BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language

Philip A. Huebner[1], **Elior Sulem**[2], **Cynthia Fisher**[1], **Dan Roth**[2]
[1]Department of Psychology, University of Illinois at Urbana-Champaign
[2] Department of Computer and Information Science, University of Pennsylvania
{huebner3, clfishe}@illinois.edu
{eliors, danroth}@seas.upenn.edu

Abstract

Transformer-based language models have taken the NLP world by storm. However, their potential for addressing important questions in language acquisition research has been largely ignored. In this work, we examined the grammatical knowledge of RoBERTa (Liu et al., 2019) when trained on a 5M word corpus of language acquisition data to simulate the input available to children between the ages 1 and 6. Using the behavioral probing paradigm, we found that a smaller version of RoBERTa-base that never predicts unmasked tokens, which we term BabyBERTa, acquires grammatical knowledge comparable to that of pre-trained RoBERTa-base - and does so with approximately 15X fewer parameters and 6,000X fewer words. We discuss implications for building more efficient models and the learnability of grammar from input available to children. Lastly, to support research on this front, we release our novel grammar test suite that is compatible with the small vocabulary of child-directed input.

1 Introduction

The rise of Transformer-based language models (TLMs) powered by multi-head self-attention (Vaswani et al., 2017) has been accompanied by impressive performance gains in virtually all areas of NLP. When pre-trained on massive datasets on the order of billions of words, models like GPT-2 (Radford et al., 2019), BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019) achieve impressive scores on benchmarks of natural language understanding tasks (Levesque et al., 2012; Wang et al., 2019). Although TLMs were developed for applications in language technology, their successes raise fundamental questions for acquisition research, including unsupervised grammar induction. To address these questions, however, we require models that are trained on developmentally plausible input, matched in quantity and quality to the input that children are exposed to.

Cognitive scientists have long debated to what extent grammatical knowledge - the categories and structures that determine how words are ordered and inflected in a given language - can be learned from exposure to language input alone (Gómez and Gerken, 2000; Harris, 1954; Reeder et al., 2017), or requires built-in linguistic knowledge or inductive biases (Lidz and Perkins, 2018; Valian, 1986). Demonstrations of strong grammatical knowledge acquired by TLMs trained on raw language data (Warstadt and Bowman, 2020; Zhang et al., 2021) make TLMs a promising new tool for developmental psychologists interested in this debate. However, with rare exceptions, most TLMs have been - and were designed to be - trained on many orders of magnitude more data than is available to children. By the time middle-class English-speaking children have acquired near adult-like grammatical knowledge (by about age 6 (Kemp et al., 2005), though improvements can be detected long after), they have been exposed to no more than 10-50M words (Hart and Risley, 1995) - at least 600 times fewer words than RoBERTa. Current TLMs thus have an enormous data advantage over children, which limits the cognitive conclusions that can be drawn from evaluating off-the-shelf models.

To remedy this, we present an acquisition-friendly version of RoBERTa, BabyBERTa, that is trained on input which is both qualitatively and quantitatively comparable to that of the average English-speaking 6-year-old. While our primary aim is to develop a useful resource for developmental psychologists, we also discuss implications for developing more efficient models that will be of broad interest to the NLP community.

1.1 Aims and Hypotheses

TLMs trained on billion of words operate in a very different regime than language-learning children. Model and training optimizations designed for such

	RoBERTa-base	BabyBERTa
parameters	125M	8M
data size	160GB	0.02GB
words in data	30B	5M
batch size	8K	16
max sequence	512	128
epochs	>40	10
max step	500	260
hardware[1]	1024x V100	1x GTX1080
training time	24hours	2hours
accuracy[2]	81.0	80.5

Table 1: A comparison between RoBERTa-base pre-trained on 30B words of web text (Liu et al., 2019), and BabyBERTa pre-trained from scratch on 5M words of child-directed input. [1]GPU(s) used for training. [2]Average accuracy on our grammar test suite.

massive scales may not work well at smaller scales. Thus, in developing BabyBERTa, we (i) used a smaller model size to avoid over-fitting on the small acquisition data; and (ii) explored a large hyper-parameter space to identify configurations that work at acquisition-scale.

We ask three questions: First, what hyper-parameters and training strategies work best for a TLM in the small-data regime? In particular, can we use insights from work in language acquisition to build more efficient - and better performing - models? Second, how quickly and to what level can a TLM acquire basic English grammatical phenomena when provided with input matched in size to that of the average English-speaking 6-year-old? Third, does child-directed language data play a special role in grammar induction compared to conventional written text data, like Wikipedia articles? Child-directed speech, characterized by shorter sentences, more formulaic constructions, repetition, and reduced lexical diversity (Kirchhoff and Schimmel, 2005; Hayes and Ahrens, 1988; Foushee et al., 2016), may facilitate identification of basic syntactic structures.

1.2 Related Work

In recent years, there has been a surge of interest in linguistic evaluations of language models (Glockner et al., 2018; Ettinger, 2020; Linzen et al., 2016; Gulordava et al., 2018; Goldberg, 2019; Tran et al., 2018; Bacon and Regier, 2019; McCoy et al., 2020; Linzen, 2020). In a review of this literature, Linzen and Baroni (2021) concluded that language models rely on a mixture of syntactic features and shallower heuristics, rather than a principled grammar. Whether the heuristics resemble those used by humans, and whether humans use more principled approaches resembling those in formal linguistics, are outstanding questions.

To isolate the effect of data size on the acquisition of grammatical knowledge, Warstadt et al. (2020b) pre-trained RoBERTa models on datasets varying in size. Their results showed that RoBERTa learns linguistic features with only a few million words, but that it takes billions of words for the model to prefer to use linguistic generalizations over surface ones. Using the same models, Zhang et al. (2021) showed that many grammatical phenomena can be acquired using 100M words of pre-training data, and that for some phenomena such as agreement, the largest improvement occurs between 1M and 10M words. In a similar study, in which left-to-right language models, some enriched with an inductive bias toward hierarchical syntax, were systematically pre-trained on datasets varying in size, Hu et al. (2020) found that many models achieved strong grammatical generalization with only several million words, well within the range of a human learner. Collectively, these results suggest that it is possible to acquire grammatical knowledge from substantially less data. However, it is not yet known if this extends to child-directed speech, which is the topic of this work.

1.3 Contribution

To evaluate the grammatical knowledge of TLMs trained on acquisition data, we created a novel test suite that is compatible with the small vocabulary typical of child-directed language. Using this test suite, we found that RoBERTa-base pre-trained from scratch did not perform well in the small-data regime, and thus explored modifications of RoBERTa-base via extensive hyper-parameter tuning. BabyBERTa has approximately 15X fewer parameters, and when trained on child-directed input - at least 6,000X fewer words than the 30B used to pre-train RoBERTa-base - scored within half a point of RoBERTa-base. See Table 1 for a comparison. Furthermore, BabyBERTa performs well above chance in a majority of 13 grammatical phenomena evaluated. This demonstrates that masked language modeling can yield strong grammatical knowledge even when the input consists of a small corpus of child-directed language.

2 Methods

2.1 BabyBERTa

We introduce a scaled-down masked language model based on RoBERTa (Liu et al., 2019), with 8M parameters, 8912 vocabulary items, and trained on no more than 30M words. Additionally, and more importantly, during training, we modified the probability of unmasking from 0.1 to 0.0 - effectively removing unmasking. We will refer to this model as BabyBERTa to highlight both its origin in RoBERTa and its use-case, small-scale language acquisition experiments. All hyper-parameters were identified by tuning BabyBERTa on a masked word prediction task using a held-out portion of our corpus of transcribed child-directed speech as input. A detailed comparison between hyper-parameters of BabyBERTa and RoBERTa is available in Appendix A. Briefly, BabyBERTa uses only 8 layers, 8 attention heads, 256 hidden units, and an intermediate size of 1024.

In line with RoBERTa, we used dynamic masking. Specifically, we duplicated each input sequence 10 times and applied a novel random mask to each. Consequently, when BabyBERTa is trained on 5M words, BabyBERTa actually receives $5 \times 10 = 50$M words, which approximates the language experience of the average English-learning 6-year-old. BabyBERTa uses fewer epochs than RoBERTa-base which we estimated to be at least 40[1].

2.2 Training Data

Our main corpus of interest is AO-CHILDES (Age-Ordered-CHILDES, Huebner and Willits, 2021). The corpus contains approximately 5M words of American-English transcribed child-directed speech obtained from the CHILDES database (MacWhinney, 2000). Transcripts were created by many different researchers, and consist primarily of in-home recordings of casual speech to children, but also in-lab activities such as book-reading.

To isolate the influence of the unique aspects of child-directed speech on model performance, we also trained BabyBERTa on corpora from two very different domains, matched approximately in size to AO-CHILDES: First, we obtained three small Wikipedia corpora by splitting a random collection of articles in English Wikipedia into three sets of approximately 500K sentences each. This method of splitting *within* articles resulted in three corpora nearly identical in vocabulary and content, which we will refer to as Wikipedia-1, Wikipedia-2, and Wikipedia-3.

Our Wikipedia corpora differ from AO-CHILDES in a number of ways: First, Wikipedia is a corpus of written, not spoken language. Second, many articles were written by professionals with topical expertise and attention to grammatical correctness. Thus, in order to further isolate the effect of domain, we included a fifth corpus, which we will refer to as AO-Newsela, based on the Newsela corpus (Xu et al., 2015). It includes 1,911 English news articles, and 4 or 5 simplified versions of each rewritten by professional annotators for children with different reading proficiency. Each simplification level 1-5 (targeted to grade-levels 2 through 12), contains close to 1M words; AO-Newsela is therefore roughly equivalent in size to AO-CHILDES. Because this corpus contains written language but is directed towards children instead of adults, it is an ideal middle ground between the spoken child-directed language in AO-CHILDES and the written adult-directed language in our Wikipedia corpora.

2.3 Vocabulary

BabyBERTa uses a sub-word vocabulary based on Byte-Pair Encoding, introduced by Sennrich et al. (2016) and later adopted in GPT-2 (Radford et al., 2019) and RoBERTa. Instead of using the original 50K vocabulary used by RoBERTa, we built a custom vocabulary of size 8192 from a concatenation of AO-CHILDES, AO-Newsela and Wikipedia-1, using the open-source Python API *tokenizers*[2]. To make our setting more relevant to the situation faced by human learners, we lowercased all corpora prior to the construction of the vocabulary. Our choice of vocabulary size is informed by studies of children's early vocabulary development, which have estimated that the average English-speaking 6-year-old has acquired a vocabulary of approximately 5,000-6,000 root-words (Biemiller, 2003)[3].

[1] This lower bound is based on dividing the number of batches that can fit the training data by the number of batches reported by Liu et al. (2019), 40×10^9 / 3×10^6 / $500\times10^3 \approx 40$, where 40×10^9 is an upper bound on the number of tokens after Byte-Pair encoding, and 3×10^6 is a lower bound on the number of tokens per batch.

[2] Available at https://github.com/huggingface/tokenizers

[3] Our vocabulary size is slightly larger than 5000-6000 considering that many items are sub-word tokens produced by Byte-Pair tokenization.

2.4 Grammar Test Suite

Our grammar test suite is inspired by BLiMP (Benchmark of Linguistic Minimal Pairs, Warstadt et al., 2020a), a behavioral probe that contains pairs of test sentences which isolate specific phenomena in syntax and morphology, such as island effects and determiner-noun agreement. Each sentence in a pair differs only by a word or a short phrase, and only one sentence in each pair is grammatically acceptable. To succeed, a model must score the grammatical sentence higher than its ungrammatical counterpart. Because of the small vocabulary used by BabyBERTa and the limited overlap between words in child-directed input and the set of words in BLiMP, we made our own grammar test suite. While retaining much of the organisational structure and philosophy of BLiMP, our evaluation data only includes words (and never sub-tokens) from the vocabulary of BabyBERTa and is therefore more sensitive to differences in grammatical knowledge acquired in our setting. We hope that our data[4] will be useful to other researchers interested in simulations of children's grammatical development.

Because one of our aims is to compare grammatical knowledge obtained by BabyBERTa trained on different corpora, we needed a single evaluation dataset that could be used with each corpus without biasing the results. Toward that end, we carefully counterbalanced every word list used to construct sentences (e.g. nouns, adjectives, verbs) such that the total number of occurrences of all word types in a given list was approximately equal (differed no more than 1K) across AO-CHILDES, AO-Newsela, and Wikipedia-1. This allowed us to draw strong conclusions about structural differences across these 3 domains, un-confounded with differences in lexical frequency.

To obtain wide coverage of grammatical phenomena, we reproduced 11 of the 12 phenomena in BLiMP. We excluded *control/raising* due to the lack of enough suitable words in our vocabulary to generate a large and diverse set of minimal pairs with the desired contrast. Within each phenomenon in BLiMP, there are multiple paradigms (types of minimal pairs), and we re-recreated at least one for each phenomenon by randomly selecting a paradigm after removing those which 1) could not be straightforwardly ported to our smaller vocabulary, and 2) did not yield high accuracies in the original work (Warstadt et al., 2020a). We also added 2 phenomena not in BLiMP ("case", and "local attractor" to challenge subject-verb agreement), for a total of 23 paradigms and 13 phenomena. Details regarding word lists, counterbalancing, templates used for generating test sentences, and additional phenomena can be found in Appendix B.

[4]Available at https://github.com/phueb/Zorro/sentences.

2.5 Evaluation Method

For each minimal pair, we computed a model's preference for the grammatical as opposed to ungrammatical sentence. The preference score was calculated by summing the cross-entropy errors at each position in the sentence (Zaczynska et al., 2020). This has the advantage of considering the test sentence as a whole, rather than just the left context of a specific position where surprisal is expected to be high for ungrammatical sentences (Warstadt et al., 2020a; Salazar et al., 2020a). We computed the accuracy by dividing the number of correct choices by the total number of pairs. To enable fair comparisons, we use this method to evaluate all models considered in this paper.

3 Results

3.1 RoBERTa is data-hungry

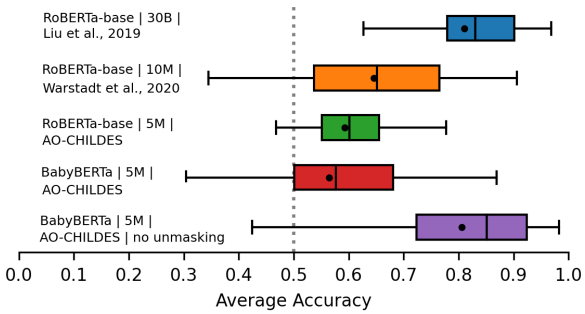

Figure 1: Average accuracy on our grammar test suite. **Blue:** RoBERTa-base pre-trained by Liu et al. (2019) on 30B words. **Orange:** RoBERTa-base pre-trained by Warstadt et al. (2020b) on 10M words. The following models were all trained on AO-CHILDES, 5M words of child-directed input: **Green:** RoBERTa-base pre-trained by us. **Red:** BabyBERTa trained with the original masking strategy, where unmasked tokens are predicted. **Purple:** BabyBERTa trained without predicting unmasked tokens. Box plots illustrate the spread of accuracies across paradigms; the center line marks the median, and the dot marks the mean. Accuracy due to chance is 0.5.

Figure 1 summarizes the accuracies across all

paradigms in our grammar test suite, for RoBERTa-base and BabyBERTa trained on different data[5]. For complete results, see Appendix F. While RoBERTa-base pre-trained on 30B words by Liu et al. (2019) performed reasonably well, the same architecture[6] performed considerably worse when trained only on 5M words of child-directed input. The drop in average accuracy is striking - from 81.1 to 59.2. We also evaluated RoBERTa-base pre-trained from scratch by Warstadt et al. (2020b) on a similarly sized dataset consisting of 10M words of English Wikipedia and Smashwords, which achieved an average accuracy of 64.5, well below 81.1. The poor performance of RoBERTa-base pre-trained from scratch by us and by Warstadt et al. (2020b) clearly indicates that RoBERTa-base does not perform well at acquisition-scale. This then raises the question of whether RoBERTa-base could be adapted to work well at this scale.

The results of a series of manual hyper-parameter tuning studies, culminating in the model we call BabyBERTa, demonstrate that RoBERTa can be straightforwardly adapted to the scale at which acquisition takes place. The full results of hyper-parameter tuning are reported in Appendix A. Briefly, we found that the most important contributions - ordered by increasing impact on performance - to obtaining a high level of accuracy on our grammar test suite are 1) using as input single non-truncated sentences, 2) model size[7], and most importantly, 3) removal of unmasking. While masked tokens are typically unmasked with a probability of 0.1, we found that setting this probability to zero - and thus removing unmasking altogether - yielded an enormous increase in accuracy, from 56.4 to 80.5 - within half a point of RoBERTa-base pre-trained on 30B words.

These results are strong evidence of the data-hungriness of RoBERTa-base, and show that a combination of downsizing the model and never predicting unmasked tokens can dramatically speed acquisition of grammatical knowledge at acquisition-scale. It follows that off-the-shelf models should not be used for language acquisition research without extensive exploration of the hyper-parameter space, and that optimizations that work well for tasks in the NLP community (i.e. predicting unmasked tokens) require scrutinizing prior to adoption in acquisition research.

What additional factors may have contributed to the success of BabyBERTa on our grammatical test suite? First, BabyBERTa was trained exclusively on single sentences, whereas RoBERTa-base was trained on multiple sentences per input. In an offline ablation study, we found that the grammatical competence of BabyBERTa is slightly compromised when training on input that consists of more than one sentence, on the order of 1-2 points. Second, it is possible that BabyBERTa's custom vocabulary - based on the corpora from which words in our test suite are sampled - has a greater coverage of the words in our test suite. However, only 24 out of the 571 content words in our test suite are not in the vocabulary of RoBERTa-base. When we probed RoBERTa-base with proper nouns capitalized - thus achieving full overlap - the accuracy increased by only 1.9 points for RoBERTa-base pre-trained by Liu et al. (2019), and 1.4 points for RoBERTa-base pre-trained by Warstadt et al. (2020b). This rather minor change in the way that word-strings are pre-processed by different models illustrates the way in which aspects of grammatical knowledge in these models is tied to particular forms rather than more abstract patterns.

3.2 Comparing domains

In this experiment, we further examined the grammatical ability of BabyBERTa, to ask whether the strong performance of BabyBERTa is specific to child-directed spoken input, or holds when trained with data that is more representative of that used in the NLP community. In particular, we compared the following three corpora, each representing a different domain of language: AO-CHILDES (spoken, child-directed), AO-Newsela (written, child-directed), and Wikipedia-1 (written, adult-directed). To ensure the validity of this comparison, we trained the model using the same number of steps in each condition. Moreover, because the total number of occurrences of each content word in all our test sentences is closely counterbalanced across all three corpora (see Appendix B for details), any observed differences can be attributed to structural as opposed to word-frequency-related differences between corpora. Further, we adopt the *probing across time* framework used by Liu et al. (2021). Instead of averaging across paradigms, we

[5]For models pre-trained by us, we chose the top-scoring model out of 3 models with a different initialization.

[6]We used the default settings in the Python package *fairseq* v0.10.0 except for batch size=256 and peak learning rate=1e-4 which resulted in higher accuracy.

[7]A combination of fewer hidden units, fewer layers, fewer attention heads, and smaller vocabulary

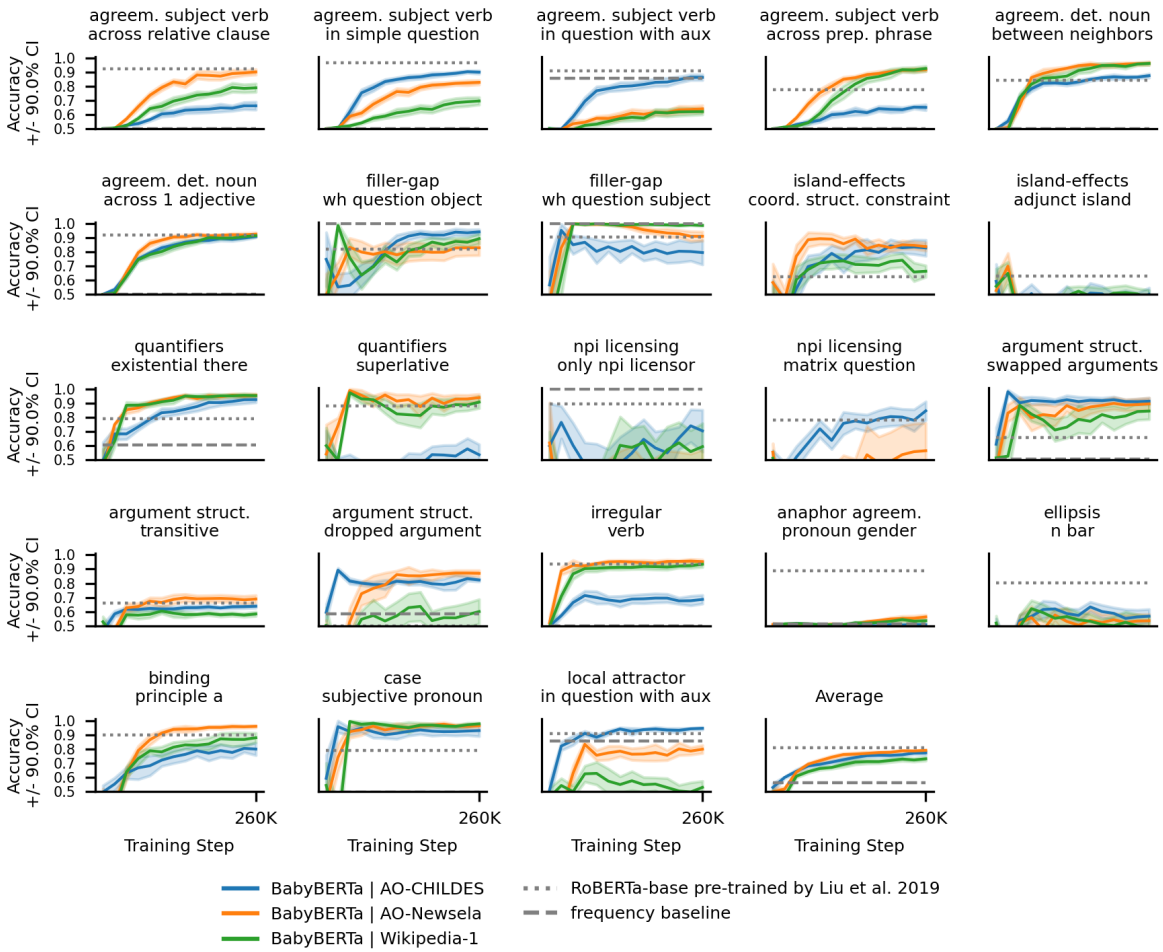

Figure 2: Accuracy on our grammar test suite separated by paradigm for BabyBERTa trained on AO-CHILDES (blue line), AO-Newsela (orange line), or Wikipedia-1 (green line). The word frequency baseline scores a sentence as grammatical if the sum of its word frequencies is greater than its counterpart sentence. Confidence intervals indicate spread across different model initializations (10 per condition).

report the accuracy within each paradigm at consecutive intervals during training. This gives a more in-depth look at when BabyBERTa acquires each grammatical phenomenon, and yields useful information for researchers wishing to compare learning trajectories between TLMs and children.

The results of our corpus comparison are shown in Figure 2. Despite variation, we observed a clear pattern: Models trained on Wikipedia-1 performed well below the others in paradigms that involve questions. This is not surprising, given that questions almost never occur in Wikipedia articles but are frequent in spoken language (at least 40% of total sentences in AO-CHILDES and no more than 1% in Wikipedia-1, see Table 6 in Appendix C). Strikingly, even though AO-CHILDES contains 4X fewer words per sentence (6.4 vs 24.7), the average accuracy of BabyBERTa trained on AO-CHILDES is higher than Wikipedia-1 (77.2 and 73.0, respectively). Furthermore, we found that BabyBERTa trained on AO-Newsela achieves the best overall accuracy (79.0). This is preliminary evidence that that adult-directed written language, which makes up the bulk of data used to train TLMs, is not necessarily the best choice for inducing grammatical knowledge. Instead, it appears that corpora containing shorter sentences (AO-CHILDES) and corpora written for pedagogical purposes (AO-Newsela) can be even more useful. This is indirect evidence that simplified language can boost grammar learning in TLMs.

3.3 Scaffolding

In addition to yielding better grammatical knowledge when training BabyBERTa on AO-CHILDES compared to Wikipedia-1, it is possible that child-directed speech is also a better *starting point* for further learning from more advanced language. For

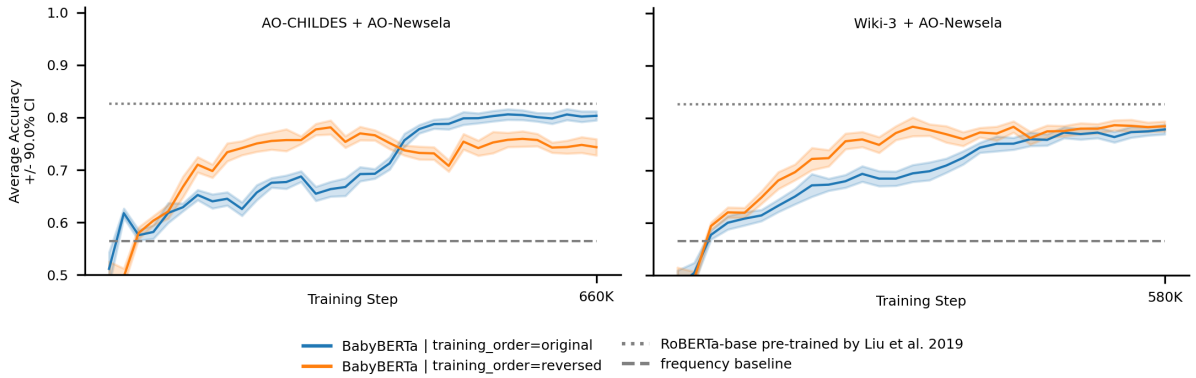

Figure 3: Average accuracy on our grammar test suite, at consecutive intervals during ordered training on concatenated corpora. **Left panel:** BabyBERTa was either trained on AO-CHILDES before AO-Newsela (blue line), or trained on AO-Newsela before AO-CHILDES (orange line). **Right panel:** BabyBERTa was either trained on Wikipedia-3 before AO-Newsela (blue line), or trained on AO-Newsela before Wikipedia-3 (orange line). The average was computed across paradigms and model initializations (10 per condition).

example, it is possible that the grammar that BabyBERTa acquires from AO-CHILDES is less tied to the peculiarities of the input, and would therefore provide better scaffolding for grammar induction after a domain-shift. Some authors have suggested that the formulaic structure of child-directed language, such as high-entropy slots in high-frequency frames, make it easier to discover part-of-speech-classes (Huebner and Willits, 2021; Cameron-Faulkner et al., 2003; Matthews and Bannard, 2010), which, in turn, could accelerate the acquisition of grammar.

To investigate the consequences of prior language experience, we examined how the ordering of examples during training might impact grammatical knowledge during and at the end of training. To accomplish this, we switched from our previous training method which samples sentences randomly to one that selects sentences in the order in which they appear in the data. We conducted two experiments: We trained BabyBERTa either on a concatenation of AO-CHILDES and AO-Newsela, or Wikipedia-3 and AO-Newsela, and manipulated the order of training (order of concatenation vs. reverse)[8].

The results are shown in Figure 3. The left panel clearly illustrates a benefit of training on AO-CHILDES before AO-Newsela compared to AO-Newsela before AO-CHILDES on our grammatical knowledge test suite. At the end of training, overall accuracy is 80.3 for models trained on AO-CHILDES first, and 74.3 for models trained on AO-Newsela first. The direction of this effect is what one would predict under the assumption that input to children aged 1-6 years but not beyond (6-12 years) scaffolds grammatical development. To test that this finding did not result simply because BabyBERTa performs better when trained on AO-Newsela *last*, and instead due to the scaffolding effects of early exposure to AO-CHILDES, we repeated the same experiment but with Wikipedia-3 in place of AO-CHILDES. The results, shown in the right panel of Figure 3 confirm this. When trained on AO-Newsela last, overall accuracy at the end of training was not statistically different from training on AO-Newsela first (77.8 ± 0.92 vs. 78.4 ± 0.94, respectively[9]). A breakdown of the results by paradigm are available in Appendix F.

3.4 Domain diversity

Finally, we asked whether we could improve the grammatical knowledge acquired by BabyBERTa by extending the training data. We were particularly interested in comparing two conditions: Training on a concatenation of three corpora from different domains (AO-CHILDES + AO-Newsela + Wikipedia-3) vs. a size-matched concatenation of our 3 Wikipedia corpora (Wikipedia-1 + Wikipedia-2 + Wikipedia-3). In line with work by Hu et al. (2020), who found an advantage of training on more diverse data in a small-data setting, we predicted that our diverse data condition would yield

[8]Note that this training method produces overall worse results on the grammar test suite compared to random sampling. This occurs because, when sampling in order, identical sentences with different mask patterns are trained as part of the same batch as opposed to different batches spread uniformly across training steps - the latter ensures that batches contain maximally diverse samples.

[9]mean ± margin of error, with $\alpha = 0.05$

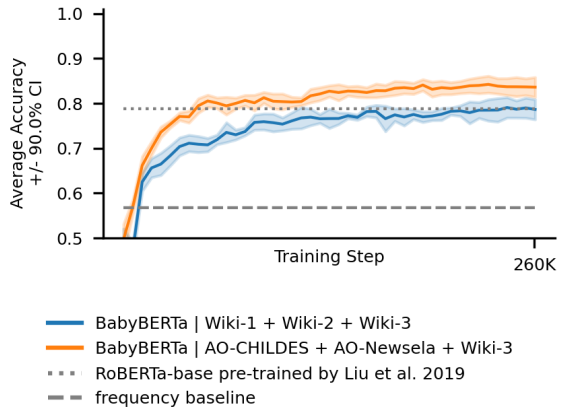

Figure 4: Average accuracy on our grammar test suite across time. BabyBERTa was trained either on Wikipedia only (blue line), or a more diverse set of corpora, AO-CHILDES + AO-Newsela + Wikipedia-3, spanning 3 different domains (orange line).

Model & Data	Unmasking	Accuracy	
		holistic	MLM
BabyBERTa			
+AO-CHILDES	no	77.9	78.0
+AO-CHILDES	yes	60.5	77.6
RoBERTa			
+AO-CHILDES	yes	59.9	72.4
+Warstadt, 2020	yes	64.4	78.1
+Liu et al., 2019	yes	82.5	90.2

Table 2: Overall accuracy on our grammar test suite for two different scoring methods. The holistic method was proposed by Zaczynska et al. (2020) and used to tune BabyBERTa. The MLM method was introduced by Salazar et al. (2020b). Overall accuracy values are averages over 10 replications (BabyBERTa), and 3 replications (RoBERTa-base except Liu et al., 2019).

better performance on our grammar test suite.

The results of this experiment are shown in Figure 4. First, we observed that BabyBERTa trained in the diverse training data condition achieved an average accuracy well above that of pre-trained RoBERTa-base, which is notable given that the amount of data BabyBERTa was exposed to is still far less than that of RoBERTa-base pre-trained on 30B words. Further, as predicted, we found that the average accuracy of BabyBERTa trained only on Wikipedia was worse (83.8 vs. 78.9). While our results support the importance of pre-training on a diverse set of data, they also point to the benefit of including simplified language in pre-training data of TLMs.

3.5 Holistic vs. MLM scoring

The results reported in previous sections were computed using a method of scoring grammaticality where each sentence in a minimal pair is input to BabyBERTa in whole and without masks. There is, however, an alternative method for scoring grammaticality, recently proposed by Salazar et al. (2020b), where each candidate sentence is input to a masked language model multiple times, each time with a mask in a different position. The score is the sum of the log-loss computed at each masked position in the sentence. This method yields improved results on BLiMP, and other tasks that require sentence-level scores. For clarity, we refer to our method as holistic scoring, and that of Salazar et al. (2020b) as MLM scoring. In this section, we compared these two methods by re-scoring BabyBERTa on our grammatical test suite using MLM scoring.

The results are shown in Table 2. We noticed an interaction between the way in which a model was trained (unmasking = yes/no) and the scoring method (holistic vs. MLM). For all models trained using the standard masking strategy, in which masked words are left unmasked 10% of the time (unmasking = yes), overall accuracy was between 8 and 16 points higher when MLM scoring was used. This includes all RoBERTa models, and BabyBERTa trained to predict unmasked tokens (unmasking = yes). The only model whose accuracy did not change noticeably is BabyBERTa (unmasking = no). These results reveal that training with unmasking introduces a handicap which makes models trained using standard masking (unmask = yes) reliant on the insertion of masks during evaluation in order to perform well. However, training without unmasking avoids this handicap, and the resulting models can be used with both evaluation methods without loss in performance. Finally, from a cognitive plausibility perspective, holistic scoring resembles much more closely the actual situation faced by humans tasked to judge grammatical acceptability; training models to perform well with holistic scoring should be considered in future work.

4 Discussion

Transformer based language models (TLMs) extract linguistic generalizations from raw, unanno-

tated language data. Their impressive scores on grammatical benchmarks (Warstadt et al., 2020a; Hu et al., 2020) can be used to estimate lower bounds for how much linguistic knowledge - in principle - can be acquired based on word-string input alone, and what kinds of architectures make this learning possible.

However, the TLMs used in such studies were not trained on language matched in quantity and quality to the input children receive. When trained on massive datasets, TLMs operate in a very different regime, in which it becomes possible to memorize a large number of individual observations, and/or support abstraction with orders of magnitude more data than are available to children. Therefore, claims about the grammatical proficiency of existing off-the-shelf TLMs cannot inform questions about children's grammatical development.

At minimum, claims about what children might learn without the aid of built-in linguistic knowledge should be based on models trained on developmentally plausible datasets. Towards this end, we developed BabyBERTa, a variation of RoBERTa-base with 15X fewer parameters. When trained only on 5M words of child-directed input, BabyBERTa achieved an overall accuracy on our grammar test suite competitive with RoBERTa-base. Performance, however, was far from perfect, and BabyBERTa was at chance in evaluating the grammaticality of sentence pairs that contrast negative-polarity-item (NPI) licensing, gender agreement, ellipsis, superlative quantifiers, and island effects involving adjuncts. More research is needed using a larger set of grammatical phenomena, and child-directed input in languages other than English, before making strong conclusions about learnability.

4.1 Limitations and Future Directions

BabyBERTa never predicts unmasked tokens, unlike RoBERTa-base which inherits this method from BERT (Devlin et al., 2019). The motivation for predicting unmasked tokens, according to Devlin et al. (2019), is to habituate the model to input which does not include mask symbols during fine-tuning. Given that unmasking is supposed to optimize performance *on downstream tasks*, it is not surprising that unmasking handicapped BabyBERTa *during pre-training*. There appears to be a fundamental trade-off between encoding useful representations during pre-training (which requires masking), and the ability to put those representations to work in the context of a downstream task (which typically does not involve masking). When predicting an unmasked token, the prediction task can be solved trivially by outputting the token that is already at the input. From the point of view of grammatical development, this is wasted computation; in order to learn about structural relationships between words, a model must learn to attend to more than one word at a time. Without unmasked tokens in the input, BabyBERTa is forced to attend to lexical context, instead of relying on the input to make predictions. However, this may leave BabyBERTa vulnerable to the distribution-shift that occurs when fine-tuning on a downstream task. More research is needed to quantify this vulnerability, and alternatives to protect against it. More generally, the evaluation on downstream tasks of TLMs pre-trained on child-directed input will allow us to further connect between progress in NLP and the psycholinguistics literature (Linzen, 2020).

To improve the validity of comparisons between spoken versus written corpora, more studies using spoken language are needed, while taking into consideration how language in these two domains differ.[10]

Lastly, while statistical relationships between linguistic forms are one of the sources of information used by children (Gómez and Gerken, 2000), they are certainly not the only type of evidence children rely upon. Our work focused on word-string input, leaving for future work the study of interactions with other linguistic information sources such as prosody (De Carvalho et al., 2017), and other modalities such as sound and vision (Goodman et al., 2007).

5 Conclusion

While child language acquisition research and NLP have largely developed independently, we think that TLMs present a promising opportunity for inter-disciplinary researchers to gain new insight into fundamental questions about the learnability of grammar. By using TLMs to study what is learnable and not learnable given language available to children, work in this area has the potential to shed new light on debates concerning the innate structure necessary for language acquisition. To support research on this front, we proposed and released BabyBERTa, a TLM trained and optimized on developmentally plausible language data.

[10] For samples of child-directed speech, see Appendix D.

Acknowledgements

This research was supported by a grant from the NICHD (HD-054448) and by Contracts FA8750-19-2-0201 and FA8750-19-2-1004 with the US Defense Advanced Research Projects Agency (DARPA). Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Geoff Bacon and Terry Regier. 2019. Does BERT agree? Evaluating knowledge of structure dependence through agreement relations. *arXiv preprint arXiv:1908.09892*.

Ellen Gurman Bard and Anne H Anderson. 1994. The unintelligibility of speech to children: effects of referent availability. *J. Child Lang.*, 21(3):623–648.

Andrew Biemiller. 2003. Vocabulary: Needed if more children are to read well. *Reading psychology*, 24(3-4):323–335.

Thea Cameron-Faulkner, Elena Lieven, and Michael Tomasello. 2003. A construction based analysis of child directed speech. *Cogn. Sci.*, 27(6):843–873.

Alex De Carvalho, Isabelle Dautriche, Isabelle Lin, and Anne Christophe. 2017. Phrasal prosody constrains syntactic analysis in toddlers. *Cognition*, 163:67–79.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Ruthe Foushee, Tom Griffiths, and Mahesh Srinivasan. 2016. Lexical complexity of child-directed and overheard speech: Implications for learning. In *CogSci*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, page 650–655.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Rebecca L Gómez and LouAnn Gerken. 2000. Infant artificial language learning and language acquisition. *Trends in cognitive sciences*, 4(5):178–186.

Noah Daniel Goodman, Joshua B Tenenbaum, and Michael C Frank. 2007. A bayesian framework for cross-situational word-learning. In *Proc. of Advances in Neural Information Processing Systems*.

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.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

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

Donald P Hayes and Margaret G Ahrens. 1988. Vocabulary simplification for children: A special case of 'motherese'? *Journal of child language*, 15(2):395–410.

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*, Stroudsburg, PA, USA. 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. Elsevier.

Nenagh Kemp, Elena Lieven, and Michael Tomasello. 2005. Young children's knowledge of the "determiner" and "adjective" categories. *J. Speech Lang. Hear. Res.*, 48(3):592–609.

Katrin Kirchhoff and Steven Schimmel. 2005. Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *The Journal of the Acoustical Society of America*, 117(4):2238–2246.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*.

Jeffrey Lidz and Laurel Perkins. 2018. Language acquisition. In *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, pages 1–49. Wiley.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annu. Rev. Linguist.*, 7(1):195–212.

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

Leo Z Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A Smith. 2021. Probing across time: What does roberta know and when? *arXiv preprint arXiv:2104.07885*.

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. 2000. *The CHILDES Project: Tools for analyzing talk. transcription format and programs*, volume 1. Psychology Press.

Danielle Matthews and Colin Bannard. 2010. Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child-directed speech. *Cognitive science*, 34(3):465–488.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Patricia A Reeder, Elissa L Newport, and Richard N Aslin. 2017. Distributional learning of subcategories in an artificial grammar: Category generalization and subcategory restrictions. *Journal of memory and language*, 97:17–29.

Douglas Roland, Frederic Dick, and Jeffrey L Elman. 2007. Frequency of basic english grammatical structures: A corpus analysis. *J. Mem. Lang.*, 57(3):348–379.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020a. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020b. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Ke Tran, Arianna Bisazza, and Christof Monz. 2018. The importance of being recurrent for modeling hierarchical structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4731–4736.

Virginia Valian. 1986. Syntactic categories in the speech of young children. *Dev. Psychol.*, 22(4):562–579.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of NeurIPS*.

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

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020a. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8(0):377–392.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R Bowman. 2020b. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Karolina Zaczynska, Nils Feldhus, Robert Schwarzenberg, Aleksandra Gabryszak, and Sebastian Möller. 2020. Evaluating German transformer language models with syntactic agreement tests. In *Proceedings of the 5th Swiss Text Analytics Conference and the 16th Conference on Natural Language Processing, SwissText/KONVENS 2020, Zurich, Switzerland, June 23-25, 2020*, volume abs/2007.03765, Zurich, Switzerland. CEUR Workshop Proceedings.

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)"*, page 1112–1125.

A Implementation Details and Reproducibility

Data and code for training BabyBERTa is available at `https://github.com/phueb/BabyBERTa`.

BabyBERTa is implemented in PyTorch with the Python package *transformers* 4.3.3 (Wolf et al., 2019). It is the result of a large hyper-parameter tuning effort, the results of which are reported in Table 3 below. Tuning was performed manually, and separately for each hyper-parameter, by sampling at least two values larger and two values smaller than the default value, and training at least 3 models in each condition. Hyper-parameters not shown in the table were not considered, and were left at their default values. We have organized the rows into three sections: Hyper-parameters for which we found values that positively impacted accuracy either considerably, or modestly, are shown in the first and second sections, respectively. Hyper-parameters which were included in our exploration but for which we did not find values yielding improved accuracy are shown in the third section of rows.

Because BabyBERTa was developed for use with single-GPU training, we did not explore large batch sizes, and found that a batch size of 16 combined with a learning rate of 1e-4 and a maximum sequence length of 128 worked well. In part due to the constraint on the maximum sequence length and due to psychological plausibility, each input sequence to the model is exactly 1 sentence. If the number of tokens (after Byte-Pair tokenization) in a sentence is longer than the maximum, we do not truncate, and instead remove the sentence altogether. Removal of sentences is rare and does not noticeably impact results. However, training on more than one sentence per input sequence does noticeably impact performance on our grammar test suite.

B Grammar test suite

Virtually all benchmarks, tasks, and evaluations used by the NLP community are based on adult vocabularies, which makes it difficult, if not impossible, to evaluate the grammatical competence of models that simulate language acquisition in children with common words that have simple meanings (e.g. *dog*, *toy*, *green*, *old* vs. *constitution*, *liability*, *exposed*, *historical*). Furthermore, because we are primarily interested in the grammatical knowledge that school-age children should possess, we used relatively simple templates for generating test sentences, as opposed to longer and more involved templates used to construct sentences in BLiMP. This should not bias performance towards models trained on language acquisition data in principle - though it might in practice - because adult-directed language is constrained by the same set of grammatical principles.

To construct test sentences, we first generated several sentence templates for each paradigm, and then inserted content words by randomly sampling from counterbalanced lists of nouns, verbs, or adjectives, depending on the part-of-speech required for a particular slot in a template. Word lists were created by 1) manually identifying whole words in BabyBERTa's vocabulary, and 2) counterbalancing. Counterbalancing was performed to ensure that content words in our test sentences are approximately equally distributed across all 5 corpora[11]. Counterbalancing total word frequency across conditions is standard procedure in modern psycholinguistics experiments, and is essential when explicitly comparing models trained on different corpora. To ensure counterbalancing worked as expected, we report the number and proportion of test words in each of our corpora in Table 4.

We added 2 phenomena not in BLiMP. In the first, which we refer to as "case", we included 2 paradigms, each contrasting a different case of pronouns (objective vs. possessive or subjective). In

[11] Because all three Wikipedia corpora consist of sentences drawn from the same set of articles, their vocabulary is virtually identical. This means that only one was needed for counterbalancing.

Hyper-parameter	Roberta-base	BabyBERTa
Considerable Improvement		
`leave_unmasked_prob`	0.1	0
layers	12	8
attention heads	12	8
hidden size	768	256
intermediate size	3072	1024
vocabulary size[1]	50265	8192
Moderate Improvement		
batch size	8K	16
sentences per sequence	unlimited	1
peak learning rate	6e-4	1e-4
weight decay	0.01	0.0
layer norm epsilon	1e-5	1e-5
add prefix space[2]	False	True
No Improvement		
allow truncation of input	False	False
weight initializer range	0.02	0.02
warm-up steps	24K	24K
`random_token_prob`	0.1	0.1
mask probability	0.15	0.15
maximum sequence length	512	128
include punctuation	True	True
epochs	<40	10

Table 3: A comparison of hyper-parameters used for pre-training RoBERTa-base, and those used to train BabyBERTa on AO-CHILDES. Hyper-parameters not shown in this table were not systematically explored by us, and were left at their default values in *transformers* v4.3.3. [1]Our vocabulary is not case-sensitive like RoBERTa-base. [2]We added a space prefix to each token in our custom trained Byte-Pair vocabulary so that the model does not treat tokens at the beginning of a sequence differently.

the second, which we refer to as "local attractor", we included 1 paradigm which does not isolate a specific grammatical rule, but contrasts a well-formed question with the same question in which the the the main verb is changed from the infinitive to agree with the subject. The result of this modification (e.g. *can the husband change ?* → **can the husband changes ?*) produces well-formed bi-and tri-grams but is not, as a whole, grammatical. We included this paradigm to distinguish models that prefer locally well-formed strings (conceptually similar to attractors) at the expense of grammatically.

One notable feature of our test sentences is their lack of semantic plausibility. This is, to some extent, also true of sentences in BLiMP, although to a lesser degree due to selectional restrictions on verbs which were not used in this work. At first, this may appear as an oversight, but we did so purposefully - following work by Gulordava et al. (2018) - in order to remove any semantic or other lexical clues that BabyBERTa might use to artificially inflate its score. In principle, the grammatical phenomena which we evaluate apply independently of the choice of content words, as long as grammatically relevant information such as gender and number are taken into consideration. Hence, a model that has acquired grammatical knowledge independent of specific lexical associations should score higher on our evaluation data than a model which relies on other kinds of information.

There are several limitations worth mentioning. First, despite our best efforts at counterbalancing lists of content words, our procedure is blind to different senses of a word. This can lead to situations in which a model trained on one corpus is familiar

with the sense of the word as used in our evaluation data, while another is not. To remedy this, we excluded content words without a strong dominant sense. Second, we did not compute agreement between our evaluation data and human judgements. However, because our templates and procedures for generating test sentences resemble those used to generate BLiMP we suspect that agreement would be close to that of BLiMP (human aggregate agreement with grammaticality is 96.4%).

Examples of grammatical and ungrammatical sentences for each of our paradigms are shown in Table 5. The full data is available at https://github.com/phueb/Zorro/sentences.

C Training data

C.1 AO-CHILDES

The AO-CHILDES was described by (Age-Ordered-CHILDES, Huebner and Willits, 2021). We did not perform additional pre-processing except for replacing periods with question marks for questions incorrectly marked with periods.

C.2 AO-Newsela

AO-Newsela is derived from raw data available in the Newsela Article Corpus, Version: 2016-01-29. The raw data is proprietary and can be requested at https://newsela.com/data/. To build AO-Newsela, we excluded Spanish articles and removed article headings.

Code for building AO-CHILDES and AO-Newsela from raw data are available at https://github.com/UIUCLearningLanguageLab.

C.3 Wikipedia

To build our Wikipedia corpora, we downloaded an English-Wikipedia dump on February, 2021, and used the python package *witokit* to extract text, available at https://github.com/akb89/witokit.

All corpora were lower-cased, and white space was inserted between the last word in every sentence and the punctuation symbol. Sentences shorter than 3 words were excluded before pretraining.

Although our corpora are roughly equal in size, it was impossible to equate each on the number of total words and the number of total sentences simultaneously. This is due to the fact that sentences in our Wikipedia corpora are much longer than sentences in AO-CHILDES and AO-Newsela. See Table 6 for a summary. Because input to BabyBERTa consists of individual sentences, and the number of training steps is therefore directly proportional to the number of sentences rather than the number of words per sentence, we controlled for differences in the number of words by halting training at a pre-determined step. Given that AO-Newsela contains the fewest number of sentences, the maximum step for all corpora is based on the maximum number of training steps possible when training BabyBERTa on AO-Newsela (260K steps). While this procedure equates the number of steps between simulations using different corpora, a model trained on any of the Wikipedia corpora nonetheless is exposed to far more words compared to the other two corpora. See Table 6 for a summary.

D Samples from AO-CHILDES

AO-CHILDES differs considerably from standard corpora used in the NLP community. This is due to at least four factors: First, language is directed to children below the age of six. Second, producers of the language in AO-CHILDES are caregivers who are unlikely to be professional writers and often have very different goals than creators of, say, Wikipedia articles. Third, the language in AO-CHILDES is spoken as opposed to written. This is an important point because speakers may employ non-standard dialects and/or contractions (e.g. *isn't*, *wanna*) that are often not represented in benchmarks, which could bias such evaluations towards models trained on written language. Fourth, spoken language does not include symbols such as parentheses, colons, dashes and other textual markers that potentially help to emphasize clause and/or phrase boundaries. We thought it would be helpful to provide some samples from AO-CHILDES, presented in Table 7 as this corpus is not typically used by the NLP community.

For more information about the frequency of syntactic constructions in child-directed language, see Roland et al. (2007)

E Additional Experiments

E.1 BLiMP

We also evaluated BabyBERTa on the Benchmark of Linguistic Minimal Pairs (BLiMP) (Warstadt et al., 2020a). The accuracies for each phenomenon are shown in Table 8, and the overall accuracy

Corpus	Test words in corpus	Proportion of test words in corpus
AO-CHILDES	3,477,273	0.206
AO-Newsela	3,406,535	0.201
Wikipedia-1	3,344,313	0.198
Wikipedia-2	3,342,946	0.198
Wikipedia-3	3,341,654	0.198

Table 4: The number and proportion of test words (content words occurring in our evaluation data) in each of our corpora. The fact that values do not differ dramatically across rows confirms that counterbalancing was successful.

Phenomenon	Paradigm	Examples	
		Well-formed	Not well-formed
Det-subject agreement	across_1_adjective	*look at this happy piece .*	*look at this happy pieces .*
	between_neighbors	*this color must be commercial .*	*this colors must be commercial .*
Subject-verb agreement	across_prepositional_phrase	*the brother by the lion is red .*	*the brothers by the lion is red .*
	across_relative_clause	*the pages that i like were dirty .*	*the page that i like were dirty .*
	in_question_with_aux	*where does the bird go ?*	*where does the birds go ?*
	in_simple_question	*what color was the piece ?*	*what color was the pieces ?*
Anaphor agreement	pronoun_gender	*she will give herself the wire .*	*she will give himself the wire .*
Argument structure	dropped_argument	*my brother moves fast .*	*my brother moves to .*
	swapped_argument	*they built the mouse that farm .*	*the mouse built that farm they .*
	transitive	*will robert eat ?*	*will robert force ?*
Binding	principle_a	*sarah thinks about herself making a tree .*	*sarah thinks about herself makes a tree .*
Case	subjective_pronoun	*they gave the person the tour .*	*the person gave they the tour .*
Ellipsis	n_bar	*allen got one roman brain and chris got two .*	*allen got one brain and chris got two roman .*
Filler-gap	question_object	*laura got the suit that the bird cut .*	*laura got what the suit cut the bird .*
	wh_question_subject	*chris reached the bear that is washing trains .*	*chris reached who the bear is washing trains .*
Irregular	verb	*sarah spoke without thinking last night .*	*sarah spoken without thinking last night .*
Island effects	adjunct_island	*what did robert eat while facing the kiss ?*	*what did robert eat the kiss while facing ?*
	coord_struct_constraint	*what did sarah and the person work for ?*	*what did sarah work for and the person ?*
Local attractor	in_question_with_aux	*can the husband change ?*	*can the husband changes ?*
NPI licensing	matrix_question	*would william ever keep the movie ?*	*william would ever keep the movie ?*
	only_npi_licensor	*only his rabbit will ever be in her magic .*	*even his rabbit will ever be in her magic .*
Quantifiers	existential_there	*there was a leg that anne made .*	*there was most leg that anne made .*
	superlative	*no bird could catch more than six plants .*	*no bird could catch at least six plants .*

Table 5: Examples of well-formed and not well-formed sentences for each paradigm in our grammar test suite. Each paradigm consists of 4,000 sentences (2,000 minimal pairs).

is shown in Table 9. Notice that the overall accuracy for BabyBERTa on BLiMP is considerably lower than the average accuracy on our grammar test suite, which closely resembles the format of BLiMP. This is most likely due to two reasons: First, the set of words used in BLiMP likely do not perfectly overlap with the words in our small corpora and/or the vocabulary of BabyBERTa. For example, many proper nouns and verbs in BLiMP simply never occur in the corpora on which BabyBERTa is trained. Second, the paradigms in BLiMP we chose to re-implement were chosen based on which received the largest accuracy scores by TLMs - meaning that its easier to score higher on our benchmark. Collectively, this suggests that the results for BLiMP underestimate the grammatical knowledge acquired by BabyBERTa.

Furthermore, when trained on AO-CHILDES, we observed that BabyBERTa only scores higher than RoBERTa-based when BabyBERTa is trained with a larger vocabulary (the original RoBERTa-base vocabulary with 50K tokens). This confirms that existing benchmarks of grammatical knowledge are biased towards large vocabularies, and thus under-estimate models trained using smaller vocabulary, or on child-directed input which is naturally constrained in the number of word types.

This leaves unanswered why BabyBERTa with a vocabulary of 8K does not out-perform RoBERTa-base on BLiMP despite our finding in the main text that BabyBERTa achieved a larger overall ac-

Corpus	Sentences	Avg sentence length		Questions (proportion)
		Sub-tokens	Words	
AO-CHILDES	723,524	7.33	6.38	0.42
AO-Newsela	442,571	22.37	15.97	0.01
Wikipedia-1	525,917	31.71	24.77	0.00
Wikipedia-2	525,903	31.71	24.78	0.00
Wikipedia-3	525,352	31.74	24.80	0.00

Table 6: Descriptive statistics for each of our corpora. The reported number of sentences was computed after excluding sentences that contain more than 128 sub-word tokens. The precise number of sentences is irrelevant, because we control for data quantity by stopping training at a pre-defined number of steps. The proportion of questions was determined based on counting question marks.

Property	Example
Interruptions and false starts	*it is about three four feet away .* *that is trouble with it if you .* *here let's find ah the gorilla .*
Dialect/grammatical error	*is that what you talking about .* *and i absolutely will never not ever eat tomatoes .* *and i absolutely will never not ever eat tomatoes .*
Contraction	*you wanna go play ?*
Nursery rhyme and song	*with a knickknack paddywack give a dog a bone...* *with an oink and a moo and a quack quack .*
Intonation marking[1]	*that is a real nice building ?* *want me to hold that !* *is it all gone !*
Made up and alternate word forms	*want to floppity ?* *what does a doggie say ?*
Interjections	*oh here's a car* *aha that's ring around the roses ?*
Onomatopoeia	*vroom vroom vroom vroom .* *they go ruff ruff ruff .*

Table 7: Examples of utterances from AO-CHILDES illustrating frequent and/or unique properties of spoken child-directed language data. [1]Intonation is often marked using exclamation or question marks, even when such marking is incompatible with grammatical rules; transcription errors also give rise to incompatible utterance boundary markers.

curacy on our own grammar test suite. We think this is related to the fact that we used a different method for scoring grammaticality, based on the pseudo-log-likelihood proposed by Salazar et al. (2020a). Instead of inputting a whole sentence to the model once and computing the sum of cross-entropy errors as we did throughout this paper, the pseudo-log-likelihood is computed by 1) creating copies of a sentence with each token masked out, and then 2) summing the log probability for each missing token over copies. In this way, a model is never given information at the input about any word it is supposed to predict at the output layer. In fact, this completely circumvents the handicap related to prediction of unmasked tokens; without this handicap, RoBERTa-base no longer performs sub-optimally relative to BabyBERTa (which derived its advantage primarily by never predicting unmasked tokens).

In sum, the method used for scoring the grammaticality of sentences can bias performance in predictable ways. It is important to evaluate grammatical knowledge using multiple benchmarks and multiple scoring methods in order to obtain an accurate picture of a model's abilities.

E.2 Unmasking curriculum

We hypothesized that a curriculum strategy might provide a middle ground between the benefit of removing unmasking during pre-training on acquisition of grammatical knowledge and the benefit provided by pre-training with standard unmasking for readying a model for fine-tuning. Inspired by the work of Bard and Anderson (1994) who found that words in child-directed speech tend to be less intelligible than words in adult speech, and the fact that auditory word recognition of infants is initially far from perfect, we gradually increased the probability that BabyBERTa predicts unmasked tokens over the course of training. Specifically, we set the probability that a masked token is left unmasked to 0.0, and linearly increased this probability to 0.1, the original value used by BERT and RoBERTa. In order to slow the curriculum, we trained BabyBERTa on a concatenation of AO-CHILDES, AO-Newsela, and Wikipedia-1. This slowed the curriculum by a factor of 3X, which we hypothesized would be necessary to reduce the handicapping due to predicting unmasked tokens.

The results of this experiment are shown in Figure 5. We hypothesized that prediction of unmasked tokens would be most detrimental for grammatical development during early training, and that by slowly increasing the probability that unmasked tokens are predicted towards the end of training, the model would be less handicapped than a model that was predicting unmasked tokens from the start. However, we found that, when BabyBERTa was trained using our curriculum strategy (blue line), while initially achieving much higher scores on our grammar test suite, performed no better than BabyBERTa with standard unmasking (orange line), at the end of training. For reference, we included a condition in which BabyBERTa was trained without unmasking throughout training (green line).

F Complete results

Due to space limitations in the main paper, we reported only average overall accuracy on the Zorro test suite in sections 3.1, 3.3, and 3.4. Below is a complete view of accuracy scores separated by paradigm.

Model	Anaphor agreement	Argument structure	Binding	Control/raising	Determiner-noun agreement	Ellipsis	Filler-gap	Irregular forms	Island effects	Npi licensing	Quantifiers	Subject-verb agreement
BabyBERTa												
+AO-CHILDES	70.3	54.8	63.3	54.8	78.8	54.4	68.4	70.3	53.0	53.9	61.3	56.0
+50K vocab	74.8	57.9	64.0	59.0	75.4	49.4	67.4	82.6	56.4	56.1	59.5	56.2
+AO-Newsela	75.6	61.5	67.9	56.1	82.9	61.4	63.9	74.2	36.1	56.5	62.8	73.3
+Wikipedia-1	62.9	59.7	70.6	60.4	80.0	74.6	61.2	75.8	52.8	54.2	54.6	68.5
+concat.[1]	74.3	64.2	72.3	59.9	87.6	69.3	70.5	81.1	48.6	60.2	57.9	83.5
RoBERTa-base												
+AO-CHILDES	65.0	57.1	63.9	59.2	69.6	65.1	55.7	69.7	56.2	58.5	67.6	57.6
+10M[2]	88.6	69.9	72.6	70.0	91.4	85.3	67.0	85.5	52.1	71.9	58.2	77.0
+30B[3]	97.3	83.5	77.8	81.9	97.0	91.4	90.1	96.2	80.7	81.0	69.8	91.9

Table 8: Accuracy on each phenomenon in BLiMP (Benchmark of Linguistic Minimal Pairs). We used the MLM-scoring method proposed by Salazar et al. (2020a) based on pseudo-log-likelihoods to ensure our results are comparable to accuracies reported in their work. We chose a random BabyBERTa model from the 10 models we trained for each corpus. [1]BabyBERTa trained on a concatenation of AO-CHILDES, AO-Newsela, and Wikipedia-1. [2]RoBERTa-base pre-trained on 10M words by Warstadt et al. (2020b). While this model is trained on data comparable in size to BabyBERTa, it uses a much larger vocabulary (50K tokens vs. 8K tokens) [3]Roberta-base pre-trained on approx. 30B words by Liu et al. (2019).

Model	**BLiMP Overall accuracy**
BabyBERTa	
+AO-CHILDES	61.6
+AO-CHILDES+50K vocab	63.2
+AO-Newsela	64.4
+Wikipedia-1	64.6
+concat.[1]	69.1
RoBERTa-base	
+AO-CHILDES	62.1
+10M[2]	74.1
+30B[3]	85.4

Table 9: Overall accuracy on BLiMP (Benchmark of Linguistic Minimal Pairs). We used the MLM-scoring method proposed by Salazar et al. (2020a) based on pseudo-log-likelihoods to ensure our results are comparable to accuracies reported in their work. We chose a random BabyBERTa model from the 10 models we trained for each corpus. [1]BabyBERTa trained on a concatenation of AO-CHILDES, AO-Newsela, and Wikipedia-1. [2]RoBERTa-base pre-trained on 10M words by Warstadt et al. (2020b). While this model is trained on data comparable in size to BabyBERTa, it uses a much larger vocabulary (50K tokens vs. 8K tokens) [3]Roberta-base pre-trained on approx. 30B words by Liu et al. (2019).

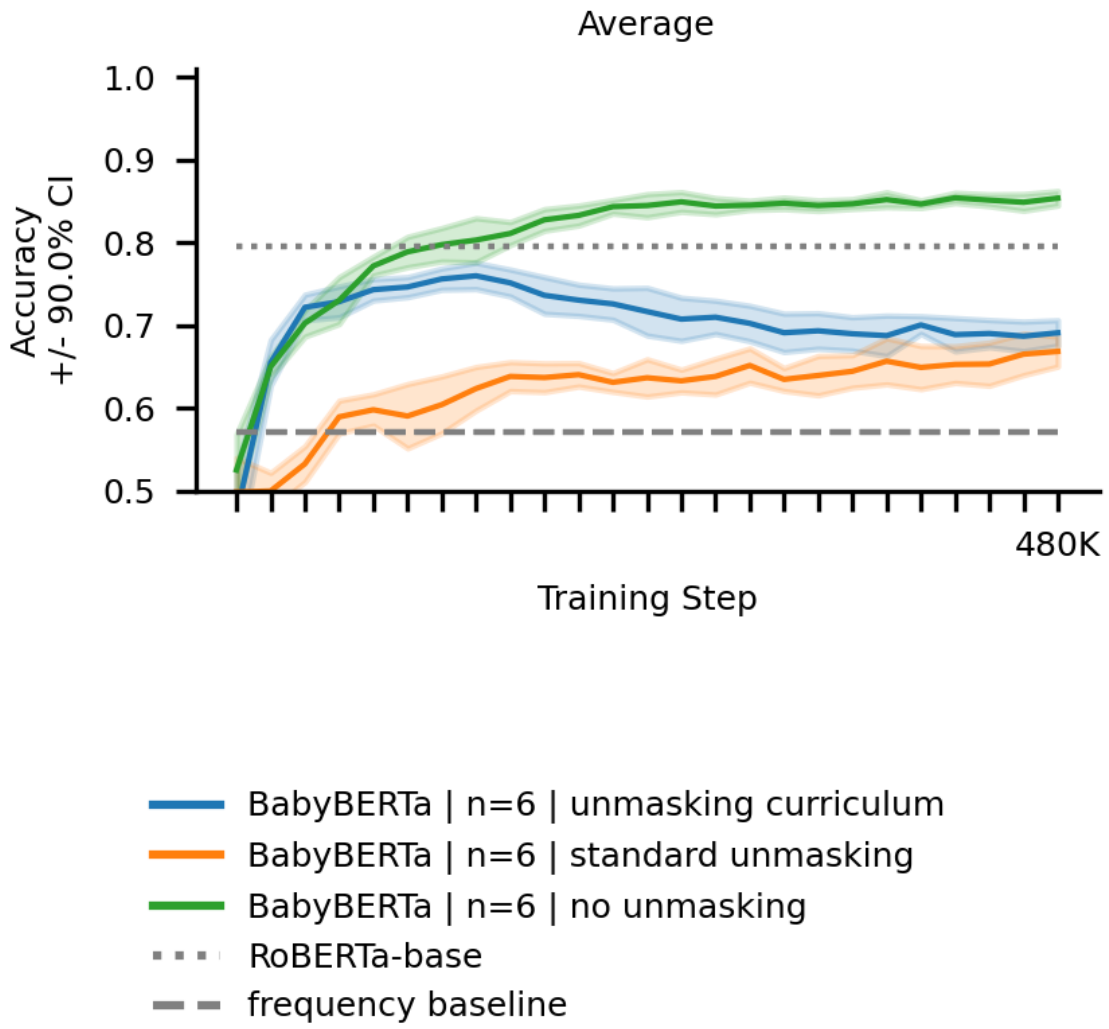

Figure 5: Average accuracy on grammar test suite across time. Comparison between 3 different unmasking strategies.

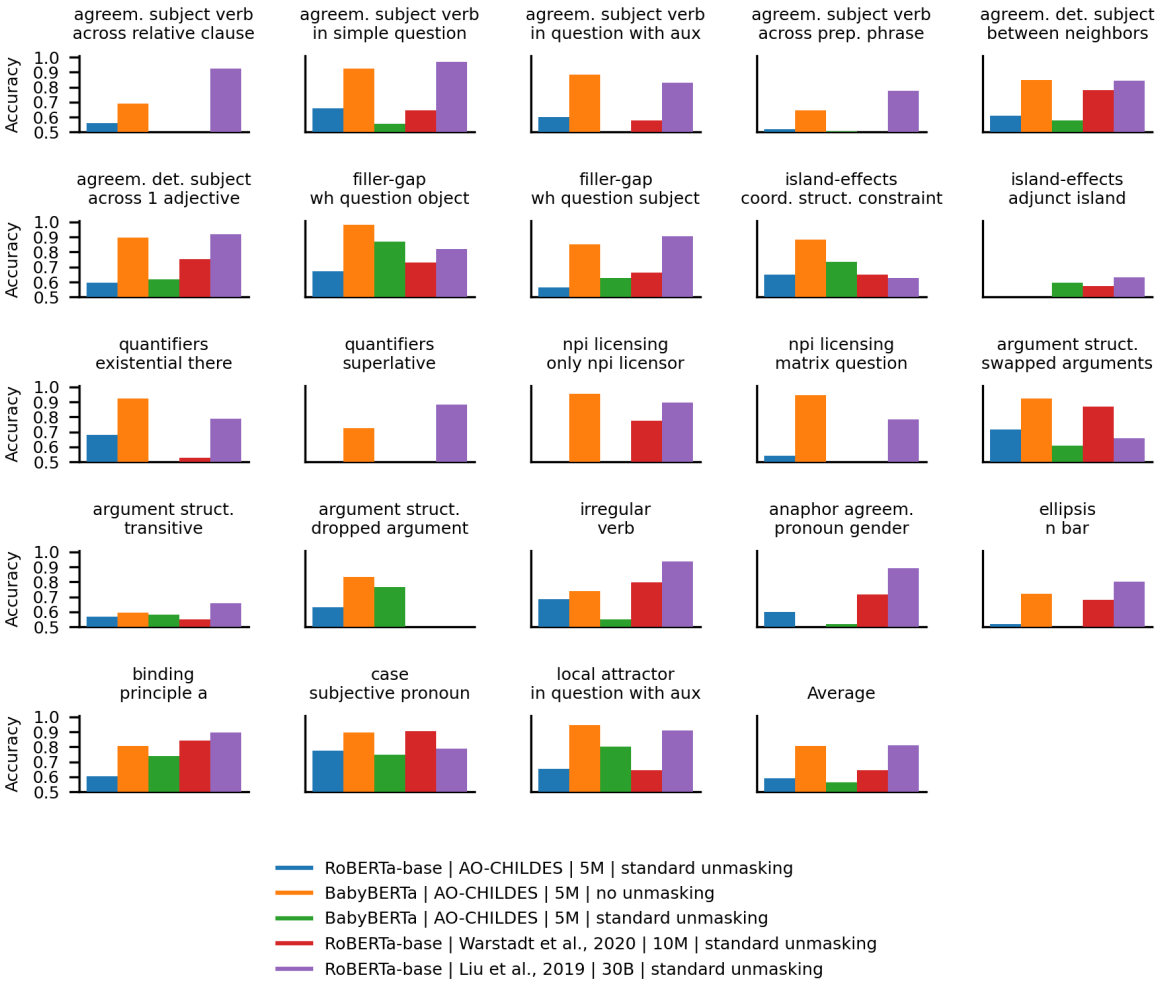

Figure 6: Accuracy on our grammar test suite separated by paradigm for RoBERTa-base pre-trained by Liu et al. (2019) on 30B words of web text (purple bar), RoBERTa-base pre-trained by Warstadt et al. (2020b) on 10M words of web text (red bar), BabyBERTa pre-trained by us on 5M words of AO-CHILDES with standard unmasking (green bar), BabyBERTa pre-trained by us on 5M words of AO-CHILDES without unmasking (orange bar), and RoBERTa-base pre-trained by us on 5M words of AO-CHILDES (blue bar). The word frequency baseline scores a sentence as grammatical if the sum of its word frequencies is greater than its counterpart sentence.

Figure 7: Accuracy on our grammar test suite separated by paradigm for BabyBERTa trained on the concatenation of AO-CHILDES and AO-Newsela in that order (blue line) or in reverse order (orange line). The word frequency baseline scores a sentence as grammatical if the sum of its word frequencies is greater than its counterpart sentence.

Figure 8: Accuracy on our grammar test suite separated by paradigm for BabyBERTa trained on the concatenation of Wikipedia-3 and AO-Newsela in that order (blue line) or in reverse order (orange line). The word frequency baseline scores a sentence as grammatical if the sum of its word frequencies is greater than its counterpart sentence.

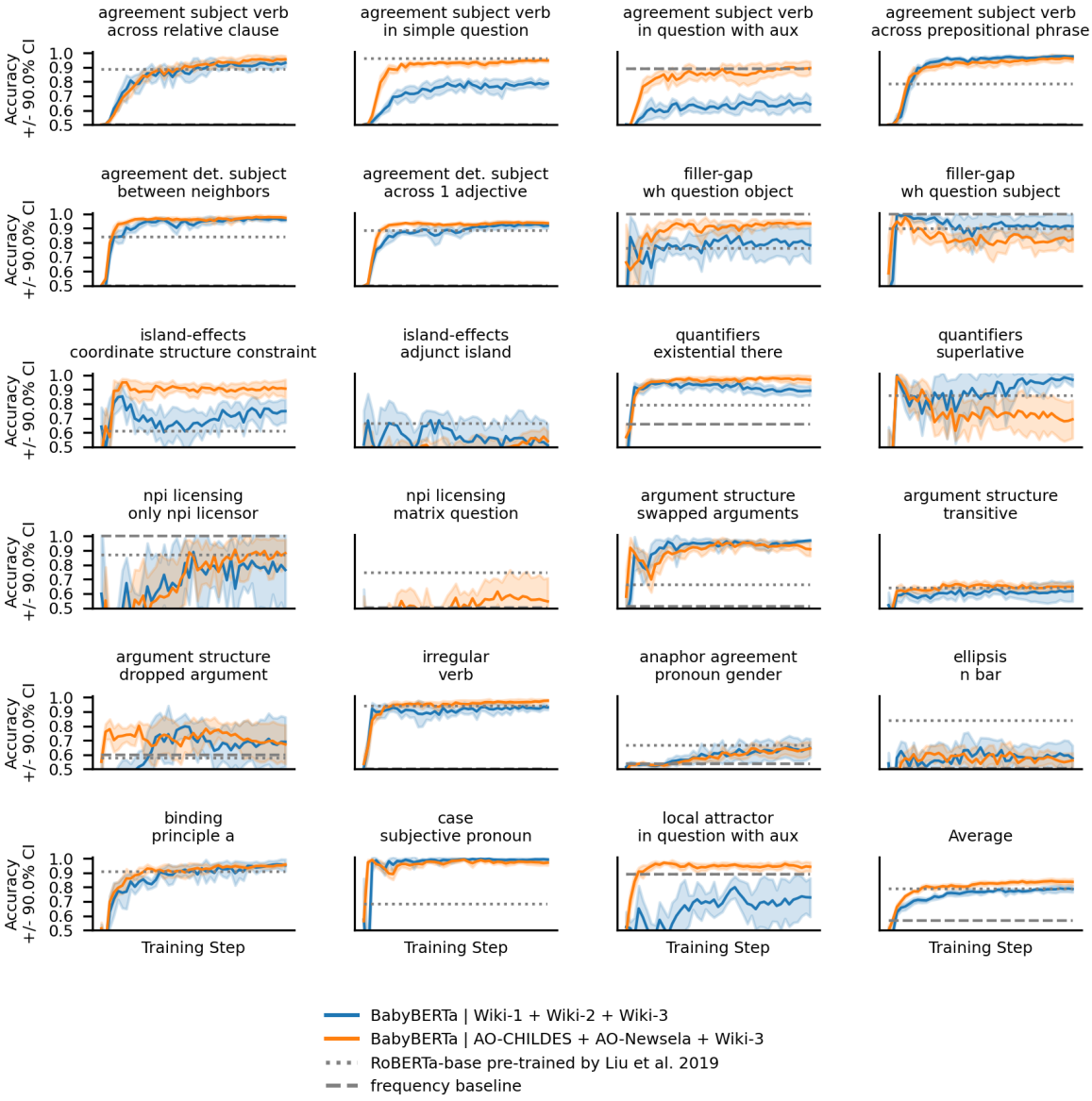

Figure 9: Accuracy on our grammar test suite separated by paradigm for BabyBERTa trained on a diverse set of corpora (AO-CHILDES + AO-Newsela + Wikipedia-1, orange line), and BabyBERTa trained on a size-matched corpus of Wikipedia articles (blue line). The word frequency baseline scores a sentence as grammatical if the sum of its word frequencies is greater than its counterpart sentence.