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

Ľuboš Kriš, Marek Šuppa

EMNLP BabyLM Workshop 2025

Kempelen Institute of Intelligent Technologies

FMFI, Comenius university, Bratislava
KINIT, Bratislava

I

Context & Motivation

- The Slavic language family has far fewer resources for creating large language models (LLMs) compared to high-resource languages like English
- As model size (number of parameters) grows, so does the demand for training data, which poses challenges for low-resource languages (LRLs) such as Slovak

Conclusion:

It is necessary to focus on the application of CL methods to improve the pre-training of language models.

II

English vs Slovak [3]

- Slovak language is based on the inflection of words. Case inflections are absent in English
- English relies more on word order and auxiliary verbs for expressing grammatical relationships

Syntax differences

	Slovak language	English language
Active	Ocko mi kúpil psa	Dad bought me a dog
Passive	Psa mi kúpil ocko	Dog was bought by dad

Cross-linguistic differences in language acquisition

- Mistakes based on verb movement and case properties
- Vocabulary acquisition

III

Creation of dataset

- For research purposes was neccessary to create new dataset which copy BabyLM dataset in english
- Six sub-datasets were created, each focusing on different aspects of language.
- General preprocessing followed by SlovakBERT procedure.
- Used various data mining and additional preprocessing methods adapted per sub-dataset.

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**

IV

Methods

- 7 language models (LMs) trained on the Strict-small dataset.

Divided into :

- 5 models: tested sorting and specific CL metric groups.
- 2 models: selected data from *strict* → *strict-small* track based on CL 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

Curricullum learning

- Combination of existing English CL metrics with new Slovak-specific metrics

Linguistic complexity

- Average word length
- Syllable/word ratio
- Punctuation density
- Conjunction ratio

Frequency complexity

- Average word frequency
- Average bi-gram frequency
- Average token frequency

Fine-tuning and Evaluation

- Fine-tuned overall 9 times (combination of 3 learning rates and 3 epochs)

2 tasks:

Question Answering (QA):
Dataset: TUKE-DeutscheTelekom/squad
Metrics: F1-score and Exact Match (EM)

Sentiment Analysis (SA):
Dataset: dgurgurov/slovak_sa
Metrics: Accuracy, Precision, Recall, F1-score

Architecture, pretraining, evaluation

- For experiment purposes was used strict-small dataset and Bert architecture
- 6 FFN and 12 attention heads (Proskurina et al., 2023).
- Sequence length (tokens) and batch size: 128
- 15% masking rate across 7 epochs (Cagatan, 2023).
- BPE tokenizer with 60,000 vocabulary of tokens.

V

Results

- CL techniques not significantly improve performance but:

Text Ordering Methods (2,3)

- Better performance of the linguistic group against the frequency group
- Indicate a potentially higher relevance of language-based features versus frequency-based features.

Application of CL methods (3,4,5)

- The ordering of the sorted sub-datasets shows worse performance
- The ordering of full data performs worse in SA tasks

Metrics as preprocessing methods (1)

- The application of the hardest complexity on the QA task show significant improvement by F1 score and the simplest texts for pretraining the model on the SA task

NOTE: numbers are type of model which where used to compare

VI

Conclusion

- Establish a foundation for cognitively inspired models in Slovak and explore Curriculum Learning (CL) for low-resource languages (LRLs).
- CL helps identify high-value training examples, improving performance over full-dataset training.
- CL methods are less effective in the Slovak language

VII

Open problems

Low-resource languages (LRLs) often lack corpora of child speech or word dictionaries. Such resources are crucial for testing and verifying generated data.

VIII

Let's Stay in Touch!

Ľuboš Kriš | AI research engineer
Kempelen Institute of Intelligent Technologies
lubos.kris@kinit.sk | www.kinit.sk