Beyond Shallow Heuristics: Leveraging Human Intuition for Curriculum Learning

Vanessa Toborek[1,2], Sebastian Müller[1,2], Tim Selbach[1], Tamás Horváth[1,2,3], Christian Bauckhage[1,2,3]
[1]University of Bonn, [2]Lamarr Institute, [3]Fraunhofer IAIS
toborek@cs.uni-bonn.de

Abstract

Curriculum learning (CL) aims to improve training by presenting data from "easy" to "hard", yet defining and measuring linguistic difficulty remains an open challenge. We investigate whether human-curated simple language can serve as an effective signal for CL. Using the article-level labels from the Simple Wikipedia corpus, we compare label-based curricula to competence-based strategies relying on shallow heuristics. Our experiments with a BERT-tiny model show that adding simple data alone yields no clear benefit. However, structuring it via a curriculum – especially when introduced first – consistently improves perplexity, particularly on simple language. In contrast, competence-based curricula lead to no consistent gains over random ordering, probably because they fail to effectively separate the two classes. Our results suggest that human intuition about linguistic difficulty can guide CL for language model pre-training.

1 Introduction

The growing scale of language models (LMs) has increased interest in training strategies that improve efficiency and convergence. Curriculum learning (CL), inspired by developmental psychology, is one such approach. CL structures training by presenting examples in a sensible order – typically from "easy" to "hard" (Elman, 1993; Bengio et al., 2009; Wang et al., 2021). While intuitively compelling and empirically useful in certain NLP tasks (Platanios et al., 2019; Nagatsuka et al., 2021), its overall impact on masked language model (MLM) pre-training remains debated (Surkov et al., 2022).

A key challenge in CL is the definition of linguistic difficulty. Unlike other domains, language difficulty may arise from multiple dimensions – such as syntax, semantics or context. In the absence of gold standards, prior work often relies on shallow heuristics (Platanios et al., 2019; Ranaldi et al., 2023). Yet, readability research suggests that no single heuristic reliably captures linguistic complexity (Battisti et al., 2020). In contrast, humans intuitively consider multiple dimensions when simplifying text. This motivates the central question for this work: *Can human-curated simple language effectively guide CL for MLM pre-training?*

Rarity	Class	Example
low	SL	She is the author of the Twilight series.
low	EL	The history of poker is the subject of some debate.
high	SL	Today, most automotive diesels are turbocharged.
high	EL	Pink Floyd watched The Beatles recording Lovely Rita.

Table 1: Sentences showing examples of high and low average word rarity for each class in the Simple Wikipedia dataset (Kauchak, 2013).

To answer this question, we study CL strategies based on article-level labels from the Simple Wikipedia corpus (Coster and Kauchak, 2011) and compare them to *competence-based CL* with shallow difficulty heuristics (Platanios et al., 2019), using BERT-tiny for MLM pre-training. Our experiments show that merely adding simple language data to training yields no overall improvement. Still, incorporating it through a label-based curriculum consistently improves not only overall perplexity but particularly the simple language perplexity. This effect vanishes when reversed: training on everyday language first is detrimental to learning, underscoring the importance of example ordering. Surprisingly, competence-based curricula show no benefit over random ordering.

Further, we find that simple and everyday language articles have similar vocabulary sizes and high lexical and distributional overlap on the chosen difficulty heuristics. This suggests that competence-based CL fails here, because the heuristics do not effectively separate the classes. In contrast, the consistent gains from label-based curricula imply that simple language encodes other useful information, providing structure that benefits pre-training when leveraged correctly. These results suggest that simple language does indeed help, when applied in a curriculum that makes use of human intuition on linguistic difficulty.

2 Related Work

A common form of data-level CL orders the data points according to a global difficulty measure. This approach has been applied to various NLP tasks such as language modelling (Nagatsuka et al., 2021; Ranaldi et al., 2023), machine translation (Platanios et al., 2019; Mohiuddin et al., 2022), and questions answering (Liu et al., 2018) using difficulty measures like input length (Nagatsuka et al., 2021; Zaremba and Sutskever, 2015), word rarity (Platanios et al., 2019), or domain similarity (Mohiuddin et al., 2022). However, the choice of metric is often intuitive and its overall effectiveness remains debated, as the work by Surkov et al. (2022) found that competence-based CL for MLM offers little to no benefit.

A parallel line of work explores the benefits of simplified language in neural network training. Mueller and Linzen (2023) show that pre-training on simple language corpora strengthens the syntactic inductive bias in encoder-decoder models. Huebner et al. (2021) demonstrate that child-directed data facilitates grammar learning for down-sized encoder-only models. Lucas et al. (2024) explore CL through a masking-based strategy, also leveraging simplified language. While these studies focus on specific linguistic gains or efficiency improvements, the role of simplified language in global, data-level curriculum design remains unexplored. We address this gap by investigating whether editorially curated simple language – such as that in Simple Wikipedia – can serve as an effective learning signal for CL, and how it compares to commonly used difficulty heuristics.

3 Methodology

We use the following experimental setup to study the effect of simple language in MLM pre-training.

Label	# tokens	# sentences
Simple (SL)	$3,395,297$	$191,318$
Everyday (EL)	$3,796,654$	$176,019$

Table 2: Dataset statistics for simple (SL) and everyday (EL) language in the Simple Wikipedia corpus.

Dataset We employ the Simple Wikipedia dataset (Coster and Kauchak, 2011), the most popular, freely available simple language corpus in English. It consists of articles from the Simple English Wikipedia in simple language (SL) and their counterparts from the English Wikipedia in everyday language (EL). Each sentence inherits the article-level label (SL or EL), which may introduce some label noise due to within-article variation in sentence complexity. Table 2 compares both classes regarding their respective number of tokens and sentences.

Difficulty Heuristics For the competence-based CL, we consider three shallow heuristics for text difficulty: sentence length, word rarity, and the Flesch Reading Ease (FRE) score (cf. Platanios et al. (2019), Ranaldi et al. (2023)). Refer to Appendix B for the details. In addition to these, we include a random baseline, where difficulty scores are sampled uniformly to isolate the effect of data ordering from the progressive data exposure.

Curriculum Strategies We compare two CL paradigms. First, following Platanios et al. (2019), we implement the *competence-based* curriculum approach. We sort the training examples according to the aforementioned difficulty measures and gradually expand the training set as model competence increases. The curriculum proceeds until the entire dataset is included. We provide the full implementation details in Appendix A.

Second, we implement two *label-based* curricula using the SL/EL distinction. The sequential strategy first trains on SL until convergence, then continues training on EL. To mitigate potential forgetting from fully replacing the training data, we propose an incremental strategy: the model is first trained on SL alone, then continues on the combined SL+EL set, each phase until convergence. We also include a reverse sequential strategy (first on EL, then SL) as a control strategy.

Training Setup We train a BERT-tiny model with two transformer layers of hidden size 128,

Strategy	Perplexity	SL Perplexity	EL Perplexity	# Updates
Baseline EL	69.25 ±4.04	59.50 ±4.38	81.78 ±4.85	658 667 ±113 192
Baseline SL+EL	69.61 ±4.87	64.15 ±5.05	76.46♪ ±5.28	665 333 ±102 111
Incremental	66.36 ±2.53	63.29 ±3.39	**71.51**♪ **±2.55**	781 333 ±83 312
Sequential	**65.31**♪ **±4.19**	**57.83**♪ **±4.52**	74.39 ±4.91	781 333 ±122 292
Anti-Sequential	70.32 ±3.97	59.24 ±4.01	81.70♪ ±4.37	682 000 ±102 274
Length	69.05 ±4.15	63.84 ±4.12	76.37 ±4.46	672 667 ±71 760
Word Rarity	66.74 ±3.48	62.48 ±3.52	74.12 ±4.12	664 666 ±72 394
FRE	68.05 ±5.22	62.53 ±4.98	75.32 ±5.88	709 333 ±105 524
Random	68.07 ±4.92	63.08 ±4.95	75.21 ±5.40	679 333 ±105 388

Table 3: Performance of BERT-tiny across baseline and CL strategies. Perplexity is reported for the full dataset and separately for the simple (SL) and everyday language (EL) subsets. Sequential label-based curriculum achieves best overall and SL perplexity. No competence-based strategy shows consistent improvement over baselines. Reported values are mean and standard deviations across 15 runs. ♪ denotes significant changes.

two attention heads, an intermediate feed-forward of size 512, a batch size of eight, and a learning rate of 10^{-4}. All models are trained until convergence, with early stopping based on validation loss. All experiments are repeated over 15 random seeds to ensure statistical robustness.

Evaluation We evaluate model performance using overall perplexity as well as SL and EL subset perplexities. This helps us assess general improvements as well as register-specific gains. Our baselines include models trained with random sampling: one on everyday language only (Baseline EL), the other on a uniform mix (Baseline SL+EL).

4 Curriculum Learning Results

We summarise the final performance of the BERT-tiny model across all training strategies in Table 3, focusing on overall, SL, and EL perplexity, as loss values are less informative. We compare each strategy against a primary baseline (Baseline SL+EL), trained on SL+EL using random data sampling, with results averaged over 15 seeds. To assess the statistical significance of our results, we apply a one-sided Wilcoxon signed-rank test for symmetric distributions, and a one-sided median bootstrap test otherwise. All p-values are adjusted using the Holm-Bonferroni method within each experiment family (baseline, label-based CL, competence-based CL), using $\alpha = 0.05$ and directional hypotheses. Appendix C details the directional hypotheses and the corresponding adjusted p-values.

Does merely adding simple language to the training data improve model performance? The results provide a clear but mixed answer. Comparing Baseline SL+EL to Baseline EL, we see a significant improvement in EL perplexity but no improvement in neither overall nor SL perplexity.

Can simple language effectively guide CL? We find clear evidence in favour of simple language guiding CL – provided that the sampling strategy is right. Among the label-based CL strategies, only the sequential variant significantly improves overall as well as SL perplexity – achieving the best scores across all strategies. Incremental improves EL perplexity, but not overall performance. To show that the improvements of the sequential strategy are not accidental, we also test its anti strategy (i.e. starting training on EL, then progressing with SL): it performs similarly to Baseline EL and yields significantly worse EL perplexity than Baseline SL+EL. Both incremental and sequential strategies require more updates than Baseline SL+EL to reach these improvements.

Are shallow text features sufficient to guide competence-based CL? We have a negative answer to this question. Across all three competence-based difficulty measures, we observe no significant improvement in perplexity compared to Baseline SL+EL. The random strategy further suggests that neither simply increasing the dataset size nor imposing an order on shallow features leads to better model performance.

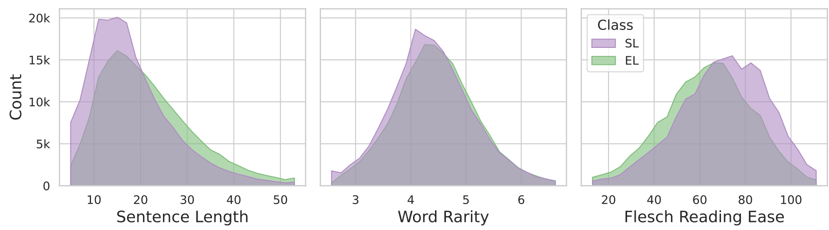

Figure 1: Distribution of sentence-level difficulty heuristics for SL and EL. None of the heuristics cleanly separates the two classes.

	SL	EL
SL	100%	96.67%
EL	86.06%	100%

Table 4: Vocabulary overlap between classes. Over 80% of EL's vocabulary is also present in SL, showing high lexical similarity.

5 Discussion

In this section we discuss the implications of the results from the previous section with regards to our three research questions.

Learning across registers: asymmetries and interference The surprisingly strong performance of Baseline EL on the SL subset suggests that EL may implicitly cover much of the SL distribution, possibly due to the compositionality of language. However, simply adding SL to the randomly ordered training data does not improve overall performance – and while it significantly improves EL perplexity, it worsens performance on SL itself. This asymmetry hints at a negative interference effect as observed in multilingual model training (Wang et al., 2020): though both classes stem from the same language, they might be different enough to cause gradient conflicts when used in the same dataset. These findings emphasise that learning patterns across language registers are not symmetric, and underscore the importance of evaluating perplexity for different subsets.

Structure matters: the effectiveness of label-based curricula Models only benefit from SL when introduced in a structured way. Sequential label-based curricula, where training begins with SL before using EL, consistently outperform other strategies in overall and SL perplexity. This aligns with the idea that simplified input can serve as a scaffold, supporting the acquisition of more complex patterns. While the effect mirrors principles observed in human learning, the underlying reason why structured exposure aids generalisation may differ in MLM.

The limits of difficulty heuristics Competence-based curricula using shallow difficulty heuristics show no clear advantage over random strategies. While this supports prior findings by Surkov et al. (2022), our analysis offers further insight. Figure 1 shows histograms comparing the distribution of shallow heuristics in SL and EL and Table 1 illustrates some examples. While it is plausible that EL has samples at the "easy" extremes, as not every sentence in everyday language is necessarily complex, we also observe SL examples at the "complex" extremes. Assuming that SL represents text that is easier to understand for humans, this highlights that the difficulty heuristics fail to meaningfully separate the two classes.

Future Directions We find that while shallow difficulty heuristics do not suffice to guide CL, the information encoded in the language classes does. Despite high lexical overlap and comparable size (Tables 2 and 4), simple language may offer more than surface-level simplicity. Prior work has shown that both humans and neural models benefit from regular, compositional input (Galke et al., 2024) and simple language might reflect just that through syntactic consistency or clearer discourse structure. Future work could explore how such compositional features manifest in simple language, and whether they can be modelled or annotated as difficulty signals – enabling broader and more effective CL strategies in MLM pre-training.

6 Conclusion

We examined whether human-curated simple language can guide CL in MLM pre-training. Our results show that label-based curricula outperform both random baselines and competence-based approaches relying on shallow difficulty heuristics. While the two language classes show high lexical and distributional overlap, their ordering – particularly when first training on simple language before moving to everyday language – leads to significant gains in model performance. This suggests that human intuition about linguistic difficulty provides more effective structure for CL than traditional surface-level heuristics.

References

Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, and Sarah Ebling. 2020. A Corpus for Automatic Readability Assessment and Text Simplification of German. In *Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020*, pages 3302–3311. European Language Resources Association.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum Learning. In *Proceedings of the 26th International Conference On Machine Learning, ICML 2009*, pages 41–48.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A New Text Simplification Task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL 2011*, volume 3, pages 665–669. Association for Computational Linguistics.

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

Rudolph Flesch. 1948. A New Readability Yardstick. *Journal of Applied Psychology*, 32:221–233.

Lukas Galke, Yoav Ram, and Limor Raviv. 2024. What Makes a Language Easy to Deep-Learn? Deep Neural Networks and Humans Similarly Benefit from Compositional Structure. *Nature Communications*, 15(1):10816.

Philip A. Huebner, Elior Sulem, Cynthia Fisher, 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, CoNLL 2021*, pages 624–646. Association for Computational Linguistics.

David Kauchak. 2013. Improving Text Simplification Language Modeling Using Unsimplified Text Data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2013*, volume 1, pages 1537–1546.

Cao Liu, Shizhu He, Kang Liu, and Jun Zhao. 2018. Curriculum Learning for Natural Answer Generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, pages 4223–4229. ijcai.org.

Evan Lucas, Dylan Gaines, Tagore Rao Kosireddy, Kevin Li, and Timothy C. Havens. 2024. Using Curriculum Masking Based on Child Language Development to Train a Large Language Model with Limited Training Data. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 221–228. Association for Computational Linguistics.

Tasnim Mohiuddin, Philipp Koehn, Vishrav Chaudhary, James Cross, Shruti Bhosale, and Shafiq Joty. 2022. Data Selection Curriculum for Neural Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1569–1582. Association for Computational Linguistics.

Aaron Mueller and Tal Linzen. 2023. How to Plant Trees in Language Models: Data and Architectural Effects on the Emergence of Syntactic Inductive Biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 11237–11252. Association for Computational Linguistics.

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 12th International Conference on Recent Advances in Natural Language Processing, RANLP 2021*, pages 989–996. INCOMA Ltd., Shoumen, Bulgaria.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom M. Mitchell. 2019. Competence-Based Curriculum Learning for Neural Machine Translation. 1:1162–1172.

Leonardo Ranaldi, Giulia Pucci, and Fabio Massimo Zanzotto. 2023. Modeling Easiness for Training Transformers with Curriculum Learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, RANLP 2023*, pages 937–948. INCOMA Ltd., Shoumen, Bulgaria.

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. Association for Computational Linguistics.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A Survey on Curriculum Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9).

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On Negative Interference in Multilingual Models: Findings and A Meta-Learning Treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 4438–4450. Association for Computational Linguistics.

Wojciech Zaremba and Ilya Sutskever. 2015. Learning to Execute. *Preprint*, arXiv:1410.4615.

A Implementation Details

We provide our implementation details for the competence-based CL strategy, where each training sample is assigned a difficulty score and the dataset is sorted accordingly. A predefined competence function then controls the fraction of data available at each training step t, gradually increasing the difficulty over time. Following Platanios et al. (2019), we adopt the square-root based competence function, which they found to be most effective:

$$c_{sqrt}(t) = \min(1, \sqrt{\frac{t(1-c_0^2)}{T}}) \in [0, 1],$$

where c_0 denotes the initial competence at $t = 0$ and T is the total number of steps in the CL phase. In our experiments, we observed that shorter competence phases tend to yield better results than longer ones. We pick $T = 50\,000$ and $c_0 = 0.05$ as function parameters. The size of the training dataset is updated every 5 000 steps depending on the current function value.

B Difficulty Heuristics

In our work we consider three popular heuristics to measure the difficulty of text for global, data-level curriculum learning (cf. Platanios et al. (2019) or Ranaldi et al. (2023)). Let S be a sentence, represented by a finite sequence of words $(w_1, w_2, \dots, w_m)$. The first heuristic, sentence length, is defined by the number of words in the sentence:

$$\text{length}(S) = |S|.$$

Next, we use the word rarity metric as proposed by Platanios et al. (2019), but normalise it by the number of words to remove its strong correlation with the sentence length:

$$\text{word rarity}(S) = -\frac{1}{|S|}\sum_{w \in S} \log\left(\frac{\text{count}_c(w)}{N}\right),$$

where N denotes the size of the vocabulary of the corpus and $\text{count}_c(w)$ the number of times w appeared in the corpus. Last, we present the Flesch Reading Ease (FRE) score as defined by Flesch (1948). It is designed to evaluate the readability of text and to return a score between 0 and 100:

$$\text{FRE}(S) = 206.835 - 1.015 \times \text{ASL} - 84.6 \times \text{ASW},$$

where ASL denotes the average sentence length, which is always the actual sentence length since we only evaluate single sentences, and ASW denotes the average syllables per word. Since the FRE was designed to evaluate text samples of 100 words, we can encounter negative FRE scores which are outside the originally defined range.

Strategy	PPL	SL PPL	EL PPL
Baseline SL+EL	.445 (w)	.996 (w)	**.004** (w)
Incremental	.598 (b)	1.00 (w)	**.008** (w)
Sequential	**.019** (w)	**.001** (w)	.126 (w)
Anti-Sequential	.252 (b)	1.00 (w)	**.008** (w)
Length	.890 (w)	.977 (w)	.899 (w)
Word Rarity	.890 (b)	.977 (w)	.718 (w)
FRE	.779 (w)	.977 (w)	.899 (w)
Random	.779 (w)	.977 (w)	.899 (w)

Table 5: Adjusted p-values for all statistical tests for the models' performance on overall perplexity (PPL), simple language perplexity (SL PPL), and everyday language perplexity (EL PPL). We choose $\alpha = 0.05$ and boldface all significant results. We further indicate which one-sided test was run: (w) Wilcoxon signed-rank test or (b) boostrap median test.

C Details on the Significance Tests

Table 5 reports the adjusted p-values for all strategies, assessing their performance relative to relevant baselines. For each comparison, we applied a one-sided test based on our directional hypotheses: (1) whether adding SL (Baseline SL+EL) *improves* over the baseline trained with EL (Baseline EL); (2) whether label-based curricula (Incremental and Sequential) *improve* over the full baseline (Baseline SL+EL); (3) whether Anti-Sequential *hurts* performance compared to Baseline SL+EL; and (4) whether competence-based strategies (Length, Word Rarity, FRE, Random) *improve* over the Baseline SL+EL.