SlovakBabyLM: Replication of the BabyLM and Sample-efficient Pretraining for a Low-Resource Language

Ľuboš Kriš and Marek Šuppa
Department of Applied Informatics, Comenius University in Bratislava
NaiveNeuron
Kempelen Institute of Intelligent Technologies
lubos.kris@kinit.sk, marek.suppa@fmph.uniba.sk

Abstract

In recent years, we have seen the creation of various specific language models (LMs) within the Slavic language family, which contain much fewer resources to create LMs, than other languages. However, with an increasing number of parameters of LM, a larger amount of text is required for good performance, which can hinder the development and creation of LMs for low-resource languages (LRL). Our research is looking for a solution in Curriculum learning (CL) methods that can help us build better models with a lower amount of text in comparison with current LMs, which can help in better pretraining of models with LRL. Therefore, we replicate the BabyLM Challenge in the Slovak language[1][2]. Additionally, apply CL methods for finding the difference in the application of CL methods on the English and Slovak languages, and evaluate whether the CL improves the performance of LM in the Strict-small track. Our experiments show that the use of CL methods as preprocessing methods is significant for improving model performance in sentiment analysis and question answering.

1 Introduction

According to a study by Joshi et al. (2020), seven languages are at the forefront that contain a large amount of data in the online space, which enables them to build large language models (LLMs). However, Slovak is not in that group. The lack of text in the online space may lead to less variability and text selection for LM pretraining. As part of the BabyLM challenge, researchers are trying to improve the pretraining on the amount of language needed for comprehension in children under 13 years of age (Warstadt et al., 2023b). The limited availability of Slovak text motivates this research. Furthermore, humans still outperform LLMs in certain language tasks, such as linguistic generalization, with much less data needed for language acquisition. (Beinborn and Hollenstein, 2024).

Our findings reveal that while CL offers a promising paradigm for language model pretraining, its direct application for ordering textual data, based on metrics developed for English, did not yield statistically significant performance gains for Slovak. This suggests a crucial divergence in how CL principles manifest across languages with differing morphological complexities.

2 Background

One of the main reasons for conducting the research is the difference in complexity between Slovak and English. The different complexities of the languages are inferred from their morphology, syntax, and semantics.

Morphology: The Slovak language is an inflection-based language that also has a higher number of consonants and a higher average number of morphemes in inflection. The changing word bases and modification of words are used to express different grammatical categories. In English, the focus is on derivational morphology, where specific suffixes form the word type (-ed, -ing); on the other hand, the Slovak language relies more on inflectional morphology to convey grammatical relationships (Panocová, 2021).

Syntax: English needs more words to form grammatical categories, but on the other hand they are based on a fixed word order. A sentence in Slovak (Adam mal'uje stenu- Adam paints the wall) can be expressed by switching the object and subject, but in English, the word order changes, but also words are added (Stenu mal'uje Adam-The Wall is painted by Adam). (Newerkla, 2010).

[1]Dataset: https://huggingface.co/datasets/ubokri/SlovakBabyLM.

[2]Code: https://github.com/baucek/Slovakbabylm/tree/main

Proceedings of the First BabyLM Workshop, pages 301–312

Semantics: A comparison of Slovak and English words from Swadesh's list (the Swadesh list is a list of provisionally universal terms (Swadesh, 1955)) showed that out of the 346 words examined, a total of 66 Slovak words and 97 English words had more meanings. Again, the authors attribute these results to the analytical nature and the plurality of English. That is, the English language uses more of the same words in different contexts (Urbániková, 2010).

The consequences of different complexities can be seen in the tokenization of text. A study examining text tokenization in 108 languages found that the GPT-3.5 model required 2.13 times more tokens to tokenize Slovak text than it did to tokenize English text (Asprovska and Hunter, 2024). Therefore, we consider the differences between English and Slovak to be key in applying solutions to the Babylm problem and strive to utilize them in the application of CL methods.

3 Related work

Human learning is based on the gradual acquisition of knowledge (from simple to complex) that is necessary for the development of skills needed for life (Skinner, 1958; Piaget, 2000). CL methods apply human learning abilities to LM by training them in a predetermined sequence of data based on their complexity (Bengio et al., 2009). This involves selecting rules that determine the order of training data. Therefore, metrics for measuring text complexity have been developed, which are grouped into a linguistically oriented group and a frequency-oriented group previously tested in English language (Edman and Bylinina, 2023; Bunzeck and Zarrieß, 2023; Warstadt et al., 2023b).

In the English language, linguistic complexity metrics didn't cause improvement of models. A study focusing on morphological and lexical complexity of words (Type/Token Ratio, Punctuation density, Mean word length, mean and max rarity of words) proved to be unreliable for overcoming the random ordering of text (Edman and Bylinina, 2023). Another study by (Bunzeck and Zarrieß, 2023) employed similar complexity metrics to rank text, including average word length, utterance length (the count of lexical tokens in the sequence), and average word frequency. These proved to be unsuccessful against random text order.

In frequency complexity metrics, a study retrained GPT-2 in the BabyLM dataset, using the average frequency of sentences and semantic similarity to rank and remove text, thus eliminating weak sentences of high frequency but semantically, which improved the performance of the BLIMP task (Borazjanizadeh, 2023). Another study first sorted text by complexity (spoken = "easier", typed = "harder"). Then, metrics such as the frequency of the BPE token were applied to the ranking, but this approach did not improve the BLIMP results (Martinez et al., 2023).

As mentioned in the chapter on the differences between English and Slovak, Slovak is a more semantically, morphologically, and syntactically complex language than English. From which we can assume that the metrics of linguistic complexity or frequency complexity may have a greater impact on the ranking of text in Slovak.

4 Creation of dataset

For the experiments, we had to create several sub-datasets, namely, 6 specific sub-datasets that focus on different parts of the language. Several different methods of data mining and preprocessing were used. Section Data mining and Preprocessing contains basic preprocessing, which was applied to all sub-datasets. However, due to the individual differences of the sub-datasets, each sub-dataset has a subsection for its own preprocessing.

4.1 Data mining and Preprocessing sub-datasets

For data mining the Python programming language (version 3.11.6) was used, and the following libraries were used: Scrapy (version: 2.11.1)[3], beautifulsoup (version 4.12.3)[4], requests (version: 2.32.3)[5], and the regex library (version 2023.10.3)[6]. However, each sub-dataset and source of data had a different way of using each library. The Google search engine was used to mine individual web pages containing specific data (a sub-dataset of fairy tales). In some cases, we did not find a single source with a large enough amount of data; therefore, text generation was used and the OpenAI library (version 1.35.10)[7].

The base preprocessing procedure was taken from the creation of SlovakBERT (Pikuliak et al., 2021):

[3] https://www.scrapy.org/
[4] https://pypi.org/project/beautifulsoup4/
[5] https://pypi.org/project/requests/
[6] https://docs.python.org/3/library/re.html
[7] https://openai.com/

- URL and email addresses were replaced with special tokens
- Elongated punctuation, newline character, and whitespaces were reduced, i.e., if there were sequences of the same punctuation mark, these were reduced to one mark (e.g., — to -).
- The whole text was removed if the text contains signs of wrong format
- The text is also removed if it is in a different language (using the Python library langdetect (version: 1.0.9)[8])

4.2 Sub-dataset of articles

For this sub-dataset, the main source of data was https://sk.wikipedia.org/ and the Scrapy library for web-crawling. Wikipedia contains additional links to the term that appears in the text. Thanks to this structure, subpages describing different subjects taught in primary schools were obtained from topics that a child may encounter during the teaching process (can be found in Appendix A). We ran 10 crawlers (9 pages individually as an initial_page and then 10 times all pages together).

The scrapy spider was forbidden to crawl webpages with the string in url: ("Zoznam:", "Kategória:", "Kategórie:", "Diskusia:") to remove subpages containing different lists. We also excluded the subpage "Hlavná_stránka" so that the spider did not go to the main Wikipedia page and therefore did not scrape the same content several times. In addition to the main preprocessing, we made further changes to the dataset. For incoherent scraped sources, sources with fewer than 55 characters were excluded. We also deleted sentences with specific strings "Zdroj:" or "FILIT", because they often contained inconsistent text. The final count of the text obtained can be seen in Table 1.

4.3 Sub-dataset of dialogues

We used the website https://www.opensubtitles.org/sk to create this sub-dataset. AWS cloud services, namely Lambda and S3 bucket, were used for data mining (Amazon Web Services, 2025a,b). For simplicity of the webpage, a for loop over and web-scraping with beautifulsoup4 were applied. The last code execution was on 2024-07-26. On that date, there were 29 852 movies and series with Slovak subtitles. After removing copies, incorrectly processed files, subtitles in a foreign language, applying preprocessing and non-subtitle-related sentences (translator's name, advertisement, etc.), we were left with 23 789 text files with 888 313 765 words. Due to the large amount of data, we took files that are larger than half of the total subtitle file to fit into the required number of words. The final count can be seen in the table 5.

4.4 Sub-dataset of literature

We used 4 websites containing freely available books in the Slovak language (Appendix A). The resources of https://eknizky.sk/ and https://www.1000knih.sk/ were not available for download through our web crawling or web scraping tools, so we chose manual selection, randomly selecting books that were in Slovak and were not poetry. Before preprocessing, we removed duplicate books due to their common names. Applied common preprocessing, then removed lines that contained information about the book, such as author name, publisher, ISBN, and EAN. Non-suitable pages, which contain 4 or more dots in a row (table of content) and the first page of the book, due to irrelevant text to our pretraining. With the library pymupdf library (version 1.25.5)[9], we deleted the footer and header based on the different position in comparison with other paragraphs. The final count of obtained text can be seen in Table 4.

4.5 Sub-dataset of fairytales

In creating the sub-dataset with fairytales, we used 7 webpages and a random Google search dealing with well-known fairytale authors (Appendix A). The process of creating this sub-dataset was very versatile because we worked individually with each source. Also, due to the lack of specific text, we chose the text generation method. To create the fairytales, we used the GPT-4o language model and OpenAI Library. We created two prompts to generate the fairytales (Appendix B). The first prompt was used to create the topic of the fairytale, and the second prompt was used to create the fairytale itself. A similar method of generating child-centered text can be found in other studies (Valentini et al., 2023; Schepens et al., 2023). However, we consider as a huge negative to achieve a cognitive plausible text. We cannot compare the accuracy of the created text with the real vocabulary of children at a given age,

[8] https://pypi.org/project/langdetect/

[9] https://pymupdf.readthedocs.io/en/latest/

because there is no dataset containing children's vocabulary in Slovak. The final count of obtained text can be seen in Table 2.

4.6 Sub-dataset of educational content

For the creation of the sub-dataset with educational content, the website https://referaty.aktuality.sk/ was the main source, and beautifulsoup4 as a tool for web-scraping due to the limited number of webpages containing content. Another reason was the exclusion of foreign languages https://referaty.aktuality.sk/cudzie-jazyky, which could contain a mix of foreign languages and Slovak. Besides common preprocessing, we delete sentences with ISBN, or 'Použitá literatúra', which describe sources with non-relevant text. Due to the high word count, we discarded sources with fewer than 300 characters to reduce the number of possible badly scraped data and to reduce the number of sources. Final gained text can be seen in the Table 4.

4.7 Sub-dataset of children communication

Datasets of children's communication already exist in english: the Child Language Data Ex- change System (CHILDES) (MacWhinney, 1998). No such dataset has been created within the Slovak language, and we are not aware of a free existing data source[10]. Therefore, we decided to mechanically generate conversations between a child and a known person using LLM. We used the gpt-4o language model to generate the conversations.

A similar set of mechanically generated conversations in English has already been created. The dataset of variations contains only the mother's site of conversation, and an LLM is used to repeat and rephrase, change words or their order from the inserted sentence used in the conversation (Haga et al., 2024). However, to better focus on the cognitive aspect of the conversation between the familiar person (FP) and the child, we decided to include features of the conversation between the FP and the child (We refer to the FP as a person in the child's close social circle, with whom interaction occurs, such as a parent, sibling, or guardian). According to the Usage-Based Theory of Language Acquisition (Tomasello, 1992). Communication must take place in an activity or game, where the child first passively and then actively participates. In a given game or interaction with a child, a familiar person assists in the proper development of language by repeating mispronounced words, describing the environment, or basically interacting with the child (Rowe and Snow, 2020).

Therefore, we decided to create 2 prompts (Appendix B). The first prompt creates a situation in which a child interacts with the FP, and a conversation takes place between them, which will serve as a basis for creating a conversation. The second prompt adopted the given topic and other relevant information such as the average number of words spoken by the child during a specific age, of the child. Subsequently, during prompt engineering, the greeting was randomly removed to reduce repetitive greetings. Our resulting dataset consisted of 4 groups of conversations that were created based on the settings of the child's age and the number of words used in a sentence by the child at that age.

The created text was saved as a .json file. The first prompt's output was stored under the key 'url', and the second prompt's output under 'page' to keep it consistent with other datasets. Since only the FP part of the conversation is used for pretraining, we didn't worry about the child's language accuracy and removed all lines starting with 'D:'.

[10]In other Slavic languages, it has been created https://talkbank.org/childes/access/Slavic/

Webpages	Number webpages	Number of words
History	16 616	6 483 417
Music	12 570	4 835 216
Chemistry	8 552	3 219 831
Sport	6 588	2 763 223
Slovak language	5 281	2 170 375
Biology	3 414	1 396 678
Physics	1 741	723 209
Civics	866	688 245
Full Subjects	93 649	35 175 212
Total	**149 277**	**57 455 406**

Table 1: Overview of scraped Wikipedia pages

Sources	Number of sources	Number of words
sikovnamamina.sk	36	41 392
rozpravkozem.sk	697	303 295
zones.sk	1 058	1 359 908
rozpravky.online	87	43 636
readmio.com/sk	1 591	359 510
svetrozpravok.sk	70	38 773
zlatyfond.sme.sk	293	671 388
Downloaded books	30	509 365
Created fairytales	3 094	1 786 974
Total	**6 956**	**4 754 731**

Table 2: Overview of fairytale sources, number of items, and word counts

Age	Number of conversations	Number of words
2 years old	9 191	478 920
3 years old	7 764	477 217
4 years old	5 433	361 184
5 years old	7 688	415 297
Total	**30 076**	**1 732 618**

Table 3: Number of conversations and words in conversations by age group

Web pages	Number of books	Number of words
zones.sk	6	246 362
eknizky.sk	210	5 628 366
greenie.elist.sk	22	423 527
1000knih.sk	44	1 383 083
Total	**272**	**7 681 338**

Table 4: Overview of scraped book sources

Domain of Sub-Dataset	Strict (Words)	Sources	Strict-small (Words)
Child-directed speech	1.7 mil	Text generation 7 webpages	470 000
Fairytales	4.7 mil	Random books Text generation	910 000
Dialogues	53.6 mil	opensubtitles.org/sk	4 000 000
Educational content	14.9 mil	referaty.aktuality.sk	1 304 000
Wiki	22 mil	sk.wikipedia.org	2 300 000
Books	7.6 mil	4 webpages	990 000
Total	**104.5 mil**		**9 974 000**

Table 5: Overview of sub-dataset domains and number of words

5 Methods

5.1 Curricullum learning criteria

The CL metrics application will consist of some of the metrics in the English versions that have already been created in the BabyLM call, plus new metrics will be created that are related to the Slovak language. The evaluation was done within a single source (downloaded book or webpage) to not divide the context of the sentences.

Linguistic complexity:

Average word length: is calculated as the number of characters divided by the number of words in a given source. This metric indicates the lexical intensity of the resource, where longer words can reduce the readability of the text.

Syllable/word ratio: is calculated as the number of syllables divided by the number of words. A given metric indicates the morphological complexity of the source, where a higher proportion of syllables per word may indicate a more complex word structure of the text.

Punctuation density: is calculated as the number of punctuation marks divided by the number of words. A low number of punctuation marks and a high number of words can mean a complex sentence structure.

Conjunction ratio: is calculated as the number of conjunctions divided by the number of words. We used 59 non-bending one-word conjunctions (Appendix C). Conjunctions link sentence constructions, which can create large sentence constructions and thus increase the syntactic complexity of the sentence (Dvonč et al., 1966).

Preposition ratio: is calculated as the number of prepositions divided by the number of words. We used 44 initial prepositions (Appendix C). Prepositions have the task of forming relations with flexible word types such as nouns or adjectives. However, at the same time, prepositions determine the case of the word they stand in front of (Dvonč et al., 1966). Inflected nouns may contain more tokens than nouns in the base form, so their presence may increase the morphological complexity of words (Asprovska and Hunter, 2024).

Frequency complexity:

Prior to actual data sorting, we extracted the frequencies of individual tokens, words, and bi-grams by splitting the words using the .split() function and removing non-alphabetical signs where appropriate.

Average word frequency: is calculated as the average of the individual word frequencies divided by the number of words in a given resource.

Average token frequency: is calculated as the average of the individual token frequencies divided by the number of words in a given resource.

Average bi-gram frequency: is calculated as the average of the individual bigram frequencies divided by the number of words in a given source.

In order to properly measure the given metrics (lower rankings == simpler sentences), we had to rescale the frequency group metrics and punctuation density metric (metric=-1*metric). The metrics were normalized using min-max normalization across the entire dataset. The order is defined as the sum of normalized values. If the experiment required the ranking of sub-datasets, the points were sum for each sub-dataset, and the ranking of sub-datasets within a single metric was determined (1 = simplest according to the metric and 6 = most complex according to the metric).

5.2 Architecture and Pretraining

The architecture of the models was based on the results of the BabyLM challenge (Warstadt et al., 2023b). Hyper-optimization of the hyperparameters prove a 1:2 ratio between feed-forward layers and attention heads as the best (Proskurina et al., 2023). Therefore, our models had 6 layers and 12 attention heads. For training, we used a sequence length of 128 tokens and a batch size of 128, both identified as optimal parameters in prior research (Cagatan, 2023; Proskurina et al., 2023). We applied a 15% masking rate across 7 training epochs, following established best practices (Cagatan, 2023). The given studies used a Vocabulary size of 40000 and 30000 (Edman and Bylinina, 2023; Opper et al., 2023), but to handle Slovak's complex morphology, we set a larger vocabulary of 60,000 tokens with byte-level BPE tokenization. To pre-train and test models, we used two graphics cards NVIDIA GeForce RTX 3090 and NVIDIA GeForce GTX 1080.

5.3 Testing parameters

Within the Babylm challenge (Warstadt et al., 2023a), researchers used several evaluation tasks. Hence, we will use sentiment analysis (SA), question answering (QA) tasks. The SA task focuses on identifying specific word combinations. The dataset `dgurgurov/slovak_sa` (Pecar et al.,

2019) was selected for the SA task and 4 performance measurements: Accuracy, Precision, Recall, and F1-score. The QA task tests the model's ability to process and generate answers by analyzing word relationships (Farea et al., 2022). Dataset `TUKE-DeutscheTelekom/squad` (Hládek et al., 2023) was selected, and the F1 score and exact match were selected as performance measurement. The models were fine-tuned for specific tasks, and each model was tested 9 times with 2 different hyperparameters (3 different learning rates[11] and 3 different epochs[12]) and then performed statistical significance testing to evaluate performance differences between model variations and the model without improvements. T-test was used for each of the selected evaluation metrics.

6 Results

For the purpose of testing CL metrics in practice, 7 LMs were created with Strict-small dataset. 5 LMs can be divided into two groups according to the criteria: sorting and specific CL metric groups. The other 2 LMs had selected data from strict track into strict-small track based on complexity.

Application of specific ordering:

1. Without ordering, without specific group metrics
2. Sub dataset ordering, both metric groups
3. Full ordering, both metric groups

Application of group metrics:

4. Full ordering, only language group metric
5. Full ordering, only frequency group metric

6.1 Application of CL methods

From the results in Tables 6 and 7, we can conclude that no group of metrics is significant for model improvement. From average of all trials (Appendix D.1), we can observe better performance of the linguistic group against the frequency group. This may indicate a potentially higher relevance of language-based features versus frequency-based features. It can be interpreted as the richness of the Slovak language, where a larger number of tokens is created. On the other hand, Individual linguistic or frequency groups of metrics can influence LM differently. Specifically, CL improvement in the SA task is based on the gradual increase in the variation of nouns and verbs as significant factors (Elgaar and Amiri, 2023).

Comparison	t-value	p-value
3. Both groups of metrics		
Accuracy	-1.59	0.924
Recall	-1.59	0.924
F1	-1.39	0.899
Precision	-1.38	0.897
4. Language group		
Accuracy	-1.00	0.827
Recall	-1.00	0.827
F1	-0.47	0.675
Precision	-0.16	0.562
5. Frequency Group		
Accuracy	-1.70	0.936
Recall	-1.70	0.936
F1	-1.35	0.894
Precision	-1.25	0.876

Table 6: Statistical significance (p>= 0.05) of models using CL methods compared to the model without CL methods in SA task

Comparison	t-value	p-value
3. Sum + Frequency (grammar)		
Exact Match	-0.53	0.694
F1	-1.24	0.875
4. Grammar Group		
Exact Match	0.04	0.484
F1	-0.54	0.699
5. Frequency Group		
Exact Match	-2.93	0.990
F1	-1.18	0.864

Table 7: Statistical significance (p>= 0.05) of models with CL methods and the model without CL methods in QA task

6.2 Text ordering methods

Models with different text ordering did not prove to be significantly better than the model without any application of the CL metrics and spcific ordering (Tables 8 and 9). The ordering of the sorted sub-datasets shows worse performance in QA performance than the model with the ordering of full data based on F1 score and exact match, and in turn, the ordering of full data performs worse in SA tasks based on F1 score, precision, accuracy, and loss (Appendix D.2). Results suggest the possibility that data ordering may affect context handling performance. In Malkin et al. (2021) demonstrate the absence of coherence and logic as a negative factor to handle longer-term dependencies between sentences and effective context work. As for SA tasks, this can be explained by the effect of variations in nouns and verbs, which by using metrics and ranking the whole dataset can effect this ranking (Elgaar and Amiri, 2023).

[11] learning rate= [5e-5,3e-5,1e-5]
[12] epochs = [5,7,10]

Comparison	t-value	p-value
2. ordering of sub-datasets		
Exact Match	-0.05	0.518
F1	-0.50	0.684
3. ordering full data		
Exact Match	-0.53	0.694
F1	-1.24	0.875

Table 8: Statistical significance (p>= 0.05) of models with specific application of CL methods and the model without CL methods in QA task

Metric	t-value	p-value
2. ordering of sub-datasets		
Accuracy	-2.54	0.983
Recall	-2.54	0.983
F1 Score	-2.28	0.974
Precision	-2.39	0.978
3. ordering full data		
Accuracy	-1.59	0.924
Recall	-1.59	0.924
F1 Score	-1.39	0.899
Precision	-1.38	0.897

Table 9: Statistical significance (p>= 0.05) of models with specific application of CL methods and the model without CL methods in SA task

6.3 Metrics as preprocessing methods

The following results show that using CL metrics for preprocessing has the highest effect among the applied improvements. The application of the hardest complexity on the QA task show significant improvement by F1 score (Table 10) and the simplest texts for pretraining the model on the SA task (Table 11). After scanning the sources used for pretraining, we can infer that resources contain long words or disjointed text, which can be reduced by selecting the simplest sources. Therefore model pretrained on text with the hardest complexity was better in QA task, and the model pretrained on text with the simplest complexity was better in the SA task. This again confirms the relationship between specific CL methods and performance improvement in specific tasks. (Elgaar and Amiri, 2023).

Metric	t-value	p-value
1. The simplest complexity		
Exact Match	-2.35	0.976
F1	-4.48	0.684
1. The hardest complexity		
Exact Match	0.76	0.233
F1	2.19	*0.030

Table 10: Statistical significance (p>= 0.05) of models with selection of the text based complexity and the model without any improvements in the QA task

Metric	t-value	p-value
1. The simplest complexity		
Accuracy	1.57	0.077
Recall	1.57	0.077
F1 Score	1.79	0.056
Precision	1.81	*0.054
1. The hardest complexity		
Accuracy	0.78	0.228
Recall	0.78	0.228
F1 Score	0.12	0.454
Precision	-0.08	0.532

Table 11: Statistical significance (p >= 0.05) of models with selection of the text based complexity and the model without any improvements in the SA task

7 Conclusion

The aim is to establish a cornerstone in the research of cognitively inspired models in the Slovak language and to point out the possibilities of applying CL to LRLs such as the Slovak language. In pursuit of this goal, a Slovak version of the BabyLM challenge (Warstadt et al., 2023a) was created. The constructed experiments demonstrated several findings. They confirmed the results of studies in English, where both sets of metrics showed no significant improvement in QA and SA tasks (Martinez et al., 2023; Bunzeck and Zarrieß, 2023; Edman and Bylinina, 2023).

CL metrics as preprocessing methods shows significant improvement against random order of text. These results can lead to use linguistic and frequency CL metrics as a potential optimization text. Similar usage of CL can be seen in data selection for abstractive text summarization (Sun et al., 2023). Applied CL identify high-value training examples, demonstrating that targeted data selection can improve model performance compared to training on the full dataset. The subtle differences noted, often visible only in decimal places (Appendix D), emphasize the challenge of discerning the impact of CL strategies when baseline performance is already competitive or when the dataset size limits the magnitude of observable improvements.

These nuances could be indicative of slight shifts in the model's learned representations, even if not leading to a statistically superior overall score. Additionally, the positive or negative effect of CL metrics in Slovak was much less significant than in English. According to (Bengio et al., 2009), CL metrics need to gradually increase the amount of more useful information with increasing learning time, which may be more complicated in Slovak language due to its linguistic complexity, where

the order produced by the metrics may be more difficult to determine compared to English. The greater semantic complexity of the Slovak language can lead to a higher number of tokens (Asprovska and Hunter, 2024). This factor can lead to subtle frequency variations that may not provide sufficiently clear signals for learning within a robust curriculum. What could have led to the failure of frequency methods.

Limitations

One of the fundamental limitations of the created dataset and results was the mechanical generation of data. This is not valid from the perspective of adapting to human learning. At the same time, LRLs may also lack dictionaries of children's speech, such as the Slovak language, which can serve as a test of the generated data or an aid to their creation (Schepens et al., 2023; Haga et al., 2024) or verification for the developmentally plausible of created text. Thus, it is not only a limitation but also a call to action for researchers to invest in improving Slovak language resources and model support, and also research in multilingual research in cognitive-inspired language models.

Within the limitations of computational resources, we could not focus on the semantic or syntactic part of the language. The results of studies with cosine similarity as an evaluator for CL metrics show that the factor would improve the evaluation, and hence we would get better results (Han and Myaeng, 2017; Borazjanizadeh, 2023)). For instance, due to the computational complexity of metrics, we could not perform the Part of Speech evaluation or more deeper analysis of text.

Acknowledgements

This work has been partially supported by grant APVV-21-0114. Additionally, we would like to thank our affiliations Kempelen Institute of Intelligent Technologies for providing technical advice, and the Faculty of Mathematics, Physics, and Informatics, for providing the computing resources used to pre-train our language models.

References

Amazon Web Services. 2025a. Amazon S3.

Amazon Web Services. 2025b. AWS Lambda. Accessed: 2025-04-12.

Marijana Asprovska and Nathan Hunter. 2024. The tokenization problem: Understanding generative ai's computational language bias. *Ubiquity Proceedings*, 4(1).

Lisa Beinborn and Nora Hollenstein. 2024. *Cognitive plausibility in natural language processing*. Springer.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Nasim Borazjanizadeh. 2023. Optimizing gpt-2 pretraining on babylm corpus with difficulty-based sentence reordering. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 356–365.

Bastian Bunzeck and Sina Zarrieß. 2023. Gpt-wee: How small can a small language model really get? In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 35–46.

Omer Veysel Cagatan. 2023. Toddlerberta: Exploiting babyberta for grammar learning and language understanding. *arXiv preprint arXiv:2308.16336*.

Ladislav Dvonč, Jozef Ružička, et al. 1966. Morfológia slovenského jazyka. *(No Title)*.

Lukas Edman and Lisa Bylinina. 2023. Too much information: Keeping training simple for babylms. *arXiv preprint arXiv:2311.01955*.

Mohamed Elgaar and Hadi Amiri. 2023. Ling-cl: Understanding nlp models through linguistic curricula. *arXiv preprint arXiv:2310.20121*.

Amer Farea, Zhen Yang, Kien Duong, Nadeesha Perera, and Frank Emmert-Streib. 2022. Evaluation of question answering systems: complexity of judging a natural language. *arXiv preprint arXiv:2209.12617*.

Akari Haga, Akiyo Fukatsu, Miyu Oba, Arianna Bisazza, and Yohei Oseki. 2024. Babylm challenge: Exploring the effect of variation sets on language model training efficiency. *arXiv preprint arXiv:2411.09587*.

Sanggyu Han and Sung-Hyon Myaeng. 2017. Tree-structured curriculum learning based on semantic similarity of text. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 971–976. IEEE.

Daniel Hládek, Ján Staš, Jozef Juhár, and Tomáš Koctúr. 2023. Slovak dataset for multilingual question answering. *IEEE Access*, 11:32869–32881.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, pages 6282–6293. Association for Computational Linguistics.

Brian MacWhinney. 1998. The childes system. *Handbook of child language acquisition*, pages 457–494.

Nikolay Malkin, Zhen Wang, and Nebojsa Jojic. 2021. Coherence boosting: When your pretrained language model is not paying enough attention. *arXiv preprint arXiv:2110.08294*.

Richard Diehl Martinez, Zebulon Goriely, Hope McGovern, Christopher Davis, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2023. Climb: Curriculum learning for infant-inspired model building. *arXiv preprint arXiv:2311.08886*.

Stefan-Michael Newerkla. 2010. Juraj dolník, všeobecná jazykoveda. opis a vysvetl'ovanie jazyka, bratislava (veda, vydavatel'stvo slovenskej akadémie vied) 2009, 376 s.

Mattia Opper, J Morrison, and N Siddharth. 2023. On the effect of curriculum learning with developmental data for grammar acquisition. *arXiv preprint arXiv:2311.00128*.

Renáta Panocová. 2021. Basic concepts of morphology i. *Košice: Vydavatel'stvo ŠafárikPress*.

Samuel Pecar, Marián Šimko, and Maria Bielikova. 2019. Improving sentiment classification in slovak language. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 114–119.

Jean Piaget. 2000. Piaget's theory of cognitive development. *Childhood cognitive development: The essential readings*, 2(7):33–47.

Matúš Pikuliak, Štefan Grivalskỳ, Martin Konôpka, Miroslav Blšták, Martin Tamajka, Viktor Bachratỳ, Marián Šimko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. 2021. Slovakbert: Slovak masked language model. *arXiv preprint arXiv:2109.15254*.

Irina Proskurina, Guillaume Metzler, and Julien Velcin. 2023. Mini minds: Exploring bebeshka and zlata baby models. *arXiv preprint arXiv:2311.03216*.

Meredith L Rowe and Catherine E Snow. 2020. Analyzing input quality along three dimensions: interactive, linguistic, and conceptual. *Journal of child language*, 47(1).

Job Schepens, Nicole Marx, and Benjamin Gagl. 2023. Can we utilize large language models (llms) to generate useful linguistic corpora? a case study of the word frequency effect in young german readers. *Preprint from PsyArXiv https://doi. org/10.31234/osf. io/gm9b6*.

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

Shichao Sun, Ruifeng Yuan, Jianfei He, Ziqiang Cao, Wenjie Li, and Xiaohua Jia. 2023. Data selection curriculum for abstractive text summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7990–7995.

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137.

Michael Tomasello. 1992. *First verbs: A case study of early grammatical development*. Cambridge University Press.

Milica Urbániková. 2010. Lexical and semantic development of the basic vocabulary in english and slovak.

Maria Valentini, Jennifer Weber, Jesus Salcido, Téa Wright, Eliana Colunga, and Katharina Kann. 2023. On the automatic generation and simplification of children's stories. *arXiv preprint arXiv:2310.18502*.

Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. 2023a. Call for papers–the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2301.11796*.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, et al. 2023b. Findings of the babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*.

A Appendix: Web-pages for sub-datasets

Sub-dataset of articles: https://sk.wikipedia.org/wiki/ObÄĎianska_nÃąuka,https://sk.wikipedia.org/wiki/Dejiny, https://sk.wikipedia.org/wiki/Sloven%C4%8Dina, https://sk.wikipedia.org/wiki/Fyzika, https://sk.wikipedia.org/wiki/Matematika, https://sk.wikipedia.org/wiki/Informatika, https://sk.wikipedia.org/wiki/Ch%C3%A9mia, https://sk.wikipedia.org/wiki/Biol%C3%B3gia, https://sk.wikipedia.org/wiki/Hudba, https://sk.wikipedia.org/wiki/%C5%A0port Sub-dataset of literature: https://www.zones.sk/, https://eknizky.sk/, https://greenie.elist.sk/, https://www.1000knih.sk/ Sub-dataset of fairytales: https://www.sikovnamamina.sk/, https://www.rozpravkozem.sk/, https://www.zones.sk/studentske-prace/rozpravky/, https://rozpravky.online/, https://zlatyfond.sme.sk/, https:

//www.readmio.com/sk/uvod, https://svetrozpravok.sk/

B Appendix: Prompts for text generation

Children's books:
First prompt:
count = 200
"role": "user", "content": f"Vytvor mi {count} názvov rozprávok. Vráť len zoznam názvov rozprávok bez ďalšieho úvodu, číslovania alebo záverečných poznámok. Témy sa nesmú opakovať.", "role": "assistant", "content": "Si spisovateľ detských rozprávok."

Second promt: "role": "user", "content": f"Vytvor rozprávku pre deti na tému:{topic}. Snaž sa využiť maximálny počet tokenov.", "role": "assistant", "content": "Si spisovateľ detských rozprávok."

Child-directed speech:
count = 200
family = 'Matkou'
age_word = 'dvojročným'
age = 2
average_number_words = 2
First prompt: "role": "user", "content": f"""Vytvor {count} situácii medzi {family} a dieťaťom, ktoré sa môžu vyskytnúť medzi {family } a {age_word } dieťaťom. Výsledok budú len dané situácie a nebudú sa opakovať. Príklad: Prebaľovanie. Dieťa nesmie byť súčasťou činnosti, ktorú je nemožné vykonať v danom roku života ({age } roky). """, "role": "assistant", "content": "Si schopný posúdiť lingvistické a motorické prejavy dieťaťa v rôznom veku."
Second prompt: "role": "user", "content": F"""Daj mi konverzáciu medzi {age_word} dieťaťom a {family } na tému:{topic}. Tvoj výsledok musí obsahovať len vytvorený dialóg, kde {family } bude označená ako {family[0].upper() }: a dieťa ako D. Správna komunikácia zo strany {family}: Komentovanie: Je dôležité opisovať, to čo sa deje v okolí. Opakovanie: Zdôrazňovať a opakovať veci, ktorým dieťa nerozumie a poskytnúť možností na neustále opakovanie nových slov alebo viet Výslovnosť: Použi slová gramaticky správne. !!!!! Prispôsobivosť: family musi prispôsobiť reč aktuálnym záujmom a potrebám dieťaťa: vety sú krátke!!!!! Priemerný počet slov vo vete u dieťaťa: {average_number_words}. Nezačínaj komunkáciu pozdravom""", "role": "assistant", "content": "Si schopný posúdiť lingvistické a motorické prejavy dieťaťa v rôznom veku."

C Appendix: List of conjunctions and prepositions

Conjunctions = (a, že, i, keby, aby, aj, ak, keď, keďže, ako, akoby, hoci, ale, alebo, lebo, ani, iba, tak, takže, teda, totižto, veď, však, žeby avšak, až, ba, bár, beztak, buď, by, či, čím, čoby, pričom čiže, čo, kým, leda, ledva, len, len čo, totiž, lenže, najprv, nech, než, nielen, no, nuž, pokiaľ, pokým, predsa, preto, pretože, síce, sotva, sťa, prí padne, poprípade, eventuálne)

Prepositions = (bez, cez, do, k, medzi, na, nad, o, od, okrem, po, pod, pre, pred, pri, proti, s, skrz, u, v, z, za, ponad, popod, popred, poza, popri, pomedzi, znad, spred, zmedzi, spod, spopred, sponad, spopod, spoza, spopri, spomedzi, zo, ku, voči, skrze, vo, so)

D Appendix: Mean of results

D.1 Application of CL methods

Models	Exact Match (%)	F1 Score (%)
1. Without ord.	1.39 (0.38)	**6.59 (0.65)**
3. Both groups	1.33 (0.29)	6.44 (0.75)
4. Language group	**1.39 (0.37)**	6.45 (1.20)
5. Frequency group	1.22 (0.37)	6.45 (0.74)

Table 12: Results of QA task for Text ordering methods (mean ± std).

Metric	1. Without ord.	3. Both groups	4. Lang. group	5. Freq. group
Loss	**0.32 (0.11)**	0.32 (0.11)	0.32 (0.10)	0.34 (0.10)
Acc (%)	**90.38 (2.65)**	90.16 (2.53)	90.23 (2.59)	89.66 (2.34)
Prec (%)	**85.45 (8.13)**	85.14 (7.90)	85.41 (8.10)	83.51 (7.81)
Rec (%)	**90.38 (2.65)**	90.16 (2.53)	90.23 (2.59)	89.66 (2.34)
F1 (%)	**87.74 (5.54)**	87.43 (5.34)	87.66 (5.49)	86.36 (5.23)

Table 13: Results of SA task for application of CL metrics (mean ± std).

D.2 Text ordering methods

Models	Exact Match (%)	F1 Score (%)
1. without ordering	**1.38 (0.38)**	**6.59 (0.65)**
2. ordering of sub-datasets	1.38 (0.41)	6.54 (0.60)
3. ordering full data	1.33 (0.29)	6.44 (0.75)

Table 14: Results of QA task for Application of CL methods: (mean ± std).

Metric	1. Without ord.	2. Sub-datasets	3. Full data
Loss	**0.32 (0.11)**	0.34 (0.10)	0.32 (0.11)
Acc (%)	**90.38 (2.65)**	89.74 (2.18)	90.16 (2.53)
Prec (%)	**85.45 (8.13)**	84.56 (7.36)	85.14 (7.90)
Rec (%)	**90.38 (2.65)**	89.74 (2.18)	90.16 (2.53)
F1 (%)	**87.74 (5.54)**	86.88 (4.85)	87.43 (5.34)

Table 15: Results of SA task for Application of CL methods: (mean ± std).

D.3 Metrics as preprocessing methods

Models	Exact Match (%)	F1 Score (%)
1. Random complexity	1.39 (0.38)	6.59 (0.65)
1. The simplest complexity	**1.60 (0.25)**	**6.95 (0.70)**
1. The hardest complexity	1.28 (0.36)	6.07 (1.25)

Table 16: Results of QA task for metrics as preprocessing methods (mean ± std).

Metric	1. Random	1. The simplest	1. The hardest
Loss	**0.32 (0.11)**	0.33 (0.10)	0.32 (0.10)
Acc (%)	**90.38 (2.65)**	90.05 (2.43)	90.32 (2.61)
Prec (%)	85.45 (8.13)	85.05 (7.76)	**85.46 (8.13)**
Rec (%)	**90.38 (2.65)**	90.05 (2.43)	90.32 (2.61)
F1 (%)	**87.74 (5.54)**	87.36 (5.20)	87.73 (5.53)

Table 17: Results of SA task for metrics as preprocessing methods (mean ± std).

E Appendix: Model settings for pretraining

BertConfig
vocab_size = 60000
hidden_size = 84
num_hidden_layers = 6
num_attention_heads = 12
intermediate_size = 1446
hidden_dropout_prob = 0.15
attention_probs_dropout_prob = 0.3
hidden_act = "gelu_new"
TrainingArguments
num_train_epochs = 7
per_device_train_batch_size = 32
per_device_eval_batch_size = 32
evaluation_strategy = "steps"
eval_steps = 1000
save_steps = 1000
logging_steps = 100
load_best_model_at_end = True
metric_for_best_model = "eval_loss"
bf16 = True
Finetuning of SA and QA
weight_decay=0.01,
per_device_train_batch_size=16,
per_device_eval_batch_size=16,
save_strategy="epoch",
evaluation_strategy="epoch",
Specific QA parameters
n_best = 20
max_answer_length = 50