line. Baby=BabyBERTa, CHI=AO-CHILDES, News=AO-NEWSELA, Wiki=Wikipedia-1.

Of these approaches, acceptability rating may be the most popular. Large acceptability rating data sets focusing on syntax, semantics, and morphology, such as BLiMP (Warstadt et al., 2020), SyntaxGym (Gauthier et al., 2020), and CoLA (Warstadt et al., 2019) lend themselves to benchmarking, and these sit alongside myriad smaller scale studies focused on specific lingusitic phenomena (e.g., Linzen et al., 2016b; Marvin and Linzen, 2018; Wilcox et al., 2018). Results have been impressive for the most part. It appears, from the logic of these studies, that many state-of-the-art neural models are capable of inducing human-like grammatical knowledge on unannotated data – like children during language acquisition.

Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP, pages 48–64
December 6, 2023 ©2023 Association for Computational Linguistics

1.1 Implications for Language Acquisition?

Neural model training differs from human language acquisition in key ways, perhaps most obviously, in that most models are trained on orders of magnitude more input (in plain text form) than humans receive (in spoken or signed form)– BERT was trained on about 3.3B forms, and Chinchilla on 1.4T, while an English-learning child only receives about 10M word per year, for a total vocabulary measured in the hundreds at age three (Fenson et al., 1994; Bornstein et al., 2004).

Recent studies have begun to address this. Can we build models that learn from input on the scale of language acquisition? Would these models then inform our understanding of human language acquisition? Warstadt and Bowman (2022) favor this perspective. They argue that a computational model that performs well on behavioral probing benchmarks when trained on ablated input, that is at least as limited as a human learner's input, is evidence that the model is a good proxy for human linguistic knowledge. Huebner et al. (2021) showed that a specially tuned model trained on only 5M tokens of child-directed speech (CDS) performs well on a purpose-designed data set. And in 2023, an aptly-named shared task, the CoNLL/CMCL BabyLM Challenge,[1] is asking participants to train on only 100M words (about the input of an adolescent) before testing on acceptability benchmarks.

1.2 Goals of the Paper

A push towards extracting performance on smaller training data is a welcome change for the field. In addition to its possible cognitive implications, the drive will also benefit efficient NLP and NLP for low-resource languages. However, while we look forward to the impending engineering advances, we also urge caution in the approaches used to draw scientific conclusions about the nature of neural models' linguistic knowledge. In particular, we take issue with Warstadt and Bowman (2022)'s assertion that "positive results from model learners are more meaningful than negative results."

Their reasoning follows that of an existence proof. If a model that strictly lacks any advantages over humans nevertheless succeeds at a task that requires human-like linguistic knowledge, then it is proof that there exists at least one model with human-like linguistic knowledge. A failure only tells us that this model failed for some reason that may or may not be relevant to the question at hand.

However, this line of reasoning requires faith in the evaluation. If there are any potentially unrecognized non-human-like ways to succeed at the task, or if the task does not truly reflect acquisition, or the task does not actually test a relevant structural property of language, then a positive result becomes inconclusive at best. Unexpected shortcuts emerging from unforeseen biases in evaluation abound across NLP (Chao et al., 2018; McCoy et al., 2019; Wang et al., 2022), so this is a realistic concern. Even the underlying reasoning that "if a (neural) model X behaves like cognitive system Y, then it is equivalent to Y" may be fraught (Guest and Martin, 2023).

In this paper,[2] we evaluate LMs as models of language acquisition on two benchmarking data sets: the widely used Benchmark of Linguistic Minimal Pairs (BLiMP; Warstadt et al., 2020), which also forms part of the evaluation for the BabyLM Challenge, and Zorro (Huebner et al., 2021), a data set inspired by BLiMP with restricted vocabulary for acquisition-inspired models trained only on CDS.

Section 2 reviews the nature of linguistic knowledge and child language acquisition. In Section 3, we introduce the BLiMP and Zorro benchmarks and subject them to baseline tests by simple non-human-like models. These establish several weaknesses in the organization and content of both benchmarks. In Section 4, we evaluate neural models on a more challenging data set derived directly from theoretical linguistics papers. We find that LMs are not necessarily human-like in terms of within- and across-model variability. Finally, Section 5 concludes with a discussion of the logical problem of behavioral probing. We argue for **(a)** benchmarks that better probe the structural knowledge of syntax, **(b)** tests that reflect the developmental findings of language acquisition, and **(c)** more baseline models.

2 Knowledge of Language and its Acquisition

One of the goals of linguistic theory is to characterize the properties that distinguish grammatical from ungrammatical sentences in a language. The empirical study of grammaticality, however, mainly relies on native speakers' acceptability judgments, which interact with other cognitive and perceptual

[1] `https://babylm.github.io/`

[2] Our evaluation code and data are available at `https://github.com/hjvm/benchmarking_acquisition.git`

systems and generally produce gradient results. For example, longer and more complex sentences, even when fully grammatical, are rated as less acceptable than shorter and simpler sentences. Nevertheless, large-scale investigations have established the structural basis of a categorical grammar (Sprouse and Hornstein, 2013). For example, syntactic constraints that prohibit certain transformational processes are shown to have a "super-additive" effect that go beyond acceptability rating due to sentence length and other non-structural factors. Furthermore, acceptability judgments collected at scale are highly consistent with the data reported in the theoretical literature typically gathered informally with few consultants (Sprouse and Almeida, 2012; Sprouse et al., 2013; Sprouse and Almeida, 2017).

The structural basis of language and its uniformity across the linguistic community can be better appreciated from the perspective of child language acquisition. Recent years have seen renewed interest in individual differences across child learners (Kidd et al., 2018), especially with respect to vocabulary acquisition (Frank et al., 2021). It is at least possible that children differ in their cognitive abilities for language and learning, but it is empirically obvious that they differ in their experience with language. Longitudinal records of child language development have made it possible to track both children's vocabulary growth, and the development of the structural aspects of their grammar. In the Providence Corpus (Demuth et al., 2006), for example, six children were recorded at regular intervals from age 1 to 3. On average, fewer than 20% of the first 100 words are shared between any two children. The overlap merely rises to about 40% for the first 1,000, which is the upper limit of a three-year old's vocabulary size (Hart and Risley, 1995; Bornstein et al., 2004). Yet these children's grammars are highly uniform even at this stage. Major syntactic categories, word order and argument structure, and the core morphological rules are firmly established before age three (Brown, 1973) on the basis of at most around 10M words per year (Hart and Risley, 1995) and a vocabulary size of only a few hundred types (Fenson et al., 1994), and all children produce similar grammatical errors during this time. Recent decades have also seen a convergence between the psychological and formal study of language development and the quantitative study of language variation in early childhood. The sociolinguist, Bill Labov, remarks that "The end result is a high degree of uniformity in both the categorical and variable aspects of language production, where individual variation is reduced below the level of linguistic significance" (2012).

The acquisition of vocabulary and grammar provide clues for investigating the capacities of LMs. Vocabulary learning is a matter of rote learning. This includes not just the arbitrary pairing of sounds and meanings, but also morphological processes (e.g., irregularity) and syntactic structures (e.g., sub-categorization, collocations, etc.). There is no escape from experience: more data results in better learning. But, the structural aspects of the grammar are different. They require form generalizations over the vocabularies.

The distinction between rote learning and structural learning (words vs. rules) is not well reflected by existing LM benchmarks including those discussed in this paper. In practice, these benchmarks are a mixture of tests for both vocabulary learning and grammar learning. Moreover, they are stochastically generated by templates: as such, a large number of test sentences are immediately available, but they lack the structural diversity that has proven revealing in the theoretical study of grammar.

Furthermore, the sentences are sometimes highly unnatural and semantically/pragmatically uncontrolled, which is precisely the confounding factor that linguists seek to neutralize when attempting to uncover the structural basis of language. Warstadt et al. (2020) are aware that their templates generate unnatural sentences, presenting the BLiMP sentence '`Sam ran around some glaciers.`' as an example. We found similar issues in Zorro, such as '`the lie on the foot is flat .`,' the first sentence in Zorro's `across_prepositional_phrase` paradigm (`lie` is a noun). The BLiMP authors state that this is not a problem because it affects both sentences in a pair, but how can we rule out unintended interactions between the grammatical phenomenon under evaluation and the semantic implausibility? Sprouse et al. (2018) find that this semantic implausibility may affect judgments of sentence well-formedness, even in the Forced Choice (FC) task used to collect the human baselines in BLiMP.

Indeed, there are already a large amount of carefully curated linguistic materials that are not only structurally diverse but also have minimized lexical and semantic confounds. Furthermore, these datasets (e.g., the Adger/LI dataset; Section 4) have been evaluated for acceptability at an individual level by a large pool of native speaker subjects and show very high convergence rates across individuals. They will be especially informative if we are to explore the structural knowledge of LMs.

3 Re-examining the Benchmarks

BLiMP (Warstadt et al., 2020)

Warstadt et al. (2020) introduce the Benchmark of Linguistic Minimal Pairs (BLiMP)[3] as a means of evaluating the linguistic knowledge of neural language models. BLiMP extends the reasoning of earlier studies (e.g., Linzen et al., 2016b; Marvin and Linzen, 2018; Wilcox et al., 2018) which use a minimal pair paradigm to approximate acceptability judgments. Instead of prompting for a acceptability judgments on individual sentences, as is most commonly done for human subjects, they present an LM with two sentences that only differ in one structural or lexical property. For a given minimal pair m_i consisting of an acceptable sentence $s_{i,1}$ and an unacceptable sentence $s_{i,2}$, if an LM evaluates $P(s_{i,1}) > P(s_{i,2})$, then the model has succeeded on m_i. An LM is scored according to the percentage of all the minimal pairs for which it identified the acceptable sentence. The minimal pair approach allows for the direct evaluation of LMs without training a binary classifier on top of them as was necessary for previous acceptability benchmarks (e.g., CoLA; Warstadt et al., 2019).

Minimal pairs need to be carefully constructed to control for length and lexical frequencies. BLiMP aims to accomplish this with automatic generation from templates, but as we discuss, it often yields sentences with low structural diversity and implausible semantics. The benchmark corpus includes data sets for 12 linguistic phenomena, including ANAPHOR AGREEMENT, ARGUMENT STRUCTURE, BINDING, CONTROL/RAISING, and others listed in the Appendix. These are further divided into 67 paradigms, each containing 1000 sentences pairs, which are meant to test variants of the phenomena, for example the phenomenon DETERMINER-NOUN AGR. contains 6 paradigms for adjacent agreement, agreement with irregular nouns, and agreement with adjectives intervening. BLiMP has become a standard NLP benchmark for this task and will be used as part of the test data for the upcoming BabyLM Challenge.

Zorro (Huebner et al., 2021)

Huebner et al. (2021) explicitly aim to evaluate the relationship between LMs and the acquisition of grammar. They introduce BabyBERTa_AO-CHILDES "an acquisition-friendly version of RoBERTa," trained on English child-directed/produced speech (CDS) approximating the total input of a typical English-learning six-year-old. They train variants on only CDS from AO-CHILDES (Huebner and Willits, 2021), a pre-processed version of English CHILDES (MacWhinney, 1991), as well as variants on larger datasets from other sources.

Because BabyBERTa_AO-CHILDES (henceforth BabyBERTa) was trained on much less text than typical large transformer models are, its vocabulary is much smaller. To mitigate the impact of out-of-vocabulary (OOV) items on their tests, the authors introduce a new grammaticality test suite, Zorro,[4] in the style of BLiMP. Sentence pairs are generated for one paradigm each for 11 of BLiMP's 12 phenomena, along with two additional phenomena. However, we show that the Zorro sentences are not only lexically simpler as intended, but their templates are also far less complex and even less varied than the sentences in the corresponding BLiMP phenomena. Full lists of paradigms for each data set can be found in the Appendix, and the full data sets themselves are made available by the benchmarks' original authors.

3.1 Linear Baselines

As noted earlier, BLiMP and Zorro tests are stochastically generated with category-based templates. This way, a large number of examples can be generated and tested, but the drawback is that all examples are essentially the same structure. Moreover, many of the structures are simple, falling considerably below the coverage of modern syntactic analyses. In fact, many examples appear solvable by strictly linear methods. The observation that such template-generated examples can be solved this way is not new to to field. For example, Kam et al. (2008) demonstrated that a bigram model will predict the grammatical sentence from template-produced pairs featuring auxiliary inversion (a structural phenomenon) as well as neural models of the time.

To take an example from BLiMP, within its SUBJECT-VERB AGR phenomenon, four of six

[3] https://github.com/alexwarstadt/blimp

[4] https://github.com/phueb/Zorro/

paradigms evaluate string-adjacent subject verb agreement that could be captured by a bigram model. The remaining two include intervening distractor nouns, but in both these and the string-adjacent paradigms, the target noun is consistently the first/leftmost noun. A single linear rule, albeit a long-distance one, is sufficient to succeed on this phenomenon. In ANAPHORA AGREEMENT, none of the sentences has any distractors at all: the test is solely about whether the anaphor (e.g., *himself/herself*) agrees with the first, and only, noun in the sentence preceding it. Success on such simple tests tells us little about the genuine grammatical capacity of LMs and distorts or dilutes summary metrics calculated over the benchmark.

We evaluate this problem quantitatively with two studies of linear rules that do not incorporate structural knowledge. We find that many, but certainly not all, paradigms are solvable with non-human-like linear approaches. These paradigms therefore do not contribute to the overall goal of evaluating whether an LM possesses linguistic knowledge. Additionally, we find that the paradigms of Zorro tend to be structurally even simpler and less internally varied than the parallel paradigms of BLiMP. It is a weaker benchmark even when accounting for the intended lexical simplicity.

3.1.1 *N*-Gram Models

The original BLiMP paper reports the accuracy of a 5-gram model trained on the 3.1B token Gigaword Corpus (Graff et al., 2003) in addition to three neural LMs and human performance. They find that the 5-gram model scores above chance (50%) on all but two phenomena but is outclassed by most of the neural LMs on most paradigms. Performance for all LMs can vary widely across paradigms within one phenomenon. In some cases, there is a clear split between the 5-gram and neural models, suggesting that the latter capture some structural property of the paradigm that the 5-gram model does not, but in other cases, the 5-gram model performs well, demonstrating that linear rules can be sufficient for completing those tasks.

Revisiting SUBJECT-VERB AGR. as an illustrative example, the Gigaword 5-gram model performs only slightly behind the neural models on each string-adjacent paradigm but far below chance in the distractor paradigms. However, the neural models also perform up to 20.5 points better in the adjacent paradigms than the distractor paradigms. The two distractor paradigms demonstrate that the neural models have learned a long-distance pattern (whether that be structural or "agree with the leftmost noun"), but the adjacent paradigms cannot show this. They, and about half of the BLiMP paradigms, are uninformative in this way.

We extend this approach to the language acquisition setting by training a 5-gram model only on AO-CHILDES and evaluating on both BLiMP and Zorro. We compare these results to BabyBERTa on these data sets.[5] To further manage lexical effects while adding minimal complexity to the model, we evaluate both a 5-gram word model (5-word), and a 5-gram model trained only on POS tags (5-tag). AO-CHILDES was tagged using GPoSTTL, a rule-based POS tagger with tokenizer and lemmatizer based on the Brill Tagger (Brill, 1992). This was used to train sklearn's CRF POS-tagger, which was then used to label the benchmark corpora. This approach was taken to avoid bringing additional knowledge from a tagger trained on larger corpora into the benchmark corpora. The downside is that the tagger is not particularly accurate on the ungrammatical benchmark sentences, which may hurt performance for the 5-tag model. In addition to the 5-word and 5-tag models, we evaluate an oracle which marks a correct prediction if either 5-word or 5-tag makes a correct prediction. Our use of POS is motivated from a developmental perspective. Syntactic categories can be formed purely distributionally as early as infancy (Mintz, 2003; Shi and Melançon, 2010; Reeder et al., 2013) and children almost never make mistakes in their use of syntactic categories (Valian, 1986). It is thus plausible to assume that the acquisition of grammatical knowledge builds on a developmentally prior stage of syntactic category learning.

The results of the 5-gram experiments are summarized in Table 1 and laid out in detail in the Appendix. We draw three conclusions from these. First, the 5-gram models perform surprisingly well relative to the BabyBERTa transformer despite its extremely non-human-like simplicity when trained on the same AO-CHILDES data. Either 5-word or 5-tag, trained on the same data as BabyBERTa, outperformed BabyBERTa on 11 of 23 Zorro paradigms and 21 of 67 BLiMP paradigms. BabyBERTa's performance appears less impressive when presented alongside even this very weak

[5] Refer to Appendix for full details. We downloaded the publicly available model checkpoints from the BabyBERTa GitHub repository and replicated the BLiMP and Zorro results hosted on the Zorro GitHub repository

Zorro	BabyBERTa	5-Word	5-Tag	Either	Oracle
# Best	–	8/23	8/23	11/23	14/23
Avg Acc	78.91%	63.44%	57.59%	–	83.43%
BLiMP	**BabyBERTa**	**5-Word**	**5-Tag**	**Either**	**Oracle**
# Best	–	18/67	10/67	23/67	48/67
Avg Acc	60.72%	50.72%	37.93%	–	68.32%

Table 1: Performance summaries for 5-grams relative to BabyBERTa on Zorro and BLiMP. Number of paradigms in which a 5-gram model outperforms BabyBERTa and overall average accuracy across paradigms are reported. Either = either 5-word or 5-tag outpeformed BabyBERTa on the entire paradigm. Oracle = sentence pairs were marked correct if either 5-word or 5-tag made the correct prediction.

baseline. The AO-CHILDES 5-gram models perform more poorly on BLiMP than the Gigaword 5-gram model, but it still achieves high accuracy on several paradigms scattered across the phenomena.

Second, 5-gram oracle outperforms 5-word, 5-tag, and BabyBERTa. The 5-gram oracle is not a fair direct comparison but provides a summary metric for correlation between 5-word and 5-tag. A high oracle score relative to the two 5-gram models indicates that they do not make the same errors. That is, errors are not necessarily attributable to the string-local limitations of 5-grams *per se* but rather to 5-gram sparsity or errors in tagging. The high oracle score is another sign that the paradigms often capture surface properties rather than structural properties that would stump 5-gram models.

Third, the 5-gram models outperform BabyBERTa on proportionately more Zorro paradigms than BLiMP paradigms. Additionally, the AO-CHILDES 5-word model achieved 78.91% performance on Zorro, while the Gigaword 5-gram model only reached 60.5% on BLiMP. If Zorro were merely accounting for the smaller vocabulary in the AO-CHILDES training data, we should expect much more similar performance on both of these measures. Taken together, these suggest that Zorro is a substantially weaker benchmark that BLiMP, and it more greatly overestimates the apparent positive results of the acquisition-inspired BabyBERTa.

3.1.2 Hand-Written Simple Rules

In addition to reporting results on 5-gram models, we created simple hand-written rules which demonstrate that the probes are solvable in principle without reference to linguistic structure. While we do not claim that such rules are akin to the state of knowledge in LMs, it is also difficult to completely rule out this possibility. On the one hand, it is still unclear how to interpret the representation of linguistic knowledge in LMs. On the other, the vast majority of training data, at least child-directed for language acquisition, is structurally simple and can in fact be handled by rule-like pattern matchers. In English CDS, the distribution of anaphora is exceedingly straightforward: almost all instances of *himself* are preceded in the sentence by the subject pronoun *he* and a (male) noun phrase with no co-referential competitors. For comparison, Zorro `adjunct_island` can be solved perfectly by always selecting the sentence where the third-last word is *the*, and many of the paradigms can be solved by tracking the index of a specific word. Others can be solved by checking for the presence of a certain word. For example, the `superlative` paradigm can be solved by accepting the sentence that contains either *more* or *fewer*. For both Zorro and BLiMP, more than one paradigm can often be solved with the exact same rule. We write simple linear rules for each Zorro and BLiMP paradigm. See the Appendix for a full list of rules.

In summary, these rules yielded 93.97% accuracy on Zorro and solved 14 of 23 Zorro paradigms with 100% accuracy. Each `agreement_` paradigm is solved with at least 96% accuracy, with the remainder due to two irregular nouns, *feet* and *children*, which do not end in the *-s* referenced by these rules. The lowest performance is 52.75% on `anaphor_agreement-pronoun_gender`, a paradigm that requires an LM to 'know' the canonical gender of English names in order to choose *himself* or *herself*. The test sentence pairs were not quite balanced, so always guessing *himself* earns more than 50%.

BLiMP proved more challenging. The rules only yielded 84.35% accuracy on average and achieved perfect scores on 14 of 67 rules. The overall high score of the hand-written simple linear rules suggests that BLiMP suffers from the same issues regarding lack of sentence variety that Zorro does, but the lower accuracy indicates that the problem is not quite as severe. In principle, we could have composed more complex rules which achieved perfect accuracy on all paradigms, however, these simpler rules better illustrate our points. The success of non-human-like simple linear rules on most paradigms on both benchmarks further emphasizes that success on the template-based behavioral task does not necessarily imply that an LM possesses linguistic knowledge.

Sentence ID	Sentence	ME Z-score
32.3.Culicover.7a.g.01	John tried to win.	1.453262
32.3.Culicover.7b.*.01	John tried himself to win.	-0.86729
33.2.bowers.7b.g.07	Sarah counted the change accurately.	1.230412
33.2.bowers.7b.*.07	Sarah accurately counted the change.	1.20698
ch8.150.*.01	Melissa seems that is happy.	-1.14131
ch8.151.g.01	It seems that Melissa is happy.	1.000644
ch8.152.g.01	Melissa seems to be happy.	1.196088

Table 2: Top: Two pairwise phenomena from the Linguistic Inquiry (LI) dataset. Bottom: One multi-condition phenomenon from the Adger dataset. The ME Z-score is the averaged Z-score transformation of the human Magnitude Estimation judgments for each of the sentences across all the experimental participants.

4 An Alternative: The LI-Adger Dataset

The LI-Adger dataset is a comprehensive collection of 519 sentence types, 300 collected by Sprouse et al. (2013) from *Linguistic Inquiry (LI) 2001-2010*,[6] a major theoretical journal in linguistics, and 219 collected by Sprouse and Almeida (2012) from Adger's (2003) *Core Syntax* textbook.[7] Each sentence type includes eight hand-constructed, semantically plausible sentences, assembled into 150 pairwise (LI) and 105 multi-condition (Adger) phenomena where each minimal pair is lexically matched. We provide an example of each in Table 2.

The LI-Adger dataset improves upon the prior two datasets in three key ways. Firstly, unlike BLiMP and Zorro, the LI-Adger sentences are controlled for semantic implausibility, which has been shown to be a strong confounding factor when eliciting human judgments (Sprouse et al., 2018). Second, the 255 total pairwise and multi-condition phenomena achieve much more diverse coverage of syntactic phenomena than the 67 paradigms in BLiMP, and the 23 paradigms in Zorro. Third, the human judgments were collected using the Magnitude Estimation (ME) task (and Likert Scale (LS) in the case of the LI sentences) in addition to Forced-Choice (FC) as in the BLiMP human baselines. We believe this to be a crucial advantage because the FC task treats sentence acceptability as functionally categorical: A sentence is only acceptable or not relative to its minimal pair counterpart, whereas tasks such as ME allow us to make comparisons within and across minimal pairs, thereby treating sentence acceptability as a gradient measure.

With this dataset, we conduct the following two tests. First, in line with Vázquez Martínez (2021), we sort the LI-Adger dataset into 2391 unique minimal pairs. We then collect pseudo log-likelihood scores for each sentence from several models evaluated by Huebner et al. (2021), and score them using the same criteria as BLiMP and Zorro. As a baseline for the models, we include Log-Likelihood and Syntactic Log-Odds Ratio (SLOR; Pauls and Klein, 2012; Lau et al., 2017) scores by a trigram model trained on the British National Corpus (BNC; 100M words) by Sprouse et al. (2018).

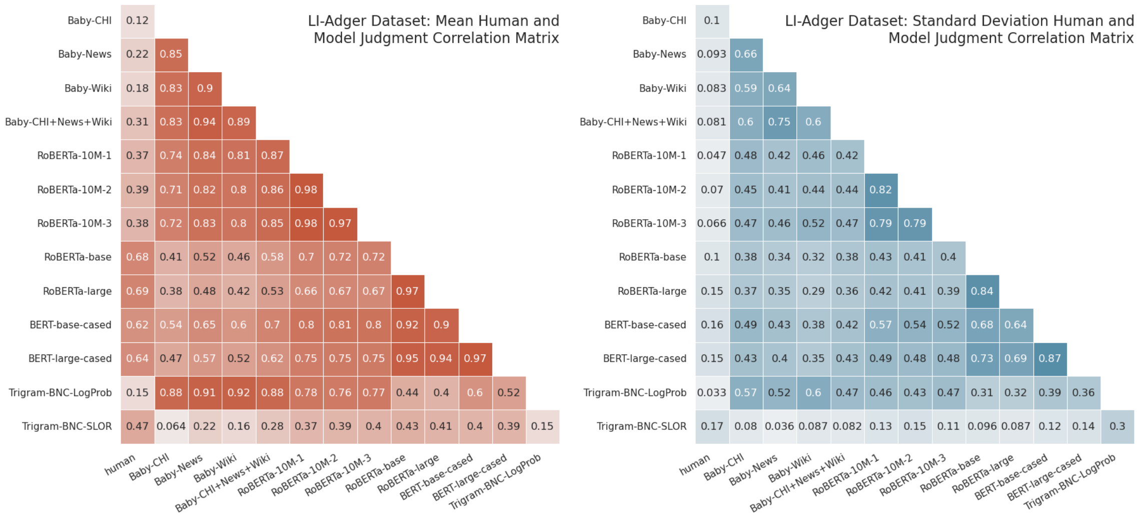
We include the results of this test in Figure 1. We observe that all models are further from the human baseline as compared to those in BLiMP (no human baselines were reported for Zorro). But more importantly, we observe that the trigram model scored using SLOR performs on par with the BabyBERTa models and approaches the performance of RoBERTa (Liu et al., 2019) trained on 10M words. If we were to adopt the "positive results from model learners are more meaningful than negative results" argument, then the trigram model is as suitable a model of language acquisition as BabyBERTa is.

Raw accuracy notwithstanding, we proceed to conduct a novel test of judgment variability on our collection of LMs. We take advantage of the structure of the LI-Adger dataset in the following way: There are 519 sentence types, and for each type there are eight sentences that retain the same syntactic structure but vary lexical items at the locus of the syntactic structure tested (e.g., the head of a verb phrase from which extraction takes place). These datasets thus allow us to contrast the consistency of human judgment across and within construction types against that of the LMs.

We z-score the LM judgments to make them comparable to the human judgments. Then, for each set of eight sentences, we take the mean and standard deviation of all the judgments for humans and each LM. We find that the models are much more variable in their judgments: The human judgments, on average, vary by 0.288 standard devia-

[6] https://www.jonsprouse.com/data/Lingua2013/SSA.data.zip

[7] https://www.jonsprouse.com/data/JoL2012/SA2012.data.zip



LI-Adger Dataset: Mean Human and Model Judgment Correlation Matrix

	human	Baby-CHI	Baby-News	Baby-Wiki	Baby-CHI+News+Wiki	RoBERTa-10M-1	RoBERTa-10M-2	RoBERTa-10M-3	RoBERTa-base	RoBERTa-large	BERT-base-cased	BERT-large-cased	Trigram-BNC-LogProb
Baby-CHI	0.12												
Baby-News	0.22	0.85											
Baby-Wiki	0.18	0.83	0.9										
Baby-CHI+News+Wiki	0.31	0.83	0.94	0.89									
RoBERTa-10M-1	0.37	0.74	0.84	0.81	0.87								
RoBERTa-10M-2	0.39	0.71	0.82	0.8	0.86	0.98							
RoBERTa-10M-3	0.38	0.72	0.83	0.8	0.85	0.98	0.97						
RoBERTa-base	0.68	0.41	0.52	0.46	0.58	0.7	0.72	0.72					
RoBERTa-large	0.69	0.38	0.48	0.42	0.53	0.66	0.67	0.67	0.97				
BERT-base-cased	0.62	0.54	0.65	0.6	0.7	0.8	0.81	0.8	0.92	0.9			
BERT-large-cased	0.64	0.47	0.57	0.52	0.62	0.75	0.75	0.75	0.95	0.94	0.97		
Trigram-BNC-LogProb	0.15	0.88	0.91	0.92	0.88	0.78	0.76	0.77	0.44	0.4	0.6	0.52	
Trigram-BNC-SLOR	0.47	0.064	0.22	0.16	0.28	0.37	0.39	0.4	0.43	0.41	0.4	0.39	0.15

LI-Adger Dataset: Standard Deviation Human and Model Judgment Correlation Matrix

	human	Baby-CHI	Baby-News	Baby-Wiki	Baby-CHI+News+Wiki	RoBERTa-10M-1	RoBERTa-10M-2	RoBERTa-10M-3	RoBERTa-base	RoBERTa-large	BERT-base-cased	BERT-large-cased	Trigram-BNC-LogProb
Baby-CHI	0.1												
Baby-News	0.093	0.66											
Baby-Wiki	0.083	0.59	0.64										
Baby-CHI+News+Wiki	0.081	0.6	0.75	0.6									
RoBERTa-10M-1	0.047	0.48	0.42	0.46	0.42								
RoBERTa-10M-2	0.07	0.45	0.41	0.44	0.44	0.82							
RoBERTa-10M-3	0.066	0.47	0.46	0.52	0.47	0.79	0.79						
RoBERTa-base	0.1	0.38	0.34	0.32	0.38	0.43	0.41	0.4					
RoBERTa-large	0.15	0.37	0.35	0.29	0.36	0.42	0.41	0.39	0.84				
BERT-base-cased	0.16	0.49	0.43	0.38	0.42	0.57	0.54	0.52	0.68	0.64			
BERT-large-cased	0.15	0.43	0.4	0.35	0.43	0.49	0.48	0.48	0.73	0.69	0.87		
Trigram-BNC-LogProb	0.033	0.57	0.52	0.6	0.47	0.46	0.43	0.47	0.31	0.32	0.39	0.36	
Trigram-BNC-SLOR	0.17	0.08	0.036	0.087	0.082	0.13	0.15	0.11	0.096	0.087	0.12	0.14	0.3

Figure 2: Correlation matrices of human judgments and LM output means (top) and standard deviations (bottom) on each sentence type on the LI-Adger dataset. Baby=BabyBERTa, CHI=AO-CHILDES, News=AO-NEWSELA, Wiki=Wikipedia-1.

tion (std. dev.) units within a given set of sentences. On the other hand, the LM that least varies is BabyBERTa Wiki, varying by 0.451 std. dev. units on average. The rest of the models nearly double the variability of the human judgments, ranging from 0.518 for RoBERTa-10M-1 to 0.554 for BERT-large-cased. Variability appears to increase rather than decrease as training size and performance increase. Surprisingly, the trigram model, when scored using log probabilities, is the closest in variability to the human judgments at 0.331 std. dev. units, but also the furthest when scored using SLOR with a variability of 0.599. Once again we find that a positive result on one test or another is not enough to draw positive conclusions.

For further illustration, we correlate the means and standard deviations of 512 sentence types across each LM and humans and plot the results in Figure 2. Both in terms of mean and standard deviations, we observe generally high correlations between the various neural LMs, but much lower correlations between the LMs and humans. This suggests that whatever the LMs are doing, good or bad, does not appear to be human. Interestingly, the BabyBERTa LMs show very high correlations with the naive trigram log-likelihood scores and very low with trigram SLOR scores, raising further suspicions that these small acquisition-inspired LMs behave like a very non-human-like model.

5 Discussion

It is widely recognized that children acquire language in ways that appear quite different from LM training. There is a growing realization that the cognitive relevance of LMs can only be established in a comparable setting. Bringing down training size requirements stands not to not only improve the applicability of such models to the study of language acquisition but also to efficient NLP on low-resource languages.

However, in this paper, we observed several weaknesses in BLiMP and Zorro, two minimal pair benchmarks for evaluating the linguistic knowledge of neural language models. We believe that it is worth critically revisiting the underlying assumption that positive results on such benchmarks are a demonstration of human-like representations or human-like language acquisition, especially if an evaluation can be solved in unintended ways, or if it does not reflect an adequately broad range of linguistic structures. It is unlikely that a behavioral probe, such as these large binarized benchmarks, can fully capture the complexity of linguistic knowledge. To this end, we made a case for also evaluating with curated benchmarking datasets: the gradient acceptability judgments from human subjects makes these effective probes for the structural basis of grammar. Together with a range of tests, from carefully constructed tests of grammaticality to probes correlating the internal state of LMs with their predictions need to complement theoretical, psycholinguistic, and neurolinguistic studies before a meaningful cognitive claim about the nature of neural language models can be made.

We end with some broader discussion about language acquisition and the cognitive interpretation of computational models. While it is now widely recognized that children learn language with only a fraction of the data needed for large LM training, merely reducing the amount of training data alone – such as the 100M word threshold in the BabyLM Challenge – still falls short of the requirement for an adequate model of language acquisition. While it is true that a native speaker's knowledge of language can be established on the basis of approximately 100 million words, child language research makes clear that not all aspects of linguistic knowledge are learned at the same time. Some, such as inflectional morphology, case marking, word order, and major transformations are acquired very early in all languages studied so far (e.g., Brown, 1973; Slobin, 2022) at an order of magnitude fewer words of input, while others are learned rather late: These include derivational morphology (Jarmulowicz, 2002), passivization (Pinker et al., 1987), control and cleft structures (Chomsky, 1969) and the dative constructions (Gropen et al., 1989) in the case of English, but these may emerge much earlier in other languages. This suggests that successful learning in the limit (e.g., 100M word) is not sufficient. For example, while a neural model of English past tense (Kirov and Cotterell, 2018) eventually learns the "add -ed" rule, it does so with over 3,000 verb lemmas. By contrast, children learn that rule before or around 3 (Kuczaj, 1977), when their vocabulary only contains around 300 or so verbs (Marcus et al., 1992). To serve as cognitive models of language, it is thus important to compare the training trajectory of LMs as a function of the training data volume against the developmental benchmarks of specific linguistic phenomena which have been amply documented in the empirical literature on child language.

Limitations

Our study is about the limitations of evaluation, so it is to be expected that our study has its limits as well. Most obviously, ours and any study would benefit from testing and reporting on a wider range of neural models and a wider range of baselines. And like most work in this area, our evaluations were only performed on English. We recommend the use of the LI-Adger data set. Like any behavioral probe, including the ones which we criticize, it can be subject to ambiguous interpretation. It has some substantial advantages, as we discuss in this paper, but also a couple of additional drawbacks. It is smaller than BLiMP or Zorro, and it has not been annotated by phenomenon. Nevertheless, it provides additional insights that those benchmarks do not. As in the paper, we recommend its use in conjunction with a wide range of other evaluation methods.

References

David Adger. 2003. *Core syntax: A minimalist approach*, volume 20. Oxford University Press Oxford.

Marc H Bornstein, Linda R Cote, Sharone Maital, Kathleen Painter, Sung-Yun Park, Liliana Pascual, Marie-Germaine Pêcheux, Josette Ruel, Paola Venuti, and Andre Vyt. 2004. Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child development*, 75(4):1115–1139.

Eric Brill. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, ANLC '92, page 152–155, USA. Association for Computational Linguistics.

Roger Brown. 1973. *A first language: The early stages.* Harvard University Press, Cambridge, MA.

Wei-Lun Chao, Hexiang Hu, and Fei Sha. 2018. Being negative but constructively: Lessons learnt from creating better visual question answering datasets. 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 431–441, New Orleans, Louisiana. Association for Computational Linguistics.

Carol Chomsky. 1969. *The acquisition of syntax in children from 5 to 10.* MIT Press.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Katherine Demuth, Jennifer Culbertson, and Jennifer Alter. 2006. Word-minimality, epenthesis, and coda licensing in the acquisition of English. *Language and Speech*, 49(2):137–173.

Larry Fenson, Philip S Dale, J Steven Reznick, Elizabeth Bates, Donna J Thal, Stephen J Pethick, Michael Tomasello, Carolyn B Mervis, and Joan Stiles. 1994. Variability in early communicative development. *Monographs of the Society for Research in Child Development*, pages i–185.

Michael C Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. 2021. *Variability and* *consistency in early language learning: The Wordbank project*. MIT Press, Cambridge, MA.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Jess Gropen, Steven Pinker, Michelle Hollander, Richard Goldberg, and Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in english. *Language*, pages 203–257.

Olivia Guest and Andrea E Martin. 2023. On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, pages 1–15.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. 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 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Betty Hart and Todd R Risley. 1995. *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing, Baltimore, MD.

John Hewitt and Christopher D. Manning. 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)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, 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.

Philip A Huebner and Jon A Willits. 2021. Using lexical context to discover the noun category: Younger children have it easier. In *Psychology of learning and motivation*, volume 75, pages 279–331. Elsevier.

Linda D Jarmulowicz. 2002. English derivational suffix frequency and children's stress judgments. *Brain and Language*, 81(1-3):192–204.

Xuân-Nga Cao Kam, Iglika Stoyneshka, Lidiya Tornyova, Janet D Fodor, and William G Sakas. 2008. Bigrams and the richness of the stimulus. *Cognitive science*, 32(4):771–787.

Evan Kidd, Seamus Donnelly, and Morten H Christiansen. 2018. Individual differences in language acquisition and processing. *Trends in cognitive sciences*, 22(2):154–169.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Stan A. Kuczaj. 1977. The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, 16(5):589–600.

William Labov. 2012. What is to be learned. *Review of Cognitive Linguistics*, 10(2):265–293.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016a. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016b. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

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*.

Brian MacWhinney. 1991. *The CHILDES language project: Tools for analyzing talk*. Lawrence Erlbaum, Mahwah, NJ.

Gary F Marcus, Steven Pinker, Michael Ullman, Michelle Hollander, T John Rosen, Fei Xu, and Harald Clahsen. 1992. Overregularization in language acquisition. *Monographs of the society for research in child development*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 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.

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.

Toben H Mintz. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91–117.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. Association for Computational Linguistics.

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968.

Steven Pinker, David S Lebeaux, and Loren Ann Frost. 1987. Productivity and constraints in the acquisition of the passive. *Cognition*, 26(3):195–267.

Patricia A Reeder, Elissa L Newport, and Richard N Aslin. 2013. From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive psychology*, 66(1):30–54.

Rushen Shi and Andréane Melançon. 2010. Syntactic categorization in french-learning infants. *Infancy*, 15(5):517–533.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Dan Isaac Slobin. 2022. *The Crosslinguistic Study of Language Acquisition: Volume 3*, volume 3. Psychology Press.

Jon Sprouse and Diogo Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's core syntax. *Journal of Linguistics*, 48(3):609–652.

Jon Sprouse and Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa*, 2(1):1.

Jon Sprouse and Norbert Hornstein. 2013. *Experimental syntax and island effects*. Cambridge University Press, Cambridge.

Jon Sprouse, Carson T Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, 134:219–248.

Jon Sprouse, Beracah Yankama, Sagar Indurkhya, Sandiway Fong, and Robert C Berwick. 2018. Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *The Linguistic Review*, 35(3):575–599.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Mycal Tucker, Peng Qian, and Roger Levy. 2021. What if this modified that? syntactic interventions with counterfactual embeddings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 862–875, Online. Association for Computational Linguistics.

Virginia Valian. 1986. Syntactic categories in the speech of young children. *Developmental psychology*, 22(4):562.

Héctor Vázquez Martínez. 2021. The acceptability delta criterion: Testing knowledge of language using the gradience of sentence acceptability. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 479–495, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022. Identifying and mitigating spurious correlations for improving robustness in NLP models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1719–1729, Seattle, United States. Association for Computational Linguistics.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*, pages 17–60. CRC Press.

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

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

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 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*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Appendix

Phenomenon	Paradigm	BabyBERTa	5-Gram			Simple
		AO-CHILDES	Word	Tag	Oracle	Rule
agreement_subject_verb	across_rel_clause	64.85	50.95	46.35	**68.95**	**96.20**
	in_simple_question	92.35	61.15	90.9	**93.9**	**98.30**
	in_question_with_aux	90.85	59	80.15	**90.9**	**98.05**
	across_prep_phrase	72.85	50	50	62.6	**98.40**
agreement_determiner_noun	between_neighbors	91.3	83.05	49.85	88.6	**98.60**
	across_1_adjective	89.85	50.45	50.05	75.05	**97.20**
filler-gap	wh_question_object	98.75	42.8	**100**	**100**	**100**
	wh_question_subject	75.7	**88.3**	**76.55**	**97.1**	**100**
island-effects	coord_struct_constr	97.05	43.35	55.6	83.85	**100**
	adjunct_island	56.15	**66.1**	**58.8**	**83.85**	**100**
quantifiers	existential_there	92.9	80.25	38.4	89.55	**100**
	superlative	64.55	45.1	**82**	**96.05**	**100**
npi_licensing	only_npi_licensor	74.1	**79.4**	3.7	**79.4**	**100**
	matrix_question	65.25	47.5	28.65	58	**100**
argument_structure	swapped_arguments	91	**92.15**	81.7	**98.85**	**100**
	transitive	60.05	**64.15**	32.65	**78.6**	58.05
	dropped_argument	79.9	**85.05**	**83.6**	**95.75**	**100**
irregular	verb	69.65	62.9	**93.6**	**96.35**	**88.40**
anaphor_agreement	pronoun_gender	51.75	49.15	1.95	50.95	**52.75**
ellipsis	n_bar	55.3	**66.6**	**63.6**	**89.9**	**100**
binding	principle_a	89.4	45.9	3.6	47.75	**100**
case	subjective_pronoun	94.7	**99.55**	**97.95**	**100**	**100**
local_attractor	in_question_with_aux	96.65	55.65	95	**99.05**	**100**
AVERAGE		78.91%	63.44%	57.59%	83.43%	93.97%
Fraction ≥ BabyBERTa		–	8/23	8/23	14/23	22/23

Table 3: Word and tag-level 5-gram models trained on AO-CHILDES plus 5-Gram Oracle and Simple Linear Rule Oracle for Zorro. 5-Gram and Simple Rule scores that are greater than BabyBERTa_AO-CHILDES are bolded

Phenomenon	Paradigm	Rule
agreement_subject_verb	across_rel_clause	2nd word ends in *s* iff 3rd last is in {*are*, *were*, *do*}
	in_simple_question	Word 2 right of {*are*, *were*} ends in *s*. Word 2 right of {*is*, *are*} does not
	in_question_with_aux	4th word ends in *s* iff 2nd is in {*are*, *were*, *do*}
	across_prep_phrase	2nd word ends in *s* iff 3rd last is in {*are*, *were*, *do*}
agreement_determiner_noun	between_neighbors	If {*these*, *those*} in sentence, next word ends in *s*. If {*this*, *that*} in sentence, next word does not
	across_1_adjective	If {*these*, *those*} in sentence, word 2 right ends in *s*. If {*this*, *that*} in sentence, word 2 right does not
filler-gap	wh_question_object	2nd word is *the*
	wh_question_subject	*who* does not immediately precede *the*
island-effects	coord_struct_constr	4th word is *and*
	adjunct_island	3rd last word is *the*
quantifiers	existential_there	Contains one of {*many*, *some*, *no*, *few*, *a*, *an*}
	superlative	Contains one of {*more*, *fewer*}
npi_licensing	only_npi_licensor	1st word is *only*
	matrix_question	Contains one of {*does*, *will*, *should*, *could*, *did*, *wouldo*}
argument_structure	swapped_arguments	1st word is *the*
	transitive	2nd last word does not end in *e*
	dropped_argument	1st word is *the*
irregular	verb	word following *had* ends in *n* or no word ends in *n*
anaphor_agreement	pronoun_gender	Sentence contains *himself*
ellipsis	n_bar	*Sentence where *and* appears farther right
binding	principle_a	4th last word ends with *ing*
case	subjective_pronoun	1st word is *the*
local_attractor	in_question_with_aux	4th word does not end with *'s*

Table 4: Simple Linear Rule descriptions for Zorro. Rules that require sentences to be compared are marked with an asterisk.

Phenomenon	Paradigm	BabyBERTa	5-Gram			Simple
		AO-CHILDES	Word	Tag	Oracle	Rule
anaphor_agreement	anaphor_gender_agreement	65.6	26.3	8	33.9	**73.9**
	anaphor_number_agreement	73.7	52.9	5.7	55.5	**80.1**
argument_structure	causative	58.5	55.2	30.7	**68.8**	**85.6**
	drop_argument	63.2	50.9	52.9	**80.8**	**77.1**
	inchoative	50.7	**56**	37.1	**73.8**	**57.1**
	intransitive	52.1	48.2	49.6	**76.3**	**73.55**
	passive_1	50.2	**52.1**	12.9	**56.4**	**59.5**
	passive_2	54	**48.4**	18.1	**56.8**	**59.6**
	transitive	55.3	51.6	36.1	**67.6**	**57.85**
binding	principle_A_case_1	43.6	**100**	7.1	**100**	**100**
	principle_A_case_2	99.9	41.5	13	48.3	99.2
	principle_A_c_command	58.7	35.7	4.2	38.1	**71.35**
	principle_A_domain_1	96.5	38.4	3.1	40.7	**100**
	principle_A_domain_2	51.4	**61.7**	2.7	**62.8**	**58.3**
	principle_A_domain_3	46.8	44.5	29.7	**61.1**	**50.4**
	principle_A_reconstruction	40.9	32.1	**53.9**	**68**	**74.1**
control_raising	existential_there_object_raising	59.1	30.5	23.4	46.5	**67.95**
	existential_there_subject_raising	51	43.4	17	**53.6**	**77**
	expletive_it_object_raising	63.3	61.2	48.3	**79.6**	**69.5**
	tough_vs_raising_1	72.2	59.1	49.6	**83.2**	**87.1**
	tough_vs_raising_2	34.4	**41.3**	18.4	**54.1**	**92.5**
determiner_noun_	a_irregular_1	66.6	48.8	37.4	61.3	**68.45**
agreement = a	a_irregular_2	87.4	74.3	12.3	77.1	73.7
	a_with_adjective_1	76.3	48.2	49.7	63.8	**95.95**
	a_with_adj_irregular_1	82.9	49	49.7	56.3	74.45
	a_with_adj_irregular_2	67	49.5	18.3	58.2	**71.8**
	a_with_adj_2	80.4	49.8	19.9	59.7	**95.6**
	a_1	72.2	64.1	48.1	**74.5**	**95.55**
	a_2	87.4	65.2	11	68.1	**96.75**
ellipsis	ellipsis_n_bar_1	58.7	**64.1**	**63.5**	**86.4**	**85.65**
	ellipsis_n_bar_2	42.8	39.9	**70.5**	**80.9**	**99.95**
filler_gap_dependency	wh_questions_object_gap	73	37	**82.4**	**89.2**	**99.95**
	wh_questions_subject_gap	79.9	49	**81.4**	**89.4**	**99.9**
	wh_vs_that_no_gap	90.9	77.2	83.8	**94.9**	**99.95**
	wh_vs_that_no_gap_long_distance	92.1	74.9	87	**95.8**	**99.7**
	wh_vs_that_with_gap	29.1	22.7	15	**33**	**100**
	wh_vs_that_with_gap_long_distance	14.9	**25.8**	12.8	**32.8**	**99.9**
irregular_forms	irregular_past_participle_adjectives	59.8	**99.4**	12.2	**99.4**	**100**
	irregular_past_participle_verbs	59.8	**99.4**	12.2	**99.4**	**100**
island_effects	adjunct_island	63.8	58.4	55.5	**82.5**	**94.5**
(coordinate_structure_	y_complex_left_branch	36.2	11.8	19.6	26.9	**97.05**
constraint = y)	y_object_extraction	56.5	41.9	37.1	**63.7**	**86.35**
	left_branch_island_echo_question	52.4	16.3	30.1	38.7	**100**
	left_branch_island_simple_question	66.6	24.5	30.3	43.8	**97.9**
	sentential_subject_island	46.1	37.3	42.8	**62.9**	**82.65**
	wh_island	47.1	**69**	**93.4**	**97.3**	**100**
npi_licensing	matrix_question_npi_licensor_present	56.4	41.1	39.5	**65.7**	**97.4**
	npi_present_1	27	**56**	26.7	**69.6**	**100**
	npi_present_2	20.3	**56.4**	**25.8**	**70.5**	**100**
	only_npi_licensor_present	71.6	**98.4**	2.4	**98.5**	**100**
	only_npi_scope	72.1	**80.4**	**79.4**	**97.2**	**100**
	sentential_negation_npi_licensor_present	73.8	**100**	0	**100**	**100**
	sentential_negation_npi_scope	81.9	40	65.3	79.6	**100**
quantifiers	existential_there_quantifiers_1	93.7	79.1	26.4	87.4	**97.3**
	existential_there_quantifiers_2	35.7	19.6	**36**	**50.6**	**96.85**
	superlative_quantifiers_1	49.5	**73**	**89.8**	**96.4**	**100**
	superlative_quantifiers_2	61.2	51.9	0.1	52	**100**
s-selection	animate_subject_passive	45.5	**48.4**	24	**58.4**	**65.25**
	animate_subject_trans	59.7	50	57.1	**78.2**	**84.65**
subject_verb_agreement	distractor_agreement_relational_noun	29	26.2	21.4	**42.1**	**50.25**
	distractor_agreement_relative_clause	35.6	28.3	30.4	**49.8**	**55.85**
	irregular_plural_subject_verb_agreement_1	67.9	33.4	51.7	62.5	53.2
	irregular_plural_subject_verb_agreement_2	66.2	51	51.9	**70.7**	59.3
	regular_plural_subject_verb_agreement_1	68.8	39.9	51.1	**72**	64.35
	regular_plural_subject_verb_agreement_2	60.1	51	55.6	**76.9**	**73.15**
AVERAGE		60.72%	50.72%	37.93%	68.32%	84.35%
Fraction ≥ BabyBERTa		–	18/67	10/67	48/67	62/67

Table 5: Word and tag-level 5-gram models trained on AO-CHILDES plus 5-Gram Oracle and Simple Linear Rule Oracle for BLiMP. 5-Gram and Simple Rule scores that are greater than BabyBERTa_AO-CHILDES are bolded

Phenomenon	Paradigm	Rule
anaphor_agreement	anaphor_gender_agreement	Does not contain *itself*
	anaphor_number_agreement	Number of words that end in *s* is even
argument_structure	causative	Does not contain one of {*appear*, *vanish*, *exist*,
	transitive	*sigh*, *rust*, *cheer*, *clash*, *fall*, *fell*, *waste*}
	drop_argument	Last word is not one of {*to*,
	inchoative	*with*, *about*, *from*,
	intransitive	*at*, *through*, *by*, *like*}
	passive_1	None of {*communicat*, *suffer*, *compet*, *shout*, *laugh*,
	passive_2	*scream*, *complain*, *compromis*, *grin*, *chat*} in sentence
binding	principle_A_case_1	*Is the shorter of the two sentences
	principle_A_case_2	*Is the longer of the two sentences
	principle_A_	(Last word ends in *s* and (1st word is any of pl_det
	c_command	or the 2nd word is *lot*)) or 2nd to last word ends in *s*)
	principle_A_domain_1	*Is the shorter of the two sentences
	principle_A_domain_2	*Is the shorter of the two sentences
	principle_A_domain_3	Does not contain *that*
	principle_A_reconstruction	4th word does not end in *ed* nor *'t*
control_raising	a_obj_raising	Does not contain one of verb_set
	a_subj	Contains one of subj_words or {*appear*, *sure*,
(existential_	subj_raising	*threaten*, *look*}
there = a)	expletive_it_object_raising	Does not contain one of verb_set
	tough_vs_raising_1	Does not contain one of subj_words, nor *apt*
	tough_vs_raising_2	Contains one of subj_words, or *apt*
determiner_noun_	a_irregular_1	Does not end in *that* followed by (one of
agreement = b	b_irregular_2	{*people*, *women*, *men*, *children*} or a word ending
	b_with_adj_irregular_1	in *ses*) nor in {*those*, *these*} followed by (a word
	b_with_adj_irregular_2	ending in *is* or not with *s* at all)
	b_with_adjective_1	Does not end in *that* followed by a word ending in
	b_with_adj_2	a letter other than *i* followed by *s* nor in
	b_1	{*those*, *these*} followed by (a word ending in
	b_2	*is* or not with *s* at all)
ellipsis	ellipsis_n_bar_1	Last word in num_quant
	ellipsis_n_bar_2	Last word has already occurred in sentence
filler_gap_	wh_questions_object_gap	Does not contain *wh*
dependency	wh_questions_subject_gap	Does not contain *wh*
	wh_vs_that_no_gap = c	Does not contain *wh*
	c_long_distance	Does not contain *wh*
	wh_vs_that_with_gap = d	Contains *wh*
	d_long_distance	Contains *wh*
irregular_forms	irregular_past_	If 1st word is *the*, then 2nd word ends in *n*,
	part_adj	otherwise 2nd word must not end in *n*
	irregular_past_part_verbs	*Is the shorter of the two sentences
island_effects	adjunct_island	Last word is not *about* and does not end in *ing*
	e_complex_left_branch	2nd word is not in mod_aux
(coordinate_structure_	e_object_extraction	2nd to last word is not *and*
constraint = e)	f_echo_question	Does not start with *Wh*
(left_branch_	f_simple_question	2nd word is not in mod_aux
island = f)	sentential_subject_island	Ends in *ing* or *ed* or *with*
	wh_island	*wh*, capitalized or not, occurs twice in sentence
npi_licensing	matrix_question_g	1st word in mod_aux
	npi_present_1	Does not contain the word *ever*
(npi_licensor_	npi_present_2	Does not contain the word *ever*
present = g)	only_g	1st word is *only*
(npi_scope = h)	only_h	1st word is *only*
	sentential_negation_g	Does not contain the word *ever*
	sentential_negation_h	Does not contain the word *ever*
quantifiers	a_quantifiers_1	Does not contain {*each*, *most*, *all*, *every*} while
	a_quantifiers_2	also containing {*one-ten*}
	superlative_quantifiers_1	*Is the longer of the two sentences
	superlative_quantifiers_2	1st word is not *no*
s-selection	animate_subject_passive	Contains one of people_groups
	animate_subject_trans	*Is the shorter of the two sentences
subject_verb_agreement	i_relational_noun	*Is the longer of the two sentences
	i_relative_clause	The number of words that ends in *s* is odd
(distractor_	irregular_j_1	Contains no word ending in a letter other than *i* and
agreement = i)		followed by *s* that is followed by a word ending in *s*
(plural_subject_	irregular_j_2	None of {*people*, *women*, *men*, *children*} is
verb_agreement = j)		followed by a word ending in *s*
	regular_j_1	*Is the shorter of the two sentences
	regular_j_2	The number of words that ends in *s* is odd

Table 6: Linear Rule descriptions for BLiMP. Rules that require sentences to be compared are marked with an asterisk. Rules sometimes span across multiple rows. If one paradigm name is split across these rows, then the rule only corresponds to that paradigm. Otherwise the rule corresponds to all the paradigms listed in these rows. All variables (e.g. verb_set, subj_words) are defined in table 7.

Name	Content
`verb_set`	{*ask*, *press*, *entic*, *prod*, *obligat*, *convinc*, *badger*, *compel*, *sway*, *order*}
`subj_words`	{*certain*, *soon*, *likely*, *unlikely*, *bound*, *about*}
`num_quant`	{*one-ten*, *many*, *few*, *several*, *more*, *some*, *lot*, *fewer*}
`mod_aux`	{*had*, *should*, *is*, *was*, *can*, *has*, *will*, *would*, *could*, *do*, *does*, *might*, *were*, *did*}
`people_groups`	{*men*, *woman*, *children*, *teacher*, *lad*, *offspring*, *student*, *customer*, *girl*, *boy*}

Table 7: The sets of words represented by the variables used in table 6