On the effect of curriculum learning with developmental data for grammar acquisition

Mattia Opper [a] and **J. Morrison** [a,b] and **N. Siddharth** [a,c]
[a] University of Edinburgh; [b] University of St Andrews; [c] The Alan Turing Institute
{m.opper,j.morrison,n.siddharth}@ed.ac.uk

Abstract

This work explores the degree to which grammar acquisition is driven by language 'simplicity' and the source modality (speech vs. text) of data. Using BabyBERTa (Huebner et al., 2021) as a probe, we find that grammar acquisition is largely driven by exposure to speech data, and in particular through exposure to two of the BabyLM (Warstadt et al., 2023) training corpora: AO-Childes and Open Subtitles. We arrive at this finding by examining various ways of presenting input data to our model. First, we assess the impact of various sequence-level complexity based curricula. We then examine the impact of learning over 'blocks'—covering spans of text that are balanced for the number of tokens in each of the source corpora (rather than number of lines). Finally, we explore curricula that vary the degree to which the model is exposed to different corpora. In all cases, we find that over-exposure to AO-Childes and Open Subtitles significantly drives performance. We verify these findings through a comparable control dataset in which exposure to these corpora, and speech more generally, is limited by design. Our findings indicate that it is not the proportion of tokens occupied by high-utility data that aids acquisition, but rather the proportion of training steps assigned to such data. We hope this encourages future research into the use of more developmentally plausible linguistic data (which tends to be more scarce) to augment general purpose pre-training regimes.

1 Introduction

Pre-training modern LLMs has become an increasingly resource intensive process, often requiring hundreds of GPU hours, and enough electricity to power a small village. These requirements have led to model creation increasingly becoming restricted to the few actors that are able to muster the resources necessary, excluding many from being able to participate in researching the field.

On the other hand, recent work (Huebner et al., 2021; Mueller and Linzen, 2023) has shown that Transformer LLMs can acquire knowledge of grammar and syntax with less data scale than was previously thought necessary, provided that they are exposed to simpler forms of language. These findings provide a hope that research on pre-training can once again become accessible to the community as a whole.

However, even if scale may not be such a strict requirement for the acquisition of linguistic knowledge, there are two tendencies exhibited by transformer models that may still be barriers to accessibility. Firstly, simply increasing the number of training steps generally yields better results. In fact, recent work by Murty et al. (2023) has shown that continuing training long past *train loss saturation* can lead to acquisition of a bias towards tree-likeness. While a fascinating finding in its own right (as hierarchical structure is considered a central feature of natural language) many groups simply won't have the GPU hours necessary to reach this point, so resources may remain a barrier. Secondly, it is often the case that simply increasing the complexity of a model can be beneficial (e.g. greater depth can aid syntactic generalisation (Mueller and Linzen, 2023)), but increasing complexity also increases cost.

This work investigates whether we can use the starting small approach to curriculum learning (Elman, 1993) combined with a small scale developmentally plausible pre-training set to aid model grammar acquisition without necessitating an increased budget of training steps. Our findings are mixed. We were unable to significantly outperform a random sampling baseline over all the pre-training corpora contained in the strict-small track. However, we are able to attribute this to the prevalence of high-utility simple speech data. We demonstrate through the use of a control corpus that in a setting where this high-utility data is more scarce, the benefits of developmentally ordered learning start to show themselves.

Proceedings of the 27th Conference on Computational Natural Language Learning: Volume 2: The BabyLM Challenge, pages 346–355

2 Related Work

Elman (1993)'s seminal early work presented the idea of starting small, whereby a model is first exposed to simpler data before moving on to more complex types of input. The idea is that complex data might get the model to learn 'false friend' heuristics that are actually harmful in the long run, but simple data might get it to learn in a way that generalises well. However, this hypothesis is not without controversy. Rohde and Plaut (1999) found that networks trained on complex sentences from the start performed better than those trained on simpler sentences initially, contradicting the starting-small hypothesis. They argue that previous studies supporting the starting small hypothesis may have terminated the training of complex networks too early. Bengio et al. (2009) train a language model using a curriculum learning strategy where only spans of text containing the first 5k most frequent words are included, then expanding to the first 10k and so on. They find that while a random sampling baseline initially achieves a superior loss, with sufficient updates the curriculum strategy comes to a better minimum and converges more stably.

These approaches have in common that they gradually reveal more and more of the dataset. An alternative approach is a single-phase curriculum where the input data is sorted by some criterion and then presented to the model in a fixed ordering. The model goes through the curriculum once, and does not revisit simpler data once it transitions to more complex data. The success of the single phase approach depends heavily on how complexity is defined, and has shown dubious results when applied to NLP (Campos, 2021; Surkov et al., 2022). Even under a developmentally plausible setting, the efficacy of the single phase approach has been shown to be mixed (Huebner et al., 2021).

3 BabyBERTa

3.1 Model and Training Details

The baseline model architecture we use in this work is an adaptation of BabyBERTa (Huebner et al., 2021). BabyBERTa is a variant of RoBERTa (Liu et al., 2019), with a few key differences:

No Unmasking: RoBERTa had used unmasking to minimise the disparity between pre-training and fine-tuning (where no mask tokens are used). Instead, BabyBERTa prioritises the finding that removing unmasking substantially improves model grammar acquisition.

No length truncation: Sequences which exceed the max length set in BabyBERTa are excluded instead of truncated. This ensures the model is only provided with whole utterances that correspond to a coherent linguistic unit.

Smaller Size: BabyBERTa is both shallower (fewer layers) and narrower (lower hidden and feed-forward size) than the original RoBERTa.

Training Data and Vocab Size: BabyBERTa is pre-trained on child directed speech and uses a substantially smaller vocabulary size in order to mimic that of a 6-year-old (theorised to be roughly 6k words).

We adopt this architecture for use in our paper with some alterations:

Increased Vocabulary: The BabyLM training corpora consist of more diverse data than AO-Childes, and encompass a wider range of developmental complexity. Consequently, a greater vocabulary size may be beneficial. We performed a grid search over vocabulary sizes 10k, 20k, 30k, 40k and 50k and found 30k to be optimal.

Increased Width: We double the hidden size and feed-forward network dimension of the original BabyBERTa from 256 to 512 and 1024 to 2048 respectively. These changes yielded slight improvements in BLiMP performance, but without them the model performed substantially worse on NLI tasks than the RoBERTa baseline provided for the challenge. However, increased width yields only minimal improvements in terms of grammar acquisition. We tested increasing the depth of the model (more layers), but found this yielded no improvements within the pre-training step budget we had available, neither did increasing the number of attention heads.

Our remaining model parameters are the defaults for RoBERTa from the transformer's library (Wolf et al., 2019). We use relative key query positional embeddings and set our max sequence length during training to 128 for efficiency reasons, and follow the no-truncation strategy. We set the learning rate to 1e-5 and the max number of steps to 120k using batch size 128. Unless stated otherwise, all our experiments utilise these same hyperparameters. We utilise dynamic masking as with the original RoBERTa, and no unmasking follow-

ing BabyBERTa in all cases without exception. While the latter choice may impact downstream performance in the fine-tuning tasks, the focus of this paper is largely on grammar acquisition as measured by the zero-shot evaluation suite and here removing unmasking proved beneficial.

4 Sequence Complexity Curricula

Our first point of investigation was to examine whether we could use sequence complexity based curricula to improve grammar acquisition. In the original BabyBERTa paper, the authors found that training on AO-Childes in its original ordering (which corresponds to age ordering, hence AO) led to better grammar acquisition than the reverse, but failed to outperform a random sampling baseline. They attribute this failure to a lack of vocabulary diversity in each batch when using age ordering. By contrast, the BabyLM pre-training corpora exhibit varying complexities (AO-Childes or Open Subtitles are on average much simpler than Wikipedia, see Figure 2), as well as variance in complexity within the corpora. Consequently, we hypothesised that we may be able to scaffold learning by presenting sequences to the model in order of complexity, while mitigating the potential issue of vocabulary and domain diversity by drawing these sequences from across all the source corpora.

4.1 Curriculum Types

We tested three kinds of curricula using different measures for complexity. As we were submitting to the strict small track, we only used sequence complexity metrics that could be easily inferred from the raw data. We call lines of the corpora 'sequences' for lack of a better term. Each corresponds to a linguistically coherent unit, but they can vary from short transcribed utterances to full articles. It is likely that better curricula can be created by using more complex and linguistically motivated metrics, but without the use of external resources this is difficult to achieve. The three types we tried are:

Entropy: Entropy favours highly likely sequences, but penalises based on length. This should order data such that the most likely shortest sequences appear first, allowing the model to learn simple local dependencies before moving to more complex data.

Unigram Probability: Orders sequences by the average unigram probability of their tokens. This is similar to entropy, except without penalising length directly. The idea here is that the model can learn good representations for highly likely tokens first and use that to inform its decision around more complicated/rarer tokens later down the line. The approach is similar to that of Bengio et al. (2009).

Block: Introduced by Nagatsuka et al. (2021) in the block curriculum, block size is increased during the course of training. This allows the model to first learn to optimise local dependencies before moving to longer range ones. The block curriculum differs from the other two in that each stage of learning does not present a subset of sequences, but rather is over the entirety of data in all the corpora, with each stage providing a greater context window for the model to consider. Secondly, by utilising blocks, each input consists of a span of tokens rather than a linguistically coherent unit like a transcribed utterance or article, and can include segments that represent partial units both at the start or end of a block. This means that the model must learn to identify the boundaries between coherent units during training, which may be a burden.

4.2 Creation

We first tokenised all sequences using the model's tokeniser, then calculated probabilities for each token using MLE, and scored each sequence, and subsequently re-ranked the data. The re-ranked sequence were then divided into different stages, by chunking according to rank. We used 4 stages for all curricula, with each stage containing a roughly equal number of sequences. Increasing this number did not yield significant improvements.

In the original block curriculum Nagatsuka et al. (2021) use block sizes 64, 128, 256 and 512, with the maximum batch size that could fit on their GPU at each step. We adopt this approach, but following initial findings that significantly smaller block sizes proved more beneficial than larger ones (potentially as a result of us limiting the max number of steps to 120k to enforce consistency across experiments), we instead switched to block sizes 16, 32, 64, 128.

In some preliminary training runs, we tested both the single phase and starting small approaches to curriculum learning. The single phase approach proved significantly inferior and exhibited a tendency towards catastrophic forgetting. Instead, we

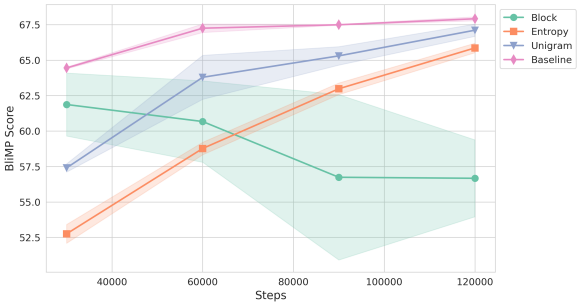

Figure 1: Zero-shot performance for curricula vs. random-sampling baseline with training (over 3 seeds).

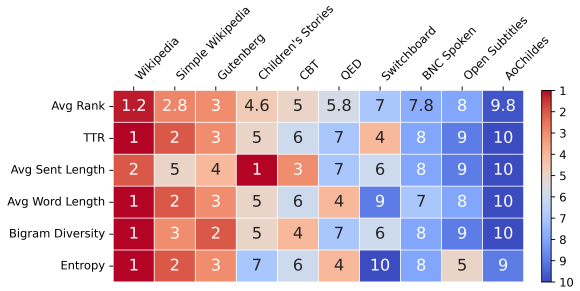

Figure 2: Heatmap ranking of the BabyLM Strict Small training corpora according to complexity measures.

used the following strategy: Each stage introduces new data for training, and the model is trained on the data in the current stage concatenated with that of all stages seen prior. This approach worked best for us. Each stage was trained on for 30k steps, totalling a combined 120k. As a baseline, we trained using random sampling over the whole data, also for 120k steps.

4.3 Summary

Figure 1 shows results. None of the curricula were able to outperform a baseline measure of simply sampling random sequences from the concatenation of all the datasets. Though the sequence complexity based curricula showed improvement throughout training, the block curriculum got worse with each stage. This raised two follow-up questions for us. First, what causes the random sampling baseline to do so well? Second, is using blocks as inputs rather than sequences causing the block curriculum to fail, or some other factor [1]?

5 Investigating Random Sampling

Why might random sampling be successful? Let us begin by examining how we present our data.

[1]The large variance exhibited by the block curriculum suggests significantly more steps would be needed to perform well.

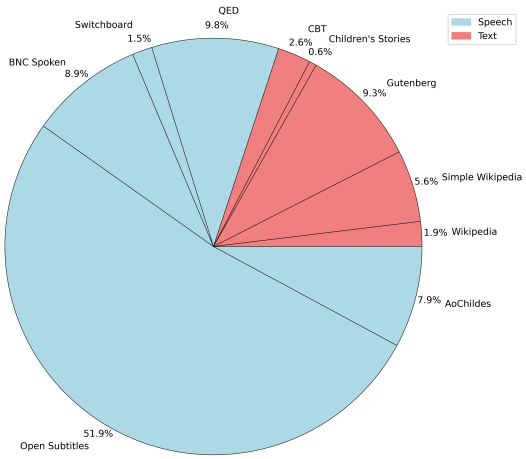

Figure 3: Distribution of line counts across the ten language corpora, with each line treated as a unique sequence. The percentages represent the proportion of total lines that each individual corpus contributes to the overall dataset.

In terms of number of tokens, the BabyLM pretraining corpora are roughly equally divided between the source modalities: text and (transcribed) speech. Though there is a slight weighting in favour of speech, which comprises 56% of total tokens. Now let us contrast this with the relative complexity of each corpus (see Figure 2). We can see that the speech corpora on average, across all metrics, contain far simpler language than the text corpora. Secondly, as we were submitting to the strict small track we do not perform any augmentation on the data, including sentence tokenisation. This means that the random sampling baseline takes as input lines from each corpus. If we examine the distribution of number of lines between corpora, we find a very different division compared with the number of tokens. Figure 3 shows the breakdown. Looking at the number of lines, the balance between transcribed speech and text data becomes highly unequal, with transcribed speech now comprising a total of 80% of all examples. Secondly, the two corpora which contain on average the simplest language (AO-Childes and Open-Subtitles) represent 59.8% of all lines, and these may be responsible for driving the majority of grammar acquisition. If this is the case, then it may explain the performance of the random sampling baseline, as it is more likely to see sequences from these two corpora than any others, while still being provided a degree of diverse examples in each batch. By contrast, when the input is treated as blocks rather

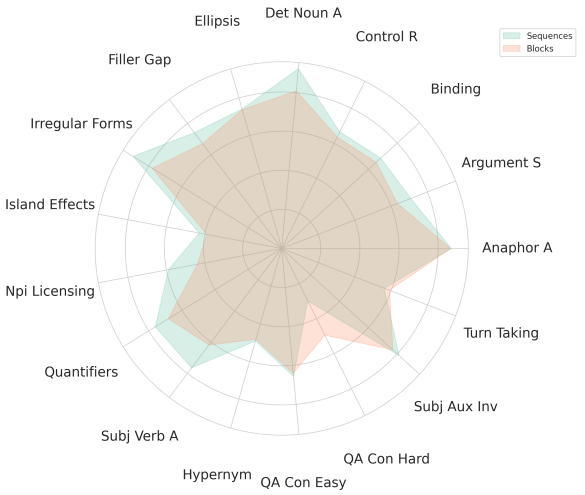

Figure 4: By-task breakdown of zero-shot performance when input data is either a linguistically coherent sequence or a block. Results averaged over 3 seeds.

than lines, the balance between speech and text inputs corresponds to the proportion of number of tokens. Alternatively, it may simply be that training on blocks requires more steps so that the model can identify linguistically coherent units.

To test this hypothesis, we train on both models, taking either blocks or lines from the corpora (henceforth referred to as sequences) as input. We train for an equal number of steps (120k). We report results for block size 32, as when trained for the full number of steps, this worked best out of all the variations tested in the block curriculum.

5.1 Summary

Even when trained for a greater number of steps we find that sequences as input still quite substantially outperform blocks. Results are shown in Figure 4 and Table 1. The only exception is on the held out tasks, however, this is due to the block variant of the model essentially having random accuracy on the QA congruence tasks (close to 50%) while the sequences variants appear to have learned to solve the easy tasks, but fail at the hard ones (see Table 7 for full results by for each task).

We can conclude from this that providing linguistically coherent units as input is beneficial for overall efficient grammar acquisition, despite the fact that the model is disproportionately being exposed to speech data, and therefore only a subset of the overall tokens throughout pre-training. However, we still need to disentangle whether it is speech that is driving this effect or the fact that the model is being presented linguistically coherent units.

Table 1: By-task breakdown of zero-shot performance between models utilising random sampling strategies where inputs are either linguistically coherent sequences or blocks. Results averaged over 3 seeds.

Tasks	Blocks	Sequences
Original	65.98 ± 1.02	**73.11 ± 0.89**
Held Out	**59.59 ± 0.6**	56.45 ± 0.88
Overall	64.1 ± 0.2	**68.21 ± 0.23**

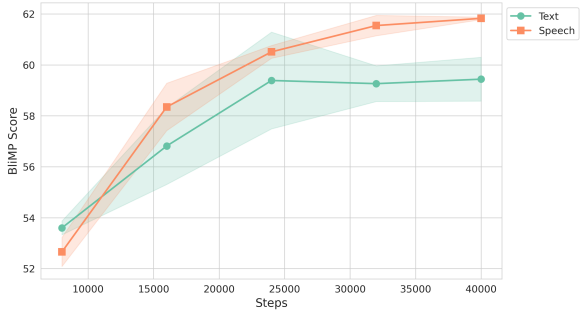

Figure 5: Zero-shot performance by step when the model is trained on either the transcribed speech or text portions of the pre-training corpora (over 3 seeds).

6 Speech vs Text

6.1 Efficient Acquisition by Modality

Prior work examining the impact of pre-training on AO-Childes (Huebner et al., 2021; Mueller and Linzen, 2023) has shown that utilising this simpler form of language enables more efficient acquisition of grammatical knowledge and encourages a bias towards hierarchical generalisation in transformer language models. As such, it is not improbable that simply over exposing the model to simpler data such as speech may be driving performance. To test this, we perform two ablations. First, we assess the impact of training on only one source modality for an equal, but reduced, number of steps to assess whether one provides a better starting point for acquisition. This instance actually in some respects favours the textual data, which contains longer sequences and therefore should provide more signal per step, as each input will contain more masks and contexts while still representing a linguistically coherent unit. Figure 5 shows results on the first comparing the two modalities when trained for 40k steps each. Training on transcribed speech consistently outperforms training on text alone, and leads to more stable improvements than just text. Indicating that it is a better starting point.

Table 2: Comparison of Ordering Effects Given Source Modality. Results averaged across 3 seeds.

Training Data	Original	Held Out	Overall
Speech → Speech+Text	**72.99 ± 0.53**	**56.26 ± 1.3**	**67.74 ± 0.77**
Text → Speech+Text	71.69 ± 0.6	**53.77 ± 2.78**	66.42 ± 1.2
Speech+Text	**73.11 ± 0.89**	**56.45 ± 0.88**	**68.21 ± 0.52**

6.2 Speech Data as a Foundation

As a second follow-up investigation, we once again trained on two different settings. In the first we train on speech first and then the concatenation of text and speech for 60k steps respectively. This is to check whether we can build a foundation from speech data alone, and then transition to including both modalities. However, here text data only occupies 10% of the overall proportion of inputs, and is only observed in the later stages of training. As a control, we also try the inverse, starting with text first and then transitioning to the concatenation of all the corpora, this means that the text data now provides 60% of all the total inputs and speech is only introduced once the model later in training, no longer acting as a foundation. Results are in Table 2. Further, weighting things towards speech improves over the text control on the original BLiMP tasks, and secondly makes performance indistinguishable from random sampling if we account for standard deviation overlap. The only area where this does not hold is in the held out tasks.

6.3 Summary

We find that transcribed speech leads to improved BLiMP performance and lower variance compared with text only data. Based on this finding, we investigated whether we could design a simple two stage curriculum where we first train the model on speech only and then transfer to the full dataset. Under this setting, performance is roughly equal to random sampling, and shows some very slight improvements compared to the reverse curriculum. This is despite the fact that the model is only exposed to the $\approx 50\%$ of total tokens contained in the text portion in the latter half of training.

7 Corpora Complexity Curricula

Having found that speech data can provide a better foundation than text, and that over exposure may be behind the random sampling baselines performance, we conduct a follow-up investigation. How much exposure to more complex data is necessary in order to achieve grammar acquisition? To probe this question, we use the same strategy

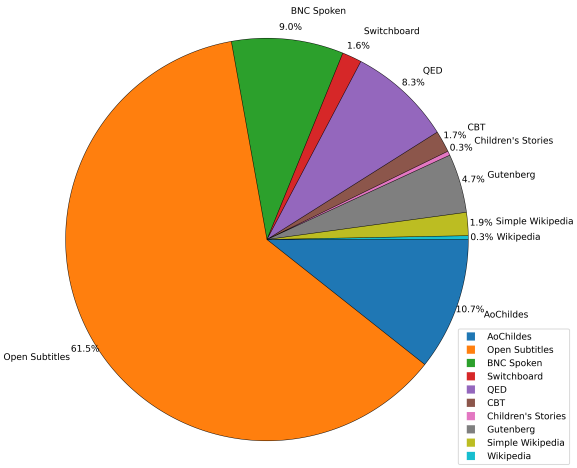

Figure 6: Proportion of total inputs comprised by each of the corpora using the corpus complexity curriculum.

for our curriculum by training on a stage and the concatenation of all previous stages. This time we define our ordering using the average rank across our various corpus complexity measures as shown in Figure 2. So our ordering starts with AO-Childes and ends with Wikipedia. The curriculum is simply the corpus complexity ordering, with two caveats. We treat BNC spoken and switchboard as one corpus, as switchboard is too small to warrant a new stage. We also do the same for CBT and children's stories, as they are very similar in terms of complexity. Using this form of curriculum further increases the model's exposure to simple data, with AO-Childes and Open Subtitles now representing 72.2% of all total training examples, compared with 59.8% before, and Wikipedia representing only 0.3% (see Figure 6). We again implement the reverse curriculum as a control measure, starting with Wikipedia and finishing with AO-Childes, and compare results to the random sampling baseline (see Table 3). The simple to complex curriculum yields marginally better results overall compared to the random sampling baseline, and the gap with the reverse curriculum is wider here than for the previous speech versus text curriculum.

However, the marginality of the increase compared to the random sampling baseline makes it

Table 3: Comparison of performance by corpora complexity ordering. Results averaged across 3 seeds.

Training Data	Original	Held Out	Overall
Simple → Complex	**74.14 ± 0.39**	**55.9 ± 0.74**	**68.77 ± 0.04**
Complex → Simple	71.89 ± 0.9	**54.29 ± 2.74**	66.72 ± 0.42
Random Sampling	**73.11 ± 0.89**	**56.45 ± 0.88**	**68.21 ± 0.52**

Table 4: Control Dataset Statistics

Name	Tokens %	Input %	Curriculum Input %
AO-Childes	4%	15.8%	26%
CBT	50%	51.8%	56%
Wikipedia	46%	32.4%	18%

difficult to make any strong claims regarding the effect of ordering. We wondered if this was because the BabyLM training data is already favourable for grammar acquisition and weighted towards speech, and whether we would observe greater benefits over random sampling in a setting where the data did not have these properties.

7.1 Summary

We wanted to test whether we could design a curriculum based on the complexity of the various pre-training corpora (see Figure 2). We find that following this approach led to improvements over the reverse, especially on the original set of BLiMP tasks, but failed to show a significant difference over random sampling. We hypothesise that this due to AO-Childes and Open Subtitles, two of the most high utility corpora for grammar acquisition, already making up a large percentage of inputs in the random-sampling setting. Thus, the introduction of a curriculum may have little impact.

8 Control Dataset

To test whether complexity ordering helped more when the training data was less optimal, we created a new dataset. It consists of the AO-Childes portion of BabyLM 10M, and the CBT and Wikipedia portions of BabyLM 100M, representing the simplest, middle, and most complex corpora respectively. We set max sequence length to 512 to allow training on as much of the data as possible. Combined, these three corpora have approximately 10 million tokens (similar to the 'strict-small' track), but with the vast majority of these coming from text data. It also means that the number of inputs that come from simpler, more beneficial data is reduced. Descriptive statistics can be found in Table 4.

We train a new tokeniser on the data, and then compare results between a random sampling baseline and corpus complexity curriculum approach described in the previous section. Both versions are trained for 120k steps, but we had to lower the batch size to 64 due to GPU memory constraints with longer sequences. Results are in Table 5, and a plot of the by task scores can be found in Figure 8. Under this setting, the curriculum approach begins to demonstrate modest, but visible improvements over random sampling, though this does not extend to the held out tasks. Figure 7 shows the performances patterns as the number of steps increases. The curriculum consistently offers slight improvements over random sampling.

Table 5: Control Dataset Results on Zero-shot Tasks. Results averaged across 3 seeds.

Training Data	Original	Held Out	Overall
Simple → Complex	**72.18 ± 0.88**	**55.52 ± 1.08**	**67.28 ± 0.52**
Random Sampling	70.77 ± 0.37	**55.88 ± 1.11**	66.38 ± 0.1

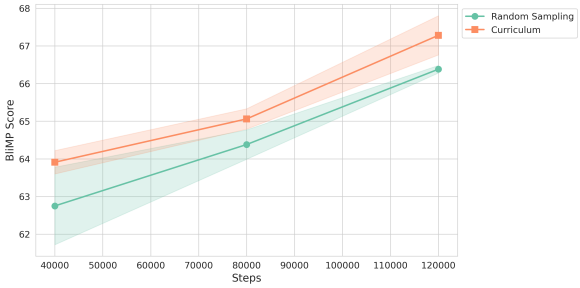

Figure 7: Zero-shot performance by step when the model is trained using the curriculum or random sampling on our control dataset (over 3 seeds).

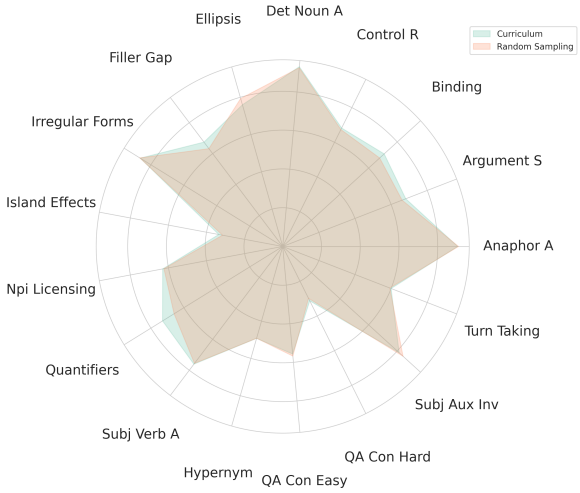

Figure 8: By-task breakdown of zero-shot performance on the control dataset curriculum vs random sampling. Results averaged across 3 seeds.

9 Summary

We wanted to test whether curriculum learning can be beneficial in a scenario where the majority of data is not high utility, i.e., simple transcribed speech. To do so, we created a control corpus where the majority of data comes from long form text. Under this setting, we find a slight, but discernable improvement from using the curriculum.

10 Conclusion

We began our exploration by attempting to design a learning curriculum to further grammar acquisition for the BabyLM strict-small track. We found that when the majority of the data is high-utility, as is the case here, curriculum learning shows no substantial benefits. However, such training data is not always available or may be dwarfed by the number of tokens of low utility data available. In these settings—common for pre-training NLP models—our results indicate some promise in starting small after all. However, extensive further experimentation, most likely requiring larger scale corpora, is necessary to properly test and verify this claim.

11 Limitations

The work presented in this paper represents an initial foray into starting-small-style learning. There are a number of extensions and further questions one could ask, building upon the work presented here, that could help shine further light on the nuances of this style of learning.

- Although the control-dataset experiments in Section 8 show better performance when starting small compared to random sampling, we don't yet definitively discount that starting large *in the same setting* does not achieve the same results. This could be remedied by constructing a careful 'complementary' large-to-small-complexity curriculum.
- Given our training regime, for both random sampling and corpus curricula, on both the original data and the control, we don't know if the eventual trends over training will resemble that reported by Rohde and Plaut (1999) or that of Bengio et al. (2009). We could explore this by attempting to train over longer horizons to see if a comparable trend emerges.
- In our submission for the competition, we used an additional technique: layer stacking (Gong et al., 2019), which involved progressively growing the model as we advanced through the curriculum (following Elman (1993)). The hypothesis was that we would be starting small in two ways: from simple data and/or a simple model. This yielded some slight improvements over only using the corpora curriculum over the entirety of the strict-small training data, which had been our previous best scoring model. We do not yet have a complete picture of how layer-stacking affects all the various training regimes discussed in this manuscript, and hence only describe the basic algorithm in the appendix A.
- Follow on work could probe how much of a token disparity can be tolerated before loseing the benefits of starting small from transcribed speech. This could be, for example, replacing CBT with a larger proportion of Wikipedia; e.g. Wiki-103 (Merity et al., 2016).

12 Full Results

While our focus here has been grammar acquisition, we present results on all tasks in Table 6. We perform favourably compared to the official RoBERTa baseline for the challenge, but one area shows a notable disparity—MSGS tasks (Warstadt et al., 2020) measuring syntactic category. This may be because our model is too shallow (RoBERTa base has 12 layers vs. our 8).

Table 6: Full results from Dynabench for our submission vs. the official RoBERTa baseline for the challenge.

Task	Ours	RoBERTa Base
Anaphor Agreement	**84**	82
Argument Struct	**70**	67
Binding	**69**	67
Control R	**70**	68
DN Agreement	**92**	91
Ellipsis	**77**	76
Filler Gap	**76**	64
Irregular Forms	87	87
Island Effects	**42**	40
NPI Licensing	**65**	56
Quantifiers	**78**	71
SV Agreement	**77**	66
Hypernym	45	**49**
QA Cong Easy	**69**	31
QA Cong Hard	**33**	32
SA Inversion	**77**	72
Turn Taking	**57**	53
CoLA	**32**	26
SST-2	87	87
MRPC	79	79
QQP	**82**	74
MNLI	73	73
MNLI-MM	74	74
QNLI	**78**	77
RTE	49	**62**
BoolQ	62	**66**
MultiRC	60	**61**
WSC	61	61
CR	**0.73**	0.43
LC	1.0	1.0
MV	**1.0**	0.98
RP	0.84	**0.94**
SC	0.16	**0.86**
CR_LC	-0.58	**-0.28**
CR_RTP	-0.92	**-0.77**
MV_LC	-1.0	**-0.99**
MV_RTP	**-0.26**	-0.79
SC_LC	-0.43	**0.16**
SC_RP	-0.59	-0.45

Acknowledgements

MO was funded by a PhD studentship through Huawei-Edinburgh Research Lab Project 10410153. We also wish to thank Victor Prokhorov for his suggestions and tireless willingness to answer questions.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Daniel Campos. 2021. Curriculum learning for language modeling. *CoRR*, abs/2108.02170.

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

Linyuan Gong, Di He, Zhuohan Li, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Efficient training of BERT by progressively stacking. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2337–2346. PMLR.

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.

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. *CoRR*, abs/1907.11692.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *CoRR*, abs/1609.07843.

Aaron Mueller and Tal Linzen. 2023. How to plant trees in language models: Data and architectural effects on the emergence of syntactic inductive biases. *ArXiv*, abs/2305.19905.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher D. Manning. 2023. Grokking of hierarchical structure in vanilla transformers. *ArXiv*, abs/2305.18741.

Koichi Nagatsuka, Clifford Broni-Bediako, and Masayasu Atsumi. 2021. Pre-training a BERT with curriculum learning by increasing block-size of input text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 989–996, Held Online. INCOMA Ltd.

Douglas L. T. Rohde and David C. Plaut. 1999. Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition*, 72:67–109.

Maxim Surkov, Vladislav Mosin, and Ivan Yamshchikov. 2022. Do data-based curricula work? In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 119–128, Dublin, Ireland. Association for Computational Linguistics.

Alex Warstadt, Leshem Choshen, Ryan Cotterell, Tal Linzen, Aaron Mueller, Ethan Wilcox, Williams Adina, and Chengxu Zhuang. 2023. Findings of the BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020. 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, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Appendices

A Layer Stacking

We grew our model during training by adding a layer when we reached each new stage of our curriculum. We cloned the existing uppermost layer at the beginning of each new stage of our curriculum, then stacked that layer on top of the existing layers of our model. Our model then proceeds to learn from our new mix of datasets for the new stage of the curriculum, with the uppermost layer most responsive to the newly revealed datasets in our curriculum. In this way we progressed from 1 to 8 layers over the course of our training regime.

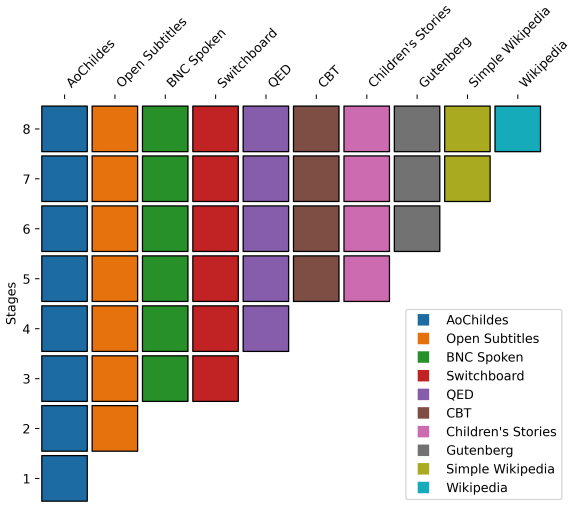

Figure 9: Our learning curriculum exposes our model to additional datasets stage-by-stage as it progresses through our training regime.

B Full Results Table

Table 7: Sequence vs block input performance on zero-shot tasks. Results are averaged across three random seeds. 4

Model	Anaphor A	Argument S	Binding	Control R	Det Noun A	Ellipsis	Filler Gap	Irregular Forms	Island Effects
Sequences	**86.78**	**70.84**	**68.58**	**66.60**	**92.31**	**74.12**	**74.20**	**89.47**	**42.91**
Blocks	86.39	63.59	65.36	63.66	80.91	73.81	67.19	77.89	39.08

Model	Npi Licensing	Quantifiers	Subj Verb A	Hypernym	QA Con Easy	QA Con hard	Subj Aux Inv	Turn Taking	**Score**
Sequences	**58.97**	**76.06**	**76.52**	**49.19**	**65.63**	29.90	**81.21**	56.31	**68.21**
Blocks	43.85	68.20	61.81	48.22	64.06	**49.50**	76.99	**59.17**	64.01