Evaluating the Existence Proof: LLMs 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

July 17, 2024

Abstract

In recent years, the technological success of large language models (LLMs) has been taken as an existence proof that language acquisition may succeed without domain-specific principles and constraints. While this argument acknowledges the important differences between LLM training and child language acquisition, its validity rests on the validity of the existence proof itself, that LLMs indeed demonstrate capacity comparable to human linguistic knowledge, the terminal state of the acquisition process. We contend that such a proof has not been delivered, in large part due to the lack of rigor in LLM evaluation and the absence of serious engagement with the empirical study of child language. When trained on child-scale input data and evaluated on widely used benchmarks, LLMs can be readily matched by simple baseline models that are demonstrably inadequate for human language. As a partial remedy, we advocate for the use of thoroughly validated datasets that more accurately reflect the scope of linguistic knowledge. On these datasets, even LLMs trained on very large amounts of data perform in a way inconsistent with human behavior. The burden of an existence proof is considerably heavier than previously realized.

1 The Promise of Existence Proofs

The rapid development of deep learning and neural large language models (LLMs) over the past decade has spurred interest in their relevance to the cognitive science of language (Linzen and Baroni 2021). An assortment of performance benchmarks, many of which will be reviewed below, have been developed to evaluate the capacity of LLMs — and with apparent success. Of course, even human-like performance does not imply that LLMs are in fact appropriate cognitive models of human language and cognition. Most obviously, LLMs require orders of magnitude more input than human children receive: BERT was trained on about 3.3 billion words, and Chinchilla on 1.4 trillion, and the recent editions of GPT models likely on orders of magnitude more, while an English-learning child only receives about a few million words per year, for a total vocabulary measured in the hundreds at age three (Fenson et al. 1994, Bornstein et al. 2004).

In a welcome move, recent studies have attempted to bring LLMs closer to the psychological setting of language acquisition. Aside from enhancing connections to the scientific study of language, smaller LLMs

*This work is a substantial revision of a paper presented at the 1st GenBench Workshop on (Benchmarking) Generalization collocated with the 2023 EMNLP conference held in Singapore (Vázquez Martínez et al. 2023). We thank the reviewers and the audience of the workshop for helpful feedback.

may also contribute to the development of technologies that serve low(er)-resourced linguistic communities. A concrete example was provided by Huebner et al. (2021): a specially tuned model trained on only 5M tokens of child-directed speech (CDS) performs well on an evaluation dataset adapted from a widely used behavioral benchmark originally designed to evaluate much larger LLMs (Warstadt et al. 2020). And in 2023, an aptly-named shared task, the BabyLM Challenge, asked participants to limit training data to 100 million words (about the input of an adolescent). The message is clear: Now that huge LLMs have already succeeded, it is time to look at small ones (Warstadt et al. 2023).

For some researchers, a successful child-sized LLM promises new understandings of human language and its acquisition. According to Portelance and Jasbi (2023), for example, LLMs would provide a proof-of-concept for what is possible "in practice" for machines and "in principle" for human learners. Warstadt and Bowman (2022) go a step further: They hold that LLMs can help understand child language acquisition as long as they succeed on *some* behavioral benchmarks even if they fail on others. This is because, according to these authors, "positive results from model learners are more meaningful than negative results." (p. 21). Since LLMs are trained on plain text without "grounding", i.e., information available to a child learner in a physical and social learning environment, positive results from such ablated conditions provide a compelling baseline of what can be learned without domain specific principles and constraints of human language.

But it is facile to assert the utility of grounding information as if its absence necessarily places LLMs at a disadvantage. Theoretically, enriching the text with additional semantic/pragmatic information does not necessarily enhance learnability (Niyogi 2006). More empirically, it is not clear whether grounding, to the extent it helps children, can similarly benefit LLMs. On the one hand, the environmental input arrives in many shapes and flavors: they often come into conflict, with differential effects on child learners across developmental stages. While social and communicative cues such as eye-gaze do provide guidance for language acquisition among toddlers (Baldwin 1991), they fail to draw attention away from perceptually privileged cues (e.g., bright colors, objects in motion) during infancy (e.g. Pruden et al. 2006). On the other, many or perhaps most elements in language (e.g., abstract words such as *of*, *can*, *good*, *time*, *think*, etc.) are not plausibly grounded in the world, yet children can learn them extremely early. And there is evidence that the richness of the sensory environment may overwhelm the learner (Gillette et al. 1999). Furthermore, language acquisition is remarkably resilient even when grounding is severely limited. The empirical literature is vast: Blind children learn language in ways comparable to sighted children (Landau and Gleitman 1985), home signers spontaneously create a combinatorial syntactic system (Goldin-Meadow and Yang 2017), and perhaps most dramatically, a tactile linguistic system can be established via touch for deaf-blind individuals (Chomsky 1986). Of course, none of this suggests that children do no make use of grounding information, just as no one has suggested that language acquisition takes place in a vacuum. But children make use of such information in highly specific ways that cannot be divorced from developmental factors, linguistic and otherwise: there is no guarantee that LLMs can benefit similarly. In fact, there is evidence that they do not: augmenting LLMs with visuo-spatial cues fail to improve learning performance (Yun et al. 2021), and an ablation experiment of multimodal learning fails to yield detectable advantage of grounding (Amariucai and Warstadt 2024).

If the potential benefit of grounding information is an over-estimation, the actual benefit of LLM training scheme over child learners has gone largely uncommented. Like most machine learning models, LLMs operate by iterative optimization over the training data. But there is no psychological evidence for such a learning process in children. If anything, empirical research has uncovered severe limitations of children's memory, attention, and other computational capacities. Even in the simplest word learning task, children can only focus on a single referent meaning even though multiple are available in the environment (Woodard et al. 2016). Five-year-olds have difficulty integrating contextual considerations in sentence comprehension and show little or no ability to recover from parsing ambiguities (Trueswell et al. 1999). In artificial language learning experiments, children generally only learn a single linguistic form even though the input by design contains variability (Hudson Kam and Newport 2009). And a long-established experimental methodology

exploits children's very cognitive limitations: failure to imitate adult sentences is taken as indication of underdeveloped knowledge of particular syntactic structures (McNeill 1970). In fact, research from several different traditions (Elman 1993, Newport 1990, Yang 2016) suggests that restrictions on children's computational capacity are beneficial for the acquisition of language.

Taken together, LLMs and human children operate under significantly different learning conditions. In the concluding remarks of this paper, we will put forward some additional considerations from child language that ought to be connected with LLM modeling. As it stands, however, LLMs and language acquisition are like apples and oranges: even if LLMs were to achieve perfectly human-like linguistic ability, they still stand to tell us nothing about the nature of language and cognition.

Our main concern in this paper, however, is narrower and more immediate. We ask: All provisos aside, have LLMs delivered an existence proof as claimed? As usual, the proof is in the pudding. To invoke the familiar goals of linguistics (Chomsky 1965), the question of explanatory adequacy is only on the table when descriptive adequacy has been achieved.

We contend that LLMs are still far from delivering an existence proof: their relevance as cognitive models of language acquisition is therefore a moot point. Our position is informed by previous efforts, individually and collectively, to investigate the validity of models of language and cognition, especially the need for well-chosen baseline, or null hypothesis, comparisons. In particular, we have found useful to evaluate language learning proposals against n-grams, the simplest statistical model possible. For example, the apparent success of a Bayesian learning model (Yang and Piantadosi 2022) is undercut by the fact that it is matched by a provably inadequate trigram model of language (Kodner et al. 2022). In the arena of LLMs, proposed tests for hierarchical syntactic structures (van Schijndel and Linzen 2018, Prasad et al. 2019) are in fact not syntactic at all: not only does a definitionally linear n-gram model succeed at the tests, but so does an LLM when the words in the test sentences are randomly scrambled (Kodner and Gupta 2020). The current paper builds on the earlier work of Vázquez Martínez (2021), the first to make use of datasets that more realistically reflect human language users' syntactic knowledge for the purposes of testing LLMs abilities: LLMs offer scant improvement over, once again, n-gram models. Overall, our past experience is concordant with findings in NLP, that LLMs often take unexpected shortcuts, exploiting unforeseen biases in the data and evaluation methods (e.g., Chao et al. 2018, McCoy et al. 2019, Wang et al. 2022). That is, LLMs may succeed at tasks in ways that are misleading and unintended by the task designers. Positive results do not necessarily demonstrate a proof-of-concept.

In this paper,[1] we evaluate LLMs 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 for the express purpose of testing LLMs as proxies for child language acquisition with restricted vocabulary and trained only on child-directed speech (CDS). Section 2 reviews some key features of child language acquisition and methodological issues in the assessment of syntactic knowledge. 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 papers in theoretical linguistics. We find that LLMs 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 benchmarks that better probe the structural knowledge of syntax — and more and better baseline models. We also make recommendations for future work on LLM modeling, especially with respect to developmental findings from child language acquisition.

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

2 Grammaticality, Acceptability, and Language Acquisition

One of the goals of linguistic theory is to characterize the properties that distinguish grammatical from ungrammatical sentences in a language. Grammaticality is most saliently observed in minimally contrasting examples: *colorless green ideas sleep furiously* and *furiously sleep ideas green colorless* form the most prominent pair (Chomsky 1957). In practice, the empirical study of grammaticality typically, and certainly most conveniently, relies on native speakers' acceptability judgments. Such practice is not without complications. Acceptability judgment reflects the complex interaction between the grammar and other cognitive and perceptual systems: It therefore generally produces gradient responses as opposed to the binary outcome of grammaticality. 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 beyond acceptability degradation due to sentence length and other non-structural factors. Furthermore, acceptability judgment data 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).

Unsurprisingly, by far the most widely used benchmark for assessing LLMs is also based on acceptability discrimination (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, Sinclair et al. 2022), mirroring the practice in the empirical study of language. In this sense, LLMs are a continuation of classical statistical language models such as n-grams. Statistical language models are trained to assign probabilities to strings: Minimal pairs can therefore be submitted for assessment in humans and machines alike when machines assign higher probability to acceptable strings than to unacceptable ones. To this end, large acceptability rating benchmarks have been developed in various forms for syntax, semantics, and morphology: these include BLiMP (Warstadt et al. 2020), SyntaxGym (Gauthier et al. 2020), and CoLA (Warstadt et al. 2019), which sit alongside smaller scale studies focused on specific linguistic phenomena (e.g., Linzen et al. 2016b, Marvin and Linzen 2018, Wilcox et al. 2018, Yedetore et al. 2023). A wide range of positive results has been reported, which has led claims of LLMs as informative cognitive models of language.

Before we examine these benchmarks and their suitability of LLM evaluation, it is once again useful to consider how children acquire stable grammars reflected in the consistency in acceptability studies. 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). 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 early 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 10 million words per year (Hart and Risley 1995) and a vocabulary size of only a few hundred types (Fenson et al. 1994). Recent research has also seen a convergence between the formal and psychological study of language and the quantitative study of language variation, use, and change. Labov, for instance, often speaks of "the enigma of uniformity:" It is remarkable that the grammars acquired by individuals in a speech community show "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; see also Labov 1972).

The acquisition of vocabulary and grammar provides guidelines for the investigation of LLMs, and indeed all models of language learning. Vocabulary acquisition 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 does result in better learning. But, the structural aspects of the grammar are different. They require form generalizations over the vocabularies: as noted above, very different vocabularies apparently yield identical grammars.

The distinction between rote learning and structural learning — roughly, words vs. rules — is not well reflected by existing LLM benchmarks including those discussed in this paper. In practice, these benchmarks are a mixture of tests for both vocabulary learning and grammar learning. Moreover, many benchmarks are stochastically generated by templates, which come with their own host of problems. While this practice makes a large number of sentences immediately available for testing — and statistically significant results may be obtained — this comes at the expense of structural diversities that are particularly revealing and the potential for compounding biases stemming from how the templates are designed. We return to this issue with concrete examples in Section 3.

Furthermore, the sentences used in LLM benchmarks are often highly unnatural and semantically and pragmatically uncontrolled. This is precisely the confounding factor that linguists seek to neutralize when attempting to uncover the structural basis of language. Authors such as Warstadt et al. (2020) are aware of the unnaturalness of their generated sentences. They present this example, '`Sam ran around some glaciers.`' Nevertheless, these researchers do not regard such semantic and pragmatic oddities as a problem, arguing that since they design sentence pairs to be minimally lexically distinct, they affect both sentences in a test pair equally. This reasoning relies on the (untested) assumptions that the injection of noise does not obscure the signal and that it does not affect ratings of acceptable and unacceptable sentences more than the other. However, we cannot rule out unintended, and indeed unknown, interactions between grammaticality and felicity in acceptability judgment. In fact, Sprouse et al. (2018) find that semantic implausibility affects sentence well-formedness results, even in forced choice tasks, conceptually similar to the LLM benchmark tasks. At the same time, we *have* seen case studies where deliberate mutilation of the signal — random permutation of words in a sentence (Kodner and Gupta 2020) — does not impede model performance in categorizing sentence types: the purported structural signal cannot be the thing, or at least the only thing, that models in the study captured.

Fortunately, there is already a large amount of material that not only reflects the realistic complexity of syntax but has also been carefully curated to minimize the confound of non-grammatical factors. Furthermore, these datasets, such as the LI-Adger dataset discussed in 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. If LLMs can truly capture the structural regularities of syntax, they should exhibit a similar degree of uniformity.

3 Re-examining the Benchmarks

BLiMP (Warstadt et al. 2020)

Warstadt et al. (2020) introduce the Benchmark of Linguistic Minimal Pairs (BLiMP)[2] 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 acceptability judgments on individual sentences, as is most commonly done with human subjects, they present a model, LLM or

[2] `https://github.com/alexwarstadt/blimp`

otherwise, with two sentences that only differ in one structural or lexical property as in (1). For a given minimal pair m_i consisting of an acceptable sentence $s_{i,1}$ (labeled in the data set as `sentence_good`) and an unacceptable sentence $s_{i,2}$ (labeled as `sentence_bad`), if a model, evaluates $P(s_{i,1}) > P(s_{i,2})$, then it has succeeded on m_i. A model is scored according to the percentage of all the minimal pairs for which it assigned higher probability to the acceptable sentence. The minimal pair approach allows for the direct evaluation of LLMs without training a binary classifier on top of them as was necessary for previous acceptability benchmarks (e.g., CoLA; Warstadt et al. 2019).

(1) Sentence pair from BLiMP's `adjunct_island` phenomenon

`sentence_good:` `Who should Derek hug after shocking Richard?`

`sentence_bad:` `Who should Derek hug Richard after shocking?`

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, this comes with the side-effect of low structural diversity and implausible semantics. The benchmark corpus includes data sets for 12 linguistic *phenomena*, including ANAPHOR AGREEMENT, ARGUMENT STRUCTURE, BINDING, CONTROL/RAISING, among others listed in Tables 3-4. These are further divided into 67 *paradigms*, each containing 1,000 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 was used as part of the test data for the 2023 BabyLM Challenge.

Zorro (Huebner et al. 2021)

Huebner et al. (2021) aim to explicitly evaluate the relationship between LLMs and the acquisition of grammar. They introduce BabyBERTa, "an acquisition-friendly version of RoBERTa," which Portelance and Jasbi (2023) describe as a "proof-of-concept" model. Variants of BabyBERTa are trained using only CDS from AO-CHILDES (Huebner and Willits 2021; BabyBERTa_AO-CHILDES), a pre-processed version of English CHILDES (MacWhinney 1991), as well as variants trained on larger datasets from other sources.

Because BabyBERTa_AO-CHILDES (henceforth BabyBERTa) was trained on much less text than RoBERTa and related models are, its vocabulary is much smaller. To mitigate the impact of out-of-vocabulary items on their tests, the authors introduce a new grammaticality test suite, Zorro,[3] in the style of BLiMP (2). Sentence pairs are generated for one paradigm each for 11 of BLiMP's 12 phenomena, along with two additional phenomena. While Zorro bills itself as a benchmark as similar to BLiMP as possible given the constraints on its vocabulary, 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 Table 2, and the full data sets themselves are made available by the benchmarks' original authors.

(2) Sentence pair from Zorro's `local_attractor-in_question_with_aux` phenomenon

`sentence_good:` `is the whale getting the person ?`

`sentence_bad:` `is the whale gets the person ?`

3.1 Linear Baselines

As noted earlier, BLiMP and Zorro tests are automatically generated from category-based templates. This way, a large number of examples can be collected and tested, but the drawback is that all examples are

[3] `https://github.com/phueb/Zorro/`

essentially the same structure. Moreover, many of the structures are simple, falling considerably below the coverage of modern syntactic analyses. In fact, many phenomena appear solvable by strictly linear methods *in practice*. Some of these do not even require structural knowledge *in principle*, and instead test lexical memorization. The observation that such template-generated examples can be solved this way is not new to the 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) about as well as neural models of the time.

Consider BLiMP's ANAPHOR AGREEMENT paradigm. This tests whether a model recognizes agreement between an anaphor (e.g., *himself*/*herself*) and its antecedent, but there are two problems here. First, the antecedent is always the first, and only noun, preceding the anaphor in the sentence (3). A linear rule like "find the leftmost noun, then make sure it matches the anaphor" should be able to solve it. This is a weakness in the template used to generate these minimal pairs.

(3) Sentence pair from BLiMPs's `anaphor_number_agreement` phenomenon

`sentence_good:` `Susan revealed herself.`

`sentence_bad:` `Susan revealed themselves.`

Second, the task relies heavily on lexical knowledge. Since the antecedent in most of the minimal pairs is an English proper name, the test reduces to memorizing which gender is conventionally assigned to the names (4). Other sentences incorporate common nouns and *itself* and *themselves*, but these too reduce to identifying human animate vs. non-human and singular vs. plural nouns. Furthermore, some minimal pairs, as in (5), are invalid. Obviously both *himself* and *herself* could agree with *dancer* in the absence of any disambiguating context, so both sentences in the pair are grammatical.

(4) Sentence pair from BLiMPs's `anaphor_gender_agreement` phenomenon

`sentence_good:` `Katherine can't help herself.`

`sentence_bad:` `Katherine can't help himself.`

(5) Invalid sentence pair from BLiMPs's `anaphor_gender_agreement` phenomenon

`sentence_good:` `That dancer wouldn't aggravate herself.`

`sentence_bad:` `That dancer wouldn't aggravate himself.`

The SUBJECT-VERB AGR phenomenon, provides another example of systemic problems. This phenomenon is widely used in behavioral probing because noun phrases intervening between the subject and verb can serve as agreement distractors. These are useful in distinguishing whether a model is attempting to enforce agreement with the closest noun or with the structural subject. Consider (6) from Linzen et al. (2016b, (9)) which uses such sentences as probes for LLM structural knowledge:

(6) The **roses** in the *vase* by the *door* **are** red.

The bold verb should agree with the bold subject plural noun, and not the intervening italicized singular nouns. While such an agreement pattern does not necessarily require a structural rule (all of Linzen et al.'s examples require agreement with the first/leftmost noun, a linear pattern), it does require a model to rely on a long distance dependency. However, BLiMP's phenomenon only partially evaluates this. It is made up of six paradigms, four of which only evaluate string-adjacent agreement with no distractor, for example (7). In the two paradigms with intervening distractor nouns, the target for agreement is always the first noun, for example (8), so a linear rule along the lines of "always agree with the first/leftmost noun" would completely suffice to solve the entire phenomenon.

(7) Sentence pair from BLiMPs's `regular_plural_subject_verb_agreement_1` phenomenon

`sentence_good:` `Most legislatures haven't disliked children.`

`sentence_bad:` `Most legislatures hasn't disliked children.`

(8) Sentence pair from BLiMPs's `distractor_agreement_relational_noun` phenomenon

`sentence_good:` `A niece of most senators hasn't descended most slopes.`

`sentence_bad:` `A niece of most senators haven't descended most slopes.`

ANAPHOR and SUBJECT-VERB AGR, among others as implemented, are just lexical and string-based phenomena – they do not actually test a model's syntactic knowledge. In principle, an n-gram model with sufficiently large n and enough training data to capture the conventional gender of English names should be able to solve them perfectly. Success on such simple tests tells us little about the genuine grammatical capacity of LLMs and distorts or dilutes summary metrics calculated over the benchmark. We evaluate this concern quantitatively with two studies of linear rules that do not incorporate structural knowledge. We find that many paradigms are indeed solvable with non-human-like linear approaches. These paradigms therefore do not contribute to the overall goal of evaluating whether an LLM 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 3.1 billion token Gigaword Corpus (Graff et al. 2003) in addition to the accuracy of three neural LLMs and humans. They find that the 5-gram model scores above chance (>50%) on all but two phenomena and even surpasses at least one neural model on 25 of the 67 paradigms and 4 of the 12 phenomena. We point out that one could therefore conclude, by the logic of the test, that the 5-gram model has "learned" a substantial amount about the linguistic phenomena, even though we know that the 5-gram model is just a linear string model with a fixed context window. Surprisingly, this was not the authors' conclusion. Despite the 5-gram model's substantially above-chance performance, they conclude "The 5-gram model's poor performance – overall and on every individual category – indicates that BLiMP is likely not solvable from local co-occurrence statistics alone" (Warstadt et al. 2020, §5).

A more detailed look at the Gigaword 5-gram model's performance proves insightful. In general, and as expected, it struggles on paradigms with long distance dependencies that stretch the limits of its small context window. Revisiting SUBJECT-VERB AGR. for illustration, its performance is only slightly weaker than the LLMs on each of the four string-adjacent paradigms, while it performs much worse, and well below chance, on the two paradigms with distractors. This makes sense. Since the distractor nouns lay within the 5-gram model's context window and the true subject often did not, the model tended to prefer agreement between the distractor and verb. It should be noted that the neural models also perform substantially worse (though still well above chance) on the distractor paradigms than on the string-adjacent paradigms.

However, the 5-gram model also performs well on some paradigms that it should not be able to. It outperforms two or three neural models on four of eight ISLAND paradigms, for example. Islands are a syntactic phenomenon *par excellence*, so this result suggests a problem with the data. At the very least, phenomena on which the 5-gram model readily scores above chance should be removed from evaluation, since they are not informative about a model's structural knowledge.

Since our focus is primarily on language acquisition, and the Gigaword 5-gram is trained on far too much data to serve as a reasonable model in that setting, we conduct our own tests with a 5-gram model with back-off trained only on AO-CHILDES. We compare these results to BabyBERTa, the acquisition-relevant

derivative of RoBERTa trained on the same data.[4] A 5-gram model trained on only AO-CHILDES is expected to be excessively limited in its vocabulary. It is likely to be missing test n-grams present in both benchmarks, which is a particular problem, since this type of model has no way to generalize over lexical categories. To give the linear baseline an additional chance to succeed while adding minimal complexity, we evaluate both a 5-gram word model (5-word) and a 5-gram part-of-speech (POS) model trained only on POS tags (5-tag). To train the POS model, AO-CHILDES was tagged using GPoSTTL, a rule-based POS tagger with a tokenizer and lemmatizer based on the Brill Tagger (Brill 1992). This was used to train the sklearn package's CRF POS-tagger, which was then used to label the benchmark corpora. This approach of training the 5-tag model only on labeled AO-CHILDES was taken to avoid bringing in additional knowledge from a tagger trained on larger corpora. The downside is that the tagger is not accurate on labeling the ungrammatical benchmark sentences, which likely hurts the performance of the 5-tag model. Nevertheless, we observe that the 5-tag model often performs well even in spite of this inherent limitation. 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.

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 on a given minimal pair. The oracle is not in itself a fair direct comparison, because it requires access to the minimal pair labels to determine whether or not either 5-gram model answered correctly. Rather, it should be interpreted as a summary metric for the relationship between 5-word and 5-tag. An oracle score only slightly above that of the two 5-gram models would indicate that they tend to succeed and fail on mostly the same minimal pairs, while a high relative oracle score indicates that the two 5-gram models tend to make complementary errors instead. 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. This is what we observe.

The results of the 5-gram experiments are summarized in Table 1 with breakdowns by Zorro and BLiMP paradigm in Tables 2-4. We draw three conclusions from these. First, the 5-gram models perform surprisingly well relative to the BabyBERTa transformer despite their extremely non-human-like simplicity when trained on the same AO-CHILDES data. Either 5-word or 5-tag outperformed BabyBERTa on 11 of 23 Zorro paradigms and 21 of 67 BLiMP paradigms. As expected, the AO-CHILDES 5-gram models perform worse on BLiMP than the Gigaword 5-gram model, but so does BabyBERTa, and both still achieves high accuracy on several paradigms scattered across the phenomena.

Second, the 5-gram oracle outperforms 5-word, 5-tag, and BabyBERTa on a majority of paradigms and on overall accuracy on both BLiMP and Zorro. This means that the 5-word and 5-tag models tended to succeed on different minimal pairs. 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 preserved BLiMP's complexity and merely accounted 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 than BLiMP, and it more greatly overestimates the apparent positive results of the acquisition-inspired BabyBERTa.

[4]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: Summary performance 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.

Phenomenon	Paradigm	BabyBERTa AO-CHILDES	5-Gram Word	5-Gram Tag	5-Gram Oracle	Simple 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 $\geq$ BabyBERTa		–	8/23	8/23	14/23	22/23

Table 2: 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 scores are bolded.

3.1.2 Hand-Written Simple Rules

In addition to reporting results on 5-gram models, we created simple hand-written rules, inspired by the pervasive "agree with the first/leftmost noun" pattern, 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 LLMs, 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 LLMs. 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*. If any of these patterns hold in the training data, then there is a real possibility that the models have learned and are exploiting them. For both Zorro and BLiMP, the same rule can often be used to solve more than one paradigm. We write simple linear rules for each Zorro and BLiMP paradigm, summarized in Tables 5-7

Phenomenon	Paradigm	BabyBERTa AO-CHILDES	5-Gram Word	5-Gram Tag	5-Gram Oracle	Simple Rule
ANAPHOR AGR	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**
ARG STRUCTURE	animate_subject_passive	45.5	**48.4**	24	**58.4**	**65.25**
	animate_subject_trans	59.7	50	57.1	**78.2**	**84.65**
	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_obj.	59.1	30.5	23.4	46.5	**67.95**
	existential_there_subj.	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**
DET-NOUN AGR (agreement = agr)	agr_irregular_1	66.6	48.8	37.4	61.3	**68.45**
	agr_irregular_2	87.4	74.3	12.3	77.1	73.7
	agr_with_adjective_1	76.3	48.2	49.7	63.8	**95.95**
	agr_with_adj_irregular_1	82.9	49	49.7	56.3	74.45
	agr_with_adj_irregular_2	67	49.5	18.3	58.2	**71.8**
	agr_with_adj_2	80.4	49.8	19.9	59.7	**95.6**
	agr_1	72.2	64.1	48.1	**74.5**	**95.55**
	agr_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**

Table 3: Word and tag-level 5-gram models trained on AO-CHILDES plus 5-Gram Oracle and Simple Linear Rule Oracle for BLiMP (ANAPHOR AGR through ELLIPSIS). 5-Gram and Simple Rule scores that are greater than BabyBERTa_AO-CHILDES scores are bolded.

Phenomenon	Paradigm	BabyBERTa AO-CHILDES	5-Gram Word	5-Gram Tag	5-Gram Oracle	Simple Rule
FILLER-GAP	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_dist.	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_dist.	14.9	**25.8**	12.8	**32.8**	**99.9**
IRREGULAR FORMS	irregular_past_participle_adj.	59.8	**99.4**	12.2	**99.4**	**100**
	irregular_past_participle_verbs	59.8	**99.4**	12.2	**99.4**	**100**
ISLAND EFFECTS (coordinate_structure_constraint = csc)	adjunct_island	63.8	58.4	55.5	**82.5**	**94.5**
	csc_complex_left_branch	36.2	11.8	19.6	26.9	**97.05**
	csc_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 (sent = sentential)	matrix_question_npi_licensor_pres.	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_pres.	71.6	**98.4**	2.4	**98.5**	**100**
	only_npi_scope	72.1	**80.4**	**79.4**	**97.2**	**100**
	sent_negation_npi_licensor_pres.	73.8	**100**	0	**100**	**100**
	sent_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**
SUBJECT-VERB AGR (agr = agreement)	distractor_agr_relational_noun	29	26.2	21.4	**42.1**	**50.25**
	distractor_agr_relative_clause	35.6	28.3	30.4	**49.8**	**55.85**
	irregular_plural_subject_verb_agr_1	67.9	33.4	51.7	62.5	53.2
	irregular_plural_subject_verb_agr_2	66.2	51	51.9	**70.7**	59.3
	regular_plural_subject_verb_agr_1	68.8	39.9	51.1	**72**	64.35
	regular_plural_subject_verb_agr_2	60.1	51	55.6	**76.9**	**73.15**
AVERAGE across Table 3 and 4		60.72%	50.72%	37.93%	**68.32%**	**84.35%**
Fraction $\geq$ BabyBERTa		–	18/67	10/67	48/67	62/67

Table 4: Word and tag-level 5-gram models trained on AO-CHILDES plus 5-Gram Oracle and Simple Linear Rule Oracle for BLiMP (FILLER-GAP through SUBJECT-VERB AGR). 5-Gram and Simple Rule scores that are greater than BabyBERTa_AO-CHILDES scores are bolded.

Phenomenon	Paradigm	Rule Description
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 rightward
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 5: Simple Linear Rule descriptions for Zorro. Rules that require sentences within a minimal pair to be compared are marked with an asterisk (in_question_with_aux).

In summary, the hand-written 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, where the 4% loss is completely attributable to the presence of two irregular nouns, *feet* and *children* in some of the minimal pairs, which do not end in the *-s* referenced by these rules. The paradigm principle_A_case_2 is solved only with 99.2% accuracy because there are several test pairs where the grammatical sentence is actually a copy of the ungrammatical sentence (e.g., Leslie imagined herself skated around the hospital.), while all valid grammatical sentences contain a word ending in -ing. The lowest performance is 52.75% on anaphor_agreement-pronoun_gender, a paradigm that requires a model to "know" the canonical gender of English names in order to choose *himself* or *herself* (the same problem is exhibited in BLiMP's equivalent (4)). The test sentence pairs were not quite balanced, so always guessing *himself* instead of *herself* yields just over 50% accuracy.

BLiMP proved more challenging. Our rules only yielded 84.35% accuracy on average and achieved perfect scores on 14 of 67 rules. Yet, this is still a high score. The general success 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 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 of both benchmarks further emphasizes that success on the template-based behavioral task does not necessarily imply that an LLM possesses linguistic knowledge.

4 An Alternative Benchmark: 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*,[5] a major theoretical journal in linguistics, and 219 collected by Sprouse and Almeida (2012) from Adger 2003 *Core Syntax* textbook.[6] Each sentence type includes eight hand-constructed sentences, assembled into 150 pairwise (LI) and 105 multi-condition (Adger) phenomena where each minimal pair is lexically matched. We provide an example from each dataset in Table 8.

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

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

Phenomenon	Paradigm	Rule Description
ANAPHOR AGR	anaphor_gender_agreement	Does not contain *itself*
	anaphor_number_agreement	Number of words that end in *s* is even
ARG STRUCTURE	animate_subject_passive	Contains one of {*men*, *woman*, *children*, *teacher*, *lad*, *offspring*, *student*, *customer*, *girl*, *boy*}
	animate_subject_trans	*Is the shorter of the two sentences
	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_c_command	(Last word ends in *s* and (st word is any of pl_det 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	existential_there_obj.	Does not contain one of {*ask*, *press*, *entic*, *prod*, *obligat*, *convinc*, *badger*, *compel*, *sway*, *order*}
	existential_there_subj.	Contains one of {*certain*, *soon*, *likely*, *unlikely*, *bound*, *about*} or {*appear*, *sure*, *threaten*, *look*}
	expletive_it_object_raising	Does not contain one of {*ask*, *press*, *entic*, *prod*, *obligat*, *convinc*, *badger*, *compel*, *sway*, *order*}
	tough_vs_raising_1	Does not contain one of {*certain*, *soon*, *likely*, *unlikely*, *bound*, *about*}, nor *apt*
	tough_vs_raising_2	Contains one of {*certain*, *soon*, *likely*, *unlikely*, *bound*, *about*}, or *apt*
DET-NOUN AGR	agr_irregular_1	Does not end in *that* followed by (one of
(agreement = agr)	agr_irregular_2	{*people*, *women*, *men*, *children*} or a word ending
	agr_with_adj_irregular_1	in *ses*) nor in {*those*, *these*} followed by (a word
	agr_with_adj_irregular_2	ending in *is* or not with *s* at all)
	agr_with_adjective_1	Does not end in *that* followed by a word ending in
	agr_with_adj_2	a letter other than *i* followed by *s* nor in
	agr_1	{*those*, *these*} followed by (a word ending in
	agr_2	*is* or not with *s* at all)
ELLIPSIS	ellipsis_n_bar_1	Last word in {*one-ten*, *many*, *few*, *several*, *more*, *some*, *lot*, *fewer*}
	ellipsis_n_bar_2	Last word has already occurred in sentence

Table 6: Linear Rule descriptions for BLiMP (ANAPHOR AGR through ELLIPSIS). Rules that require sentences to be compared are marked with an asterisk.

Phenomenon	Paradigm	Rule
FILLER-GAP	wh_questions_object_gap	Does not contain *wh*
	wh_questions_subject_gap	Does not contain *wh*
	wh_vs_that_no_gap	Does not contain *wh*
	wh_vs_that_no_gap_long_dist.	Does not contain *wh*
	wh_vs_that_with_gap	Contains *wh*
	wh_vs_that_with_gap_long_dist.	Contains *wh*
IRREGULAR FORMS	irregular_past_participle_adj.	If 1st word is *the*, then 2nd word ends in *n*, otherwise 2nd word must not end in *n*
	irregular_past_participle_verbs	*Is the shorter of the two sentences
ISLAND EFFECTS	adjunct_island	Last word is not *about* and does not end in *ing*
(coordinate_structure_	csc_complex_left_branch	2nd word is not in {*had*, *should*, *is*, *was*, *can*, *has*, *will*, *would*, *could*, *do*, *does*, *might*, *were*, *did*}
constraint = csc)	csc_object_extraction	2nd to last word is not *and*
	left_branch_island_echo_question	Does not start with *Wh*
	left_branch_island_simple_question	2nd word is not in {*had*, *should*, *is*, *was*, *can*, *has*, *will*, *would*, *could*, *do*, *does*, *might*, *were*, *did*}
	sentential_subject_island	Ends in *ing* or *ed* or *with*
	wh_island	*wh*, capitalized or not, occurs twice in sentence
NPI LICENSING	matrix_question_npi_licensor_pres.	1st word in {*had*, *should*, *is*, *was*, *can*, *has*, *will*, *would*, *could*, *do*, *does*, *might*, *were*, *did*}
(sentential=sent)	npi_present_1	Does not contain the word *ever*
	npi_present_2	Does not contain the word *ever*
	only_npi_licensor_pres.	1st word is *only*
	only_npi_scope	1st word is *only*
	sent_negation_npi_licensor_pres.	Does not contain the word *ever*
	sent_negation_npi_scope	Does not contain the word *ever*
QUANTIFIERS	existential_there_quantifiers_1	Does not contain {*each*, *most*, *all*, *every*} while
	existential_there_quantifiers_2	also containing {*one-ten*}
	superlative_quantifiers_1	*Is the longer of the two sentences
	superlative_quantifiers_2	1st word is not *no*
SUBJECT-VERB AGR	distractor_agr_relational_noun	*Is the longer of the two sentences
(agreement = agr)	distractor_agr_relative_clause	The number of words that ends in *s* is odd
	irregular_plural_subject_verb_agr_1	Contains no word ending in a letter other than *i* and followed by *s* that is followed by a word ending in *s*
	irregular_plural_subject_verb_agr_2	None of {*people*, *women*, *men*, *children*} is followed by a word ending in *s*
	regular_plural_subject_verb_agr_1	*Is the shorter of the two sentences
	regular_plural_subject_verb_agr_2	The number of words that ends in *s* is odd

Table 7: Linear Rule descriptions for BLiMP (FILLER-GAP through SUBJECT-VERB AGR). Rules that require sentences to be compared are marked with an asterisk.

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 8: 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.

While the LI-Adger dataset serves as a behavioral probe, and is thus subject to the inherent weaknesses of such approaches, it improves upon prior template-based benchmark data sets in three key ways. First, unlike BLiMP and Zorro, the LI-Adger sentences were hand-constructed by theoretical linguists who are well aware of the contribution of non-grammatical factors such as lexical frequency, sentence parsing strategies, and semantic and pragmatic plausibility to the behavioral measure of acceptability rating (Sprouse et al. 2018). Therefore, best efforts have been made to control for these confounds and zero in on the structural properties of the grammar. The strong confirmation from large scale judgment data provides methodological support for such practices, as well as the empirical status of these datasets. 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. They were not generated from templates nor even devised by a single research team, since they were collected from across the literature. 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 the Forced-Choice (FC) paradigm that was used for the BLiMP human baselines. This is 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 2,391 unique minimal pairs and collect pseudo log-likelihood scores for each sentence from several models evaluated by Huebner et al. (2021). We score them them using the same criteria as BLiMP and Zorro, where a model is expected to assign higher probability to the grammatical member of a grammatical-ungrammatical sentence pair. 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 10 million words. If we were to adopt Warstadt and Bowman's (2022) "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 LLMs. 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.

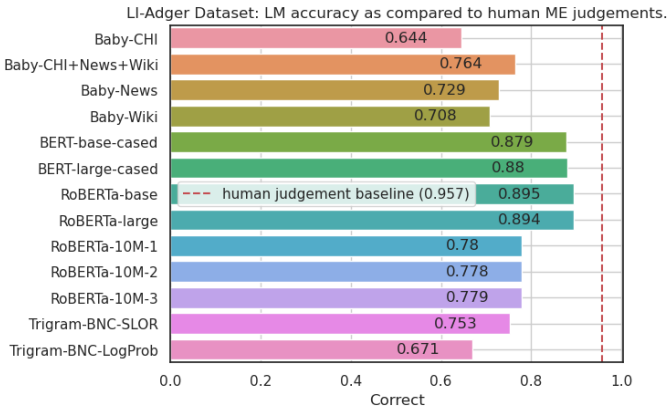

Figure 1: LM performance on the LI-Adger dataset. Human performance is marked by the vertical line. Baby=BabyBERTa, CHI=AO-CHILDES, News=AO-NEWSELA, Wiki=Wikipedia-1.

We z-score the LLM and trigram model 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 model. We find that the models are much more variable in their judgments: The human judgments, on average, vary by 0.288 standard deviation (std. dev.) units within a given set of sentences. On the other hand, the LLM 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. This within-model variability provides another data point in addition to raw accuracy, which can be compared between people and machines, one that casts LLMs in a less human-like light.

Exploring this further, 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 deviation, we observe generally high correlations between the various LLMs, but much lower correlations between the LLMs and humans. As humans are the odd ones out, this suggests that whatever the LLMs learn does not appear to be human-like. Interestingly, the BabyBERTa LLMs show very high correlations with the naive trigram log-likelihood scores and very low correlations with trigram SLOR scores, raising further suspicions that these small acquisition-inspired LLMs behave like a very non-human-like model.

5 Conclusions and Prospects

The development of LLMs has generated remarkable technological advances and intellectual excitement. But a clear-eyed assessment of their capacity and limitations is still necessary, not least because of the closed nature of most of the leading LLM models, and the fact that there is very little theoretical understanding of how and why these models work.

The claim of LLMs as an existence proof for language acquisition appears premature. The scientific relevance of LLMs is predicated upon, among others, the assumption that despite the differences in the learning mechanisms and learning environment, LLMs already demonstrate human-like linguistic abilities. By critically revisiting the commonly used evaluation benchmarks, we contend that current LLMs still have

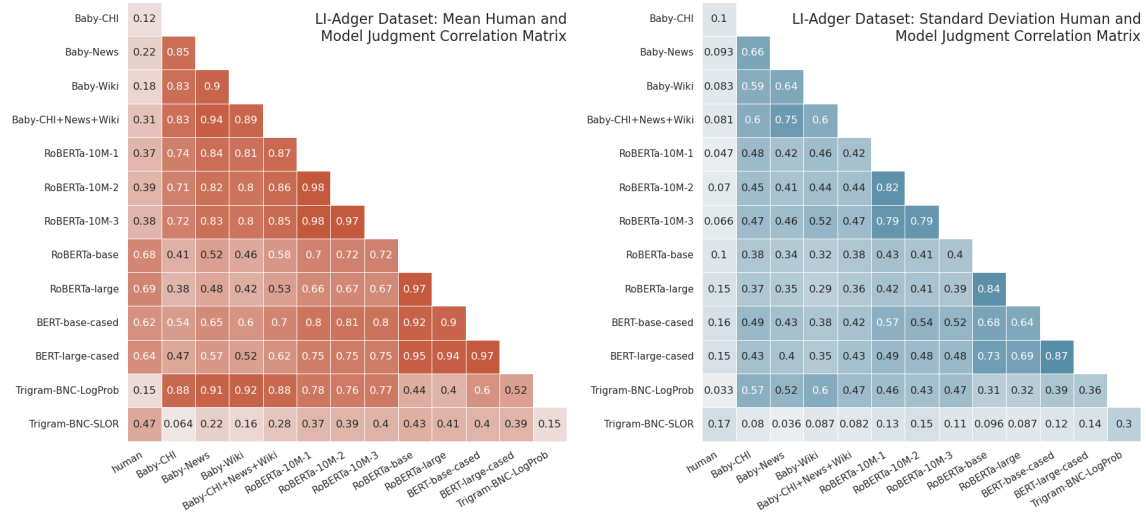

Figure 2: Correlation matrices of human judgments and LM output means (top) and standard deviations (bottom) on each sentence type on the LI-Adger dataset. Baby=BabyBERTa, CHI=AO-CHILDES, News=AO-NEWSELA, Wiki=Wikipedia-1.

some ways to go. The inaccurate assessment of LLMs stems from the weakness of existing benchmarks, which can be met in unintended ways and in any case do not reflect a representative range of linguistic phenomena. Furthermore, we suggest that binarized tasks be complemented with professionally curated datasets that include gradient acceptability judgments from human subjects.

Our study is about the limitations of evaluation, so it is useful to point out the limitations of our study 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: we have only considered very simple n-gram language models, while the classic NLP toolbox contains many others. For example, the problem of aligning words and images is a well-studied machine learning problem: think Google image search. It would be informative to compare classic non-neural methods against LLMs in the modeling of lexical acquisition (Vong et al. 2024). And like most work in this area, our evaluations were only performed on English. Fortunately, recent years have seen the development of LI-Adger-type datasets for specific phenomena (Sprouse and Hornstein 2013) as well as for other languages (Kush et al. 2019, Chen et al. 2020, Fukuda et al. 2022, Al-Aqarbeh and Sprouse 2023). All behavioral probes, of course, should be submitted for rigorous assessment.

We end with some broader discussion about language acquisition and the cognitive interpretation of computational models such as LLMs. While it is now widely recognized that children learn language with only a fraction of the data needed for LLM training, merely reducing the amount of training data alone – such as the 100 million word threshold in the BabyLM Challenge – still falls far short of the requirement for an adequate model of language acquisition. It is true that a native speaker's knowledge of language can be established on the basis of approximately 100 million words, but child language research makes clear that not all aspects of linguistic knowledge are learned at the same time: the developmental trajectory of language acquisition matters as much as the final grammar attained (Pinker 1984, Yang 2002). 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), 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. For example, English-learning children use the passive structure relatively late (Borer and Wexler 1987, Pinker et al. 1987) but no such delay is found in the passives

of Sesotho (Demuth 1989) and Inuktitut (Allen and Crago 1996) learning children.

Successful learning in the "limit" (e.g., 100 million word), therefore, is not an inadequate measure of LLMs as cognitive models of language acquisition. In this sense, the current discussion of LLMs and their cognitive relevance compares unfavorably against the earlier connectionist approach to language, which was deeply concerned with the empirical study of child language and subsequently led to many empirical discoveries (McClelland and Patterson 2002, Pinker and Ullman 2002). Let us not forget that the major thrust of the original PDP model for morphological learning (Rumelhart and McClelland 1986) was an account of the abrupt emergence of productivity and over-regularization (Ervin and Miller 1963, Brown 1973), previously attributed to symbolic rules. Disappointingly, child language development has received virtually no attention in the LLM literature. A few case studies, however, are already informative. 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. At the same time, child language acquisition often exhibits non-monotonic learning trajectories, with important implications for learning theories: once again, the phenomenon of over-regularization in English past tense but similar patterns are found in phonological, syntactic, and semantic acquisition (e.g., Richter 2021, Bowerman 1988, Crain and Thornton 2000). However, to the best of our knowledge, LLM learning shows no similar developmental trends as performance monotonically improves as the training data size increases (Belth et al. 2021, Zhang et al. 2021, Kodner et al. 2023).

In our understanding, the research goals of the LLM community in fact align well with the developments in theoretical linguistics in past thirty years, as both aim to minimize principles and constraints unique to language (Chomsky 1995, Hauser et al. 2002, Berwick and Chomsky 2016). In our own work, again individually and collectively, we have developed distributional learning models as alternative accounts to a domain specific Universal Grammar (Yang 2002, 2016, Li et al. 2021, Belth et al. 2021, Kodner 2022, Heuser et al. 2024). These efforts may find useful points of contact with LLMs, so long as the goals of descriptive and explanatory adequacy are broadly shared and respected.

References

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

Rania Al-Aqarbeh and Jon Sprouse. 2023. Island effects and amelioration by resumption in jordanian arabic: An auditory acceptability-judgment study. *Syntax*.

Shanley EM Allen and Martha B Crago. 1996. Early passive acquisition in inuktitut. *Journal of child language*, 23(1):129–155.

Theodor Amariucai and Alex Warstadt. 2024. Acquiring linguistic knowledge from multimodal input. In *Proceedings of the 27th Conference on Computational Natural Language Learning: The BabyLM Challenge*, pages 128–141.

Dare A. Baldwin. 1991. Infants' contribution to the achievement of joint reference. *Child Development*, 62.

Caleb Belth, Sarah Payne, Deniz Beser, Jordan Kodner, and Charles Yang. 2021. The greedy and recursive search for morphological productivity. In *Proceedings of CogSci 2021*.

Robert Berwick and Noam Chomsky. 2016. *Why only us: Language and evolution*. MIT Press, Cambridge, MA.

Hagit Borer and Kenneth Wexler. 1987. The maturation of syntax. In Thomas Roeper and Edwin Williams, editors, *Parameter setting*, pages 123–172. Reidel, Berlin.

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.

Melissa Bowerman. 1988. The 'no negative evidence' problem: How do children avoid constructing an overly general grammar? In John A. Hawkins, editor, *Explaining language universals*, pages 73–101. Basil Blackwell, Oxford.

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.

Zhong Chen, Yuhang Xu, and Zhiguo Xie. 2020. Assessing introspective linguistic judgments quantitatively: the case of the syntax of chinese. *Journal of East Asian Linguistics*, 29:311–336.

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

Carol Chomsky. 1986. Analytic study of the tadoma method: Language abilities of three deaf-blind subjects. *Journal of Speech, Language, and Hearing Research*, 29(3):332–347.

Noam Chomsky. 1957. *Syntactic structures*. Mouton, The Hague.

Noam Chomsky. 1965. *Aspects of the theory of syntax*. MIT Press, Cambridge, MA.

Noam Chomsky. 1995. *The minimalist program*. MIT Press, Cambridge, MA.

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.

Stephen Crain and Rosalind Thornton. 2000. *Investigations in universal grammar: A guide to experiments on the acquisition of syntax and semantics*. MIT Press, Cambridge, MA.

Katherine Demuth. 1989. Maturation and the acquisition of the sesotho passive. *Language*, pages 56–80.

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.

Jeffrey L Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99.

Susan M Ervin and Wick R Miller. 1963. Language development. In Harold Stevenson, editor, *Child psychology: The sixty-second yearbook of the National Society for the Study of Education*, pages 108–143. University of Chicago Press.

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.

Shin Fukuda, Nozomi Tanaka, Hajime Ono, and Jon Sprouse. 2022. An experimental reassessment of complex np islands with np-scrambling in japanese. *Glossa: a journal of general linguistics*, 7(1).

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.

Jane Gillette, Henry Gleitman, Lila Gleitman, and Anne Lederer. 1999. Human simulations of vocabulary learning. *Cognition*, 73(2):135–176.

Susan Goldin-Meadow and Charles Yang. 2017. Statistical evidence that a child can create a combinatorial linguistic system without external linguistic input: Implications for language evolution. *Neuroscience and Biobehavioral Reviews*, 81(Part B):150 – 157.

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.

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.

Marc D Hauser, Noam Chomsky, and W Tecumseh Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579.

Annika Heuser, Héctor Vázquez Martínez, and Charles Yang. 2024. The learnability of syntactic islands. In *Proceedings of the 54th Northeast Linguistic Society Meeting*, page Forthcoming.

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.

Carla L Hudson Kam and Elissa L Newport. 2009. Getting it right by getting it wrong: When learners change languages. *Cognitive psychology*, 59(1):30–66.

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.

Jordan Kodner. 2022. What learning Latin verbal morphology tells us about morphological theory. *Natural Language and Linguistic Theory*, 41:733–792.

Jordan Kodner, Spencer Caplan, and Charles Yang. 2022. Another model not for the learning of language. *Proceedings of the National Academy of Sciences*, 119(29):e2204664119.

Jordan Kodner and Nitish Gupta. 2020. Overestimation of syntactic representationin neural language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1757–1762.

Jordan Kodner, Salam Khalifa, Sarah Payne, and Zoey Liu. 2023. Re-Evaluating the Evaluation of Neural Morphological Inflection Models. In *Proceedings of the 45th Annual Meeting of the Cognitive Science Society (CogSci)*, volume 45, pages 3259–3267, Sydney, NSW, Australia. Cognitive Science Society.

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

Dave Kush, Terje Lohndal, and Jon Sprouse. 2019. On the island sensitivity of topicalization in norwegian: An experimental investigation. *Language*, 95(3):393–420.

William Labov. 1972. *Sociolinguistic patterns*. University of Pennsylvania Press, Philadelphia.

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

Barbara Landau and Lila R Gleitman. 1985. *Language and experience: Evidence from the blind child*, volume 8. Harvard University Press, Cambridge, MA.

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.

Daoxin Li, Lydia Grohe, Petra Schultz, and Charles Yang. 2021. The distributional learning of recursive structures. In *Proceedings of the 45th Boston University Conference on Language Development*, pages 471–485, Somerville, MA. Cascadilla Press.

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.

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.

James L. McClelland and Karalyn Patterson. 2002. Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences*, 6(11):465–472.

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.

David McNeill. 1970. *The acquisition of language: The study of developmental psycholinguistics*. Harper and Row, New York.

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

Elissa Newport. 1990. Maturational constraints on language learning. *Cognitive Science*, 14(1):11–28.

Partha Niyogi. 2006. *The computational nature of language learning and evolution*. MIT Press, Cambridge, MA.

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. 1984. *Language learnability and language development*. Harvard University Press, Cambridge, MA.

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

Steven Pinker and Michael T. Ullman. 2002. The past and future of the past tense. *Trends in Cognitive Science*, 6(11):456–463.

Eva Portelance and Masoud Jasbi. 2023. The roles of neural networks in language acquisition. *PsyarXiv preprint https://doi.org/10.31234/osf.io/b697*.

Grusha Prasad, Marten Van Schijndel, and Tal Linzen. 2019. Using priming to uncover the organization of syntactic representations in neural language models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 66–76.

Shannon M Pruden, Kathy Hirsh-Pasek, Roberta Michnick Golinkoff, and Elizabeth A Hennon. 2006. The birth of words: Ten-month-olds learn words through perceptual salience. *Child development*, 77(2):266–280.

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.

Caitlin Richter. 2021. *Alternation-sensitive phoneme learning: Implications for children's development and language change*. Ph.D. thesis, University of Pennsylvania.

David E. Rumelhart and James L. McClelland. 1986. On learning the past tenses of English verbs. In James L. McClelland, David E. Rumelhart, and the PDP Research Group, editors, *Parallel distributed processing: Explorations into the microstructure of cognition. Volume 2: Psychological and biological models*, pages 216–271. MIT Press, Cambridge, MA.

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.

John C. Trueswell, Irina Sekerina, Nicole M. Hill, and Marian L. Logrip. 1999. The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition*, 73(2):89–134.

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

Marten van Schijndel and Tal Linzen. 2018. A neural model of adaptation in reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4704–4710, Brussels, Belgium. Association for Computational Linguistics.

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.

Héctor Vázquez Martínez, Annika Lea Heuser, Charles Yang, and Jordan Kodner. 2023. Evaluating neural language models as cognitive models of language acquisition. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 48–64, Singapore. Association for Computational Linguistics.

Wai Keen Vong, Wentao Wang, A Emin Orhan, and Brenden M Lake. 2024. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511.

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, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

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.

Kristina Woodard, Lila R Gleitman, and John C Trueswell. 2016. Two-and three-year-olds track a single meaning during word learning: Evidence for propose-but-verify. *Language learning and development*, 12(3):252–261.

Charles Yang. 2002. *Knowledge and learning in natural language*. Oxford University Press, Oxford.

Charles Yang. 2016. *The price of linguistic productivity: How children learn to break rules of language*. MIT Press, Cambridge, MA.

Yuan Yang and Steven T. Piantadosi. 2022. One model for the learning of language. *Proceedings of the National Academy of Sciences*, 119(5):e2021865119.

Aditya Yedetore, Tal Linzen, Robert Frank, and R. Thomas McCoy. 2023. How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9370–9393, Toronto, Canada. Association for Computational Linguistics.

Tian Yun, Chen Sun, and Ellie Pavlick. 2021. Does Vision-and-Language Pretraining Improve Lexical Grounding? In *Proceedings of EMNLP*. ArXiv: 2109.10246.

Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125.