Mmi01 at The BabyLM Challenge: Linguistically Motivated Curriculum Learning for Pretraining in Low-Resource Settings

Maggie Mi
Department of Computer Science
The University of Sheffield
zmi1@sheffield.ac.uk

Abstract

This paper presents our findings for the BabyLM Challenge (Warstadt et al., 2023). Our exploration is inspired by vanilla curriculum learning (Bengio et al., 2009) and we explored the effect of linguistic complexity in forming the best curriculum for pre-training. In particular, we explore curriculum formations based on dependency-based measures (dependents per token, average dependency distance) and lexical-based measures (rarity, density, dispersion and diversity). We found that, overall, models pretrained using curriculum learning were able to beat the performance of a non-curriculum learning pre-trained model. Furthermore, we notice using different linguistic metric for measuring complexity lead to advantageous performance for some tasks, but not all. We share our results and analysis in the hope that it can provide beneficial insights for future work.

1 Introduction

Currently, pretraining language models (LMs) involve training models on large, diverse datasets before fine-tuning them on specific downstream tasks. As a byproduct of this procedure, datasets have grown substantially beyond developmentally plausible amounts. For instance, the recently released large variant of LLAMA-2 has 70 billion parameters and it was pre-trained with 2 trillion tokens (Touvron et al., 2023). This amount of data is well over the amount of exposure a child would have. Gilkerson et al. (2017) find that on average, a child aged 48-mo would be exposed to 12,128 tokens, from solely their parents. Calculations show LLAMA2's pretraining data is 165,000 times more than this developmental-plausible quantity.

Therefore, the goal of this task is to use human-development plausible methods for pretraining smaller-sized language models. In particular, we combine intuitions from linguistics and curriculum learning to explore whether different curricula designs affect models' performance. To do this, we investigate two strands of complexity measures, namely, structural complexity and lexical complexity.

Our research questions (RQs) are as follows:

1. Do pre-training LMs using CL produce better performance? If so:

2. Are linguistic complexity measures helpful in designing curricula for CL?

3. Which linguistic metric is advantageous and which is less? Is one strand of complexity measure inherently better than the other?

To answer RQ1, we aim to compare a baseline non-CL model to the results of CL-pretrained models. For RQ2, we make a similar comparison but this time using the results of a model that is trained on a random curriculum. For the last RQ, we make inter-model comparisons.

We provide an analysis of curriculum designs and the novel aspects of our work (§ 2). Following this, we explain the linguistic metrics in detail and provide details of our approach (§ 3). In § 4, we present our findings and discussions, before finally summarising the paper in § 5.

2 Related Works

Curriculum learning (CL) was first proposed by Bengio et al. (2009). The idea behind curriculum learning comes from the pedagogical observation that animals and humans learn better when knowledge is presented in a meaningfully organised way. For instance, starting with simple examples and gradually advancing to more complex ones (Skinner, 1958; Sweller, 1994; Krueger and Dayan, 2009). In the language modelling experiment carried out by Bengio et al. (2009), a corpus replacement method was used to make the data

Proceedings of the 27th Conference on Computational Natural Language Learning: Volume 2: The BabyLM Challenge, pages 269–278
December 6-7, 2023 ©2023 Association for Computational Linguistics

increasingly difficult. This way of pertaining was found to be more effective, producing improved results.

There have been then numerous works have explored using CL as the pretraining approach for language models. Whilst some works reported CL as beneficial to pretraining, others have reported the opposite results. Nagatsuka et al. (2021) investigated a CL-based pretraining scheme that utilises the length of the input text as the measure of "difficulty" in curriculum design. It was found that using length-based curriculum training alongside using the maximum available batch size, models achieved drastically faster convergence speed, and higher scores on downstream tasks (Nagatsuka et al., 2021, 2022) .

Curriculum design greatly varies in each work. Linguistic features that have been used in curriculum formation include Parts-of-Speech (POS) information, n-gram frequency (Platanios et al., 2019), average number of dependents per word in the sentence parse tree (Jafarpour et al., 2021), edit distance (Kadotani et al., 2021; Chang et al., 2021). However, arguably, the most common curriculum formations are based on measures of frequency (Liu et al., 2018) and text length (Tay et al., 2019; Cirik et al., 2016).

Comparing curriculum learning studies becomes challenging due to the inherent variability in curriculum choices across different tasks. However, it is undeniable that the arrangement of data holds significance. As a result, in distinction from prior research, our work is oriented towards investigating diverse linguistic features in curriculum formation. Notably, we investigate 5 different measures of linguistic complexity. They are:

- Average dependency distance (ADD)
- Dependents per word (DPW)
- Lexical rarity (RARITY)
- Lexical density (DENSITY)
- Lexical Evenness (DISPERSION)
- Lexical diversity (TTR)

We choose these measures of linguistic complexity to address the multi-dimensionality of measuring language complexity. In particular, we consider not only lexical (vocabulary-based) information, but also syntactical (structural-based) complexity measures. To the best of our knowledge, this study is the first to consider curriculum formation using such a comprehensive set of measures. Moreover, we focus our experimentation specifically on low-resource, data-constrained scenarios. As a result, we adopt a simple CL approach to reflect these settings.

3 Methodology

Our submission considers GPT-2 models (Radford et al., 2019) pretrained using curricula formed by various linguistic measures detailed in § 2. The pretraining approach involves sequentially training the model using ten different curriculum levels of the dataset, with each level building upon the previous one in terms of difficulty. Each model is pretrained three times, with a random seed used each time.

3.1 Curricula Formations

We used the 10M words dataset provided by the task authors for the STRICT-SMALL track of the Challenge. As detailed in the task description, the dataset consists of 10 excerpts, sourced from 10 different corpora of mixed domains (Warstadt et al., 2023). We consider all of the models to qualify for the LOOSE track, and only the evenness and lexical diversity models are legible for the STRICT-SMALL track. This is due to the fact that we use existing scripts from `textcomplexity`[1], which makes use of external tools, such as POS taggers trained on much more data than the given amount for linguistic complexity calculations.

For each part of the overall dataset, a score for each linguistic metric was calculated. As an example, Table 1 provides the TTR scores of each subset of the data. Curriculum formation is based on this ranking, with the "easiest", or in this case, the least lexically diverse data being Open Subtitles and the "hardest" being the Wikipedia data. Using the same idea, other curricula were formed using each linguistic measure.

3.1.1 Syntactic Diversity (DPW)

DPW quantifies the average number of syntactic dependents (i.e., words that depend on another word for their grammatical function) in a given text per word. A DPW score indicates that, on average, each word in a sentence has a large number of syntactic dependents. This means that the sentence has a complex and intricate syntactic structure, with many words relying on each other to convey meaning and grammatical relationships. Sentences with

[1] https://github.com/tsproisl/textcomplexity

high DPW scores tend to be more challenging for humans to process and understand (Hawkins, 1994; Grodner and Gibson, 2005; Gibson, 1998).

3.1.2 Syntactic Proximity (ADD)

ADD is mathematically defined as (Liu et al., 2009):

$$\text{ADD} = \frac{1}{n-s}\sum_{i=1}^{n-s}|DD_i|$$

where:

- n is the total number of tokens in the sentence
- s is the total number of sentences in the document
- DD_i is the dependency length of the i-th syntactic link

Conceptually, this is calculating a ratio of calculating the total lengths of dependency links in a sentence to the total number of dependencies links in the same sentence. It gives an indication of how closely related the words are in a sentence syntactically. A lower average dependency distance suggests that the words in a sentence tend to be more closely connected, indicating a more compact sentence structure. Conversely, a higher average dependency distance suggests more complex and possibly longer distances between heads and their dependents in a sentence (Oya, 2011).

3.1.3 Lexical Rarity (RARITY)

As detailed in `textcomplexity`, rarity was calculated with the help of the COW frequency list (Schäfer, 2016). More frequent lexical items were given a smaller score.

3.1.4 Lexical Density (DENSITY)

Lexical density is calculated as the proportion of content words to function words. We consider a higher score on this metric as data that is harder to learn since it is more likely to be information-heavy.

3.1.5 Lexical Evenness (DISPERSION)

Dispersion is measured using Gini-based dispersion (Gini, 1912). It measures how evenly tokens of the same type are distributed in the text (Blombach et al., 2022). The Gini-based dispersion for a single type is computed as

$$1 - \frac{Gini}{Gini_{max}}$$

where $Gini$ is the Gini coefficient of the distances between tokens of the same type, and $Gini_{max}$ is the maximum value for a type with frequency f in a text of length N.

The formula for $Gini_{max}$ is:

$$\text{Gini\_max} = \frac{(N-f)\cdot(f-1)}{f\cdot N}$$

where

- N is the length of the entire text (total number of tokens in the text)
- f is the frequency of the type (number of times a particular token appears in the text)

In this work, evenness serves to illustrate the arrangement or spread of token types within a text.

3.1.6 Lexical Diversity (TTR)

Type-token ratio (TTR) is used to measure lexical diversity. It is calculated by dividing the number of unique words (types) to the total number of words (tokens) present in the text (Templin, 1957). This can be thought of as measuring the richness of the vocabulary of the corpus. A higher TTR indicates a more diverse vocabulary with a greater range of unique words in the text. Conversely, a lower TTR suggests a more repetitive or limited use of vocabulary.

TTR is given by :

$$\text{TTR} = \frac{\textit{Number of different word types}}{\textit{Total number of tokens}}$$

Table 1: TTR scores of each subset of the 10M words dataset, shown in increasing order.

Subset	TTR Score
open subtitles	1.623
bnc spoken	2.034
aochildes	2.068
qed	2.966
cbt	3.450
children_stories	3.570
switchboard	3.997
gutenberg	4.149
simple wikipedia	4.491
wikipedia	5.678

3.2 Model Description

We use the provided data to train a Unigram-16000 tokeniser, and our experiments all use this tokeniser.

In this Challenge, we focus specifically on smaller settings of the models. All models featured

in this work are trained on architectures with 12 layers and 12 attention heads [2]. Our focus is directed towards this smaller setting since smaller models typically require less computational power and memory, making them more accessible and cost-effective for researchers with limited resources.

3.3 Model Evaluation

All models undergo evaluation on The Benchmark of Linguistic Minimal Pairs (BLiMP) benchmark as well as SuperGLUE and MSGS tasks. We run each evaluation suite three times for every model. Each run uses a different random seed.

BLiMP is an evaluation suite that tests LMs' abilities on a range of grammatical phenomena in the English language (Warstadt et al., 2020a). For BLiMP tasks, a zero-shot evaluation approach is used, allowing the models to be assessed without any additional fine-tuning. On the other hand, to gauge the models' performance on SuperGLUE tasks, they are subjected to fine-tuning using the respective datasets.

SuperGLUE is a benchmark that comprises challenging language understanding tasks. Inspired by GLUE, SuperGLUE aims to address the limitations of the original GLUE benchmark (Wang et al., 2018), which had gradually lost its challenge due to the improving capabilities of LMs.

Mixed Signals Generalization Set (**MSGS**) assess whether language models exhibit preferences for certain aspects of language, such as linguistic features (e.g., specific sentence structures) or surface features (e.g., word positioning). The MSGS dataset evaluates whether language models can identify and detect these linguistic and surface features and whether they prioritize linguistic features over surface features, which is a crucial aspect of human language understanding abilities (Warstadt et al., 2020b).

Taken together, these evaluation suites provide insights into the models' general language understanding capabilities as well as their adaptability and performance on specific downstream tasks. The code for this task's evaluation originates from *eval-harness* by Gao et al. (2021). Furthermore, as a fascinating aspect of cognitive modelling, we assess our models' capability to predict the **age of word acquisition (AoA)**. Based on the work of Portelance et al. (2023), computing this metric involves an estimation of the average surprisal of words in child-directed utterances sourced from CHILDES. Models are then evaluated using leave-one-out cross-validation. The metric used to measure prediction is mean absolute deviation (MAD). A lower MAD score indicates that the model's predictions are closer to the actual age of acquisition, signifying better performance on the task. Conversely, a higher MAD score suggests that the model's predictions are less accurate.

Baselines: The two baseline models we use are a model trained without CL (NONCL) and a model trained on a randomly formed curriculum (RANDOM). The non-CL model represents a conventional approach, where the model is trained on all available data simultaneously for a fixed number of steps (50000 in this case). On the other hand, the CL model trained on a randomly formed curriculum serves as a comparison to understand how much improvement linguistically justified curricula can provide.

4 Results

The results of models can be seen in Tables 2, 3, 4. We provide the performance results for the supplement BLiMP tasks and MSGS tasks (see Table 6 and 7). The analysis of the main BLiMP, SuperGLUE and AoA prediction tasks serves as a representative basis, and the conclusions drawn from these tasks can be extended to the results presented in the Appendices. The analysis presented takes into consideration the results of all evaluation metrics, however, we mainly focus on the BLiMP, SuperGLUE and AoA benchmarks; MSGS and the supplement BLiMP tasks will be referred to on a needs basis.

4.1 Non-CL vs. CL

By comparing the non-curriculum learning pretrained baseline model (NONCL) with models pretrained using curriculum learning, we observe that the latter exhibit slightly better performance. For most of the tasks, CL models (RANDOM, ADD, DPW, DISPERSION, DENSITY, RARITY, TTR) outperform NONCL. Higher scores are observed in these systems on BLiMP tasks such as ANA. AGR, ARG. STR, QUANTIFIERS and SuperGLUE tasks such as QQP, BoolQ, and MultiRC indicating that curriculum learning leads to better performance. Although the improvements are not substantial in some cases and there exist also situations where

[2]The code for curriculum formation and training can be found on Github: https://github.com/mi-m1/BabyLM-Entry.

Model	ANA. AGR	ARG. STR	BINDING	CTRL. RAIS	D-N AGR	ELLIPSIS	FILLER. GAP	IRREGULAR	ISLAND	NPI	QUANTIFIERS	S-V AGR
NONCL baseline	50.19	58.53	46.26	55.57	50.31	38.41	28.94	47.96	45.71	45.74	30.52	48.35
RANDOM baseline	61.28	59.53	48.74	56.87	49.24	40.36	28.99	56.49	52.14	23.70	38.01	46.91
ADD	**64.37**	59.53	47.03	55.04	49.58	37.30	29.05	48.50	48.13	31.49	55.15	**48.93**
DISPERSION	63.19	59.78	46.82	57.39	49.58	39.32	28.94	52.60	51.08	45.27	**46.68**	**48.93**
DPW	63.80	**59.86**	**49.80**	56.79	49.38	37.88	28.99	**59.29**	51.97	30.11	43.39	47.77
DENSITY	65.56	59.66	45.83	**57.56**	49.25	40.07	29.76	49.86	50.31	41.48	42.25	**48.93**
RARITY	59.01	59.78	49.01	57.11	49.47	38.51	29.03	56.28	50.75	18.91	41.60	48.43
TTR	59.00	59.13	44.99	56.99	**49.78**	36.76	**30.58**	47.75	49.00	**49.52**	33.08	48.55

Table 2: Table showing BLiMP results of models. All results are average performance accuracy over three runs. **Bold** values are results that are the best performance achieved average for the given task. These values are also statistically significantly better than the baseline CL model tested with Welch's t-test ($p < 0.05$).

the NONCL model has exceeded CL models, for instance, in CoLA and MRPC. Comparisons between the random CL baseline model (RANDOM) and models trained on structured curriculum suggest that training data on increasing lexical complexity can contribute to improved performance, albeit to a limited extent.

Since the models are trained on small amounts of data, they are likely to overfit. Future investigations can explore more computationally complex methods, such as competency-based scheduling functions to make more robust decisions on when to expose a new level of curriculum to the model (Platanios et al., 2019).

4.2 Best and Worst Curriculum Design

Considering the similarity of the results and the diverse nature of the evaluation tasks, we determine the best model as the one that outperforms the baseline CL model statistically significantly in the highest number of tasks. We find that the best curriculum depends on the evaluation suite. On BLiMP tasks, the best curriculum is found to be DENSITY; ADD on SuperGLUE tasks; TTR on MSGS tasks. Interestingly, the curriculum that demonstrated the fewest instances of outperforming the baseline across all evaluation suites is DISPERSION. From these observations, organising pretraining data according to syntactic complexity is perhaps more advantageous on the SuperGLUE and MSGS tasks, whereas lexical information is more effective for gaining the knowledge required to perform well on BLiMP tasks. The best aggregate model is found to be pretrained by ADD curriculum. This could indicate that exposing data incrementally to the model based on sentence structure is a modest choice for curriculum design.

4.3 Curriculum Design Variation

The variation in performance between each model is observed to be diverse across all evaluation schemes. On average, the gap in performance between the best and worst CL model on SuperGLUE tasks (3.072) and MSGS tasks (4.670) is smaller than on BLiMP (7.440) and supplement BLiMP tasks (5.763). This difference in spread shows that models perform more consistently on finetuning tasks than BLiMP ones. We attribute this to the nature of the evaluation tasks. SuperGLUE comprises a variety of natural language understanding tasks, but they may share certain linguistic or semantic characteristics that make them more predictable for models to generalize across tasks. On the other hand, BLiMP tasks are designed to test specific linguistic phenomena, making them more challenging and potentially leading to greater variation in model performance. Furthermore, given that a portion of the dataset comprises transcribed spoken speech, the exposure to intricate linguistic structures may be restricted, as spoken language tends to be less complex than written language. For instance, Chang and Bergen (2022) find that the average mean sentence length in the CHILDES corpus is 4.5 tokens. This adds plausibility to the fact that spoken language contains simpler syntactic structures.

4.4 Age-of-Acquisition Prediction Results

We find that some of the results for AoA predictions are statistically insignificant. In particular, we see that the models are unable to predict AoA for Overall and Nouns categories. Out of the results that are statistically significant, ADD is able to predict predicates more accurately than the NONCL model and functions words more accurately than the RANDOM model. DISPERSION and DENSITY models have higher accuracy on function words predictions than ADD model.

4.5 Difficulties

Overall, there are fewer instances where the models are able to exceed the CL baseline on SuperGLUE tasks. However, the hardest tasks, whereby models achieved the lowest scores are mostly BLiMP tasks.

Model	CoLA (MCC)	SST-2	MRPC (F1)	QQP (F1)	MNLI	MNLI-mm	QNLI	RTE	BoolQ	MultiRC	WSC
NONCL baseline	69.48	83.27	65.35	69.65	58.07	57.89	56.74	55.22	60.40	48.63	58.63
RANDOM baseline	68.92	83.14	63.65	71.30	57.76	58.84	58.30	50.84	64.08	52.39	61.45
ADD	68.56	83.07	62.90	70.56	58.56	58.94	57.58	**55.89**	62.52	50.38	**61.45**
DISPERSION	68.53	82.94	63.09	70.28	58.33	**59.44**	57.98	52.86	62.84	49.65	61.45
DPW	68.92	82.15	**63.65**	69.66	58.36	59.30	58.18	53.87	62.38	50.16	61.45
DENSITY	**69.12**	82.94	57.63	70.27	**58.94**	57.69	57.60	51.18	62.89	50.93	57.83
RARITY	67.71	82.87	59.32	**73.79**	58.31	57.89	56.39	51.18	62.24	51.92	61.45
TTR	69.09	82.35	60.83	73.53	58.61	59.20	55.89	53.20	57.81	48.67	60.24

Table 3: Table showing SuperGLUE results of models. All results are average performance accuracy over three runs. Matthews correlation is reported for CoLA; F1 scores are reported for MRPC and QQP; the rest are accuracy scores. **Bold** values are results that are the best-performing model for the given task. These values are also statistically significant, tested using Welch's t-test ($p < 0.05$).

Model	Overall (591 words)	Nouns (322 words)	Predicates (167 words)	Function words (102 words)
NONCL baseline	2.053	1.970	1.867	2.619
RANDOM baseline	2.050	1.968	1.850	2.640
ADD	2.051	1.970	1.851*	2.637*
DISPERSION	2.053	1.973	1.854	2.632*
DPW	2.051	1.971	1.847	2.640
DENSITY	2.051	1.970	1.852	2.632*
RARITY	2.049	1.969	1.845	2.637
TTR	2.052	1.969	1.862	2.626

Table 4: Table showing Age-of-Acquisition prediction results of models. The scores are mean absolute deviation in months across Leave-One-Out (LOO) cross-validation folds. Lower MAD scores denotes higher accuracy. Values* are results that are significant, tested using Welch's t-test ($p < 0.005$).

Namely, NPI (lowest = 18.91), FILLER GAP, and QUANTIFIERS (lowest = 33.08), as can be seen in Table 2).

As noted by Warstadt et al. (2020a), tasks such as NPI licensing and Quantifiers require in-depth semantic knowledge. LMs seem to lack such knowledge, as they tend to make errors that produce contradictory language and show a lack of understanding of assumptions and ideas (Marvin and Linzen, 2018). Interestingly, upon inspecting the predictions made by the models, it appears that there is a strong preference for constructions that contain the adverb "*ever*". In fact, all the predictions made by the models incorporated this adverb. The predictions for the Quantifier task also exhibit consistent patterns of ungrammatically. For instance, they do not seem to know superlative quantifiers cannot be embedded under negation.

Table 5 provides examples that illustrate these judgements. Taken together, this effectively shows the models have not been able to generalise conditions for NPI licensing, which is, that NPIs prefer not to occur in positive sentences and are restricted to specific contexts, primarily negative environments. In addition, the models seem to have also not learned that NPI licensing environments exist and can take the form of negation and negative quantifiers. Similarly, the model has not learned the required knowledge for resolving the right quantifier constructions.

In this light, solely relying on CL with varying kinds of lexical complexity for forming curricula may not be sufficient. Additional efforts are required to explicitly introduce language models with the knowledge necessary for completing both semantic and syntax tasks successfully. This draws questions to LMs' abilities to generalise syntactical patterns in language. Whilst this 10M-word corpus might be sufficient for humans acquiring language, LMs perhaps require more targeted training and additional data.

5 Conclusion

In this work, we investigated different CL curricula. We find that linguistically-motivated curriculum formation produces better results than (1) a non-CL pretrained model, and (2) a CL model trained on a randomly formed curriculum. In addition, we provide an analysis of the impact of linguistic curriculum on evaluation tasks. The findings underscore the potential of leveraging linguistic principles to address the challenges posed by sequential learn-

ID	Prediction
npi_licensing_9	"Should Mitchell ever know Eva?"
npi_licensing_43	"Sharon has ever climbed down a hill."
quantifiers_62	"There weren't most gates looking like most photographs."

Table 5: Example BLiMP predictions made by the models

ing tasks and pave the way for further research in this promising direction. One possible direction to explore is the adaptive CL approach, which dynamically adjusts the curriculum based on the model's learning progress and task complexities. This could involve incorporating feedback mechanisms to fine-tune the curriculum during training for optimal task mastery. With this work as a foundation, we hope it can provide insights to linguistically-oriented pertaining works.

6 Limitations

We would like to point out that more advanced features, such as discourse features and additional semantic features provided by Lee et al. (2021) form promising areas of exploration. Arguably, including these features will paint a more representative of linguistic complexity. However, as a starting point, we frame our work to first isolate each "dimensionality" of linguistic complexity, and explore each one's effect in pretraining independently.

Acknowledgements

This work is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications are funded by the UK Research and Innovation grant EP/S023062/1. We are grateful to the reviewers for their contributions and feedback. Special thanks to Aline Villavicencio for the insights and directions. Additional thanks to Ed Gow-Smith, Dylan Phelps, Bohua Peng and members of the CDT for making this work happen.

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.

Andreas Blombach, Stephanie Evert, Fotis Jannidis, Steffen Pielström, Leonard Konle, and Thomas Proisl. 2022. *Digital Humanities 2022*, 2022 edition, page 130–134. University of Tokyo.

Ernie Chang, Hui-Syuan Yeh, and Vera Demberg. 2021. Does the order of training samples matter? improving neural data-to-text generation with curriculum learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 727–733, Online. Association for Computational Linguistics.

Tyler A. Chang and Benjamin K. Bergen. 2022. Word Acquisition in Neural Language Models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Volkan Cirik, Eduard H. Hovy, and Louis-Philippe Morency. 2016. Visualizing and understanding curriculum learning for long short-term memory networks. *ArXiv*, abs/1611.06204.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Jill Gilkerson, Jeffrey A. Richards, Steven F. Warren, Judith K. Montgomery, Charles R. Greenwood, D. Kimbrough Oller, John H. L. Hansen, and Terrance D. Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, 26(2):248–265.

C. Gini. 1912. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche. [Fasc. I.]*. Studi economico-giuridici pubblicati per cura della facoltà di Giurisprudenza della R. Università di Cagliari. Tipogr. di P. Cuppini.

Daniel Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, 29(2):261–290.

John A Hawkins. 1994. *A performance theory of order and constituency*. 73. Cambridge University Press.

Borna Jafarpour, Dawn Sepehr, and Nick Pogrebnyakov. 2021. Active curriculum learning. In *Proceedings*

of the First Workshop on Interactive Learning for Natural Language Processing, pages 40–45, Online. Association for Computational Linguistics.

Sora Kadotani, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. 2021. Edit distance based curriculum learning for paraphrase generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 229–234, Online. Association for Computational Linguistics.

Kai A Krueger and Peter Dayan. 2009. Flexible shaping: How learning in small steps helps. *Cognition*, 110(3):380–394.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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-18*, pages 4223–4229. International Joint Conferences on Artificial Intelligence Organization.

Haitao Liu, Richard Hudson, and Zhiwei Feng. 2009. Using a chinese treebank to measure dependency distance. 5(2):161–174.

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

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.

Koichi Nagatsuka, Clifford Broni-Bediako, and Masayasu Atsumi. 2022. Length-based curriculum learning for efficient pre-training of language models. *New Gen. Comput.*, 41(1):109–134.

Masanori Oya. 2011. Syntactic dependency distance as sentence complexity measure. *Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics*.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. 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 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.

Eva Portelance, Yuguang Duan, Michael C. Frank, and Gary Lupyan. 2023. Predicting age of acquisition for children's early vocabulary in five languages using language model surprisal. *Cognitive science*, 47 9:e13334.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Roland Schäfer. 2016. CommonCOW: Massively huge web corpora from CommonCrawl data and a method to distribute them freely under restrictive EU copyright laws. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4500–4504, Portorož, Slovenia. European Language Resources Association (ELRA).

Burrhus F Skinner. 1958. Reinforcement today. *American Psychologist*, 13(3):94.

John Sweller. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4:295–312.

Yi Tay, Shuohang Wang, Anh Tuan Luu, Jie Fu, Minh C. Phan, Xingdi Yuan, Jinfeng Rao, Siu Cheung Hui, and Aston Zhang. 2019. Simple and effective curriculum pointer-generator networks for reading comprehension over long narratives. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4922–4931, Florence, Italy. Association for Computational Linguistics.

Mildred C. Templin. 1957. *Certain Language Skills in Children: Their Development and Interrelationships*, new edition edition, volume 26. University of Minnesota Press.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas

Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. 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, 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*, 8:377–392.

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.

A Additional Results

Results on the supplement BLiMP tasks and MSGS tasks are shown in Table 6 and Table 7, respectively.

	HYPERNYM	QA CONGRUENCE EASY	QA CONGRUENCE TRICKY	SUBJECT AUX INVERSION	TURN TAKING
NONCL baseline	50.85	27.60	32.73	62.63	51.31
RANDOM baseline	50.85	34.90	28.69	59.32	50.00
ADD	49.42	**37.50**	29.09	**67.16**	46.43
DISPERSION	50.81	28.65	30.30	64.16	50.95
DPW	50.58	30.73	28.69	57.37	48.10
DENSITY	49.92	32.81	30.71	65.37	47.98
RARITY	**50.54**	34.38	28.08	61.77	49.40
TTR	50.50	31.25	**32.32**	61.06	**50.83**

Table 6: Table showing results of supplement BLiMP tasks. All results are average performance accuracy over three runs. **Bold** values are results that are the best performance achieved average for the given task. These values are also statistically significantly better than the baseline CL model tested with Welch's t-test ($p < 0.05$).

	CR_CTRL	LC_CTRL	MV_CTRL	RP_CTRL	SC_CTRL	CR_LC	CR_RTP	MV_LC	MV_RTP	SC_LC	SC_RP
NONCL	59.64	79.23	82.98	98.85	60.58	54.61	23.22	29.39	23.92	40.82	35.40
RANDOM	59.78	93.15	82.34	99.75	60.21	51.90	24.95	23.59	26.22	40.82	34.69
ADD	**61.19**	**98.30**	76.38	99.75	60.18	48.78	23.43	22.81	22.81	40.84	30.09
DISPERSION	58.34	93.07	79.17	99.72	59.35	50.91	24.69	22.67	23.48	40.84	31.64
DPW	59.12	87.27	79.97	99.64	59.25	42.76	23.68	23.41	25.26	40.82	37.18
DENSITY	59.39	88.63	75.69	99.75	**61.39**	50.98	26.02	22.66	24.53	40.80	33.75
RARITY	58.70	92.71	76.71	**99.82**	59.96	49.84	24.49	27.18	25.62	40.82	35.10
TTR	59.17	85.87	81.26	99.09	59.20	46.57	**27.83**	**28.82**	**26.92**	**40.84**	**39.42**

Table 7: Results of MSGS evaluation. All results are Matthews correlation coefficients (MCCs). All results are average performance accuracy over three runs. **Bold** values are results that are average MCC for the given task. These values are also statistically significantly better than the baseline CL model tested with Welch's t-test ($p < 0.05$).