BabyLMs for isiXhosa: Data-Efficient Language Modelling in a Low-Resource Context

Alexis Matzopoulos **Charl Hendriks** **Hishaam Mahomed** **Francois Meyer**
Department of Computer Science
University of Cape Town
{mtzale001, hmhis005, hndcha033}@myuct.ac.za, francois.meyer@uct.ac.za

Abstract

The BabyLM challenge called on participants to develop sample-efficient language models. Submissions were pretrained on a fixed English corpus, limited to the amount of words children are exposed to in development (<100m). The challenge produced new architectures for data-efficient language modelling, which outperformed models trained on trillions of words. This is promising for low-resource languages, where available corpora are limited to much less than 100m words. In this paper, we explore the potential of BabyLMs for low-resource languages, using the isiXhosa language as a case study. We pretrain two BabyLM architectures, ELC-BERT and MLSM, on an isiXhosa corpus. They outperform a vanilla pretrained model on POS tagging and NER, achieving notable gains (+3.2 F1) for the latter. In some instances, the BabyLMs even outperform XLM-R. Our findings show that data-efficient models are viable for low-resource languages, but highlight the continued importance, and lack of, high-quality pretraining data. Finally, we visually analyse how BabyLM architectures encode isiXhosa.

1 Introduction

Large language models (LLMs) are trained on trillions of words (Touvron et al., 2023). Humans are much more efficient language learners – children are exposed to less than 100 million words of speech/text by age 13 (Gilkerson et al., 2017). This mismatch motivated the establishment of the BabyLM challenge (Warstadt et al., 2023), a shared task in which participants were invited to propose data-efficient language modelling techniques. Submissions were pretrained on a fixed corpus of developmentally plausible English (e.g. child-directed speech, educational content) and ranked according to performance on natural language understanding (NLU) benchmarks.

The top submissions comfortably outperformed standard Transformer-based (Vaswani et al., 2023) models pretrained on the same fixed corpus, even surpassing state-of-the-art pretrained language models (PLMs) trained on orders of magnitude more data. The main aims of the BabyLM challenge was to build cognitively plausible models of language acquisition and enable compute-limited language modelling research (Warstadt et al., 2023). In this paper, we investigate an additional opportunity arising from the shared task: its potential to improve LMs for low-resource languages.

Model	POS	NER	NTC
Pretrained on 13m isiXhosa words			
RoBERTa	$87.0_{\pm 0.1}$	$85.4_{\pm 0.4}$	$\mathbf{97.6}_{\pm 0.5}$
MLSM	$87.4_{\pm 0.1}$	$87.0_{\pm 0.4}$	$95.4_{\pm 0.2}$
ELC-BERT	$\mathbf{87.7}_{\pm 0.5}$	$\mathbf{88.6}_{\pm 0.6}$	$95.0_{\pm 0.3}$
Massively multilingual pretraining			
XLM-R	88.1	88.1	89.2
Afro-XLMR	**88.7**	89.9	97.2
Nguni-XLMR	88.3	**90.4**	**98.2**

Table 1: BabyLM performance on isiXhosa tasks, compared to a RoBERTa baseline trained from scratch and three large-scale multilingual PLMs. We **boldface** best per-category performance and underline best overall.

BabyLMs aim to optimise performance on a limited training budget. For the BabyLM challenge, this was simulated by creating a constrained English corpus. For low-resource languages, such constraints represent the reality of their NLP resources. Most languages do not have publicly available corpora consisting of trillions of words, so out of necessity they operate on a limited training budget. The data-efficiency of BabyLMs therefore presents a promising opportunity to achieve real-world performance gains for certain languages.

To investigate BabyLMs in a low-resource context we turn to isiXhosa, a South African language with over 22 million speakers (Eberhard et al.,

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 266–274
January 20, 2025.

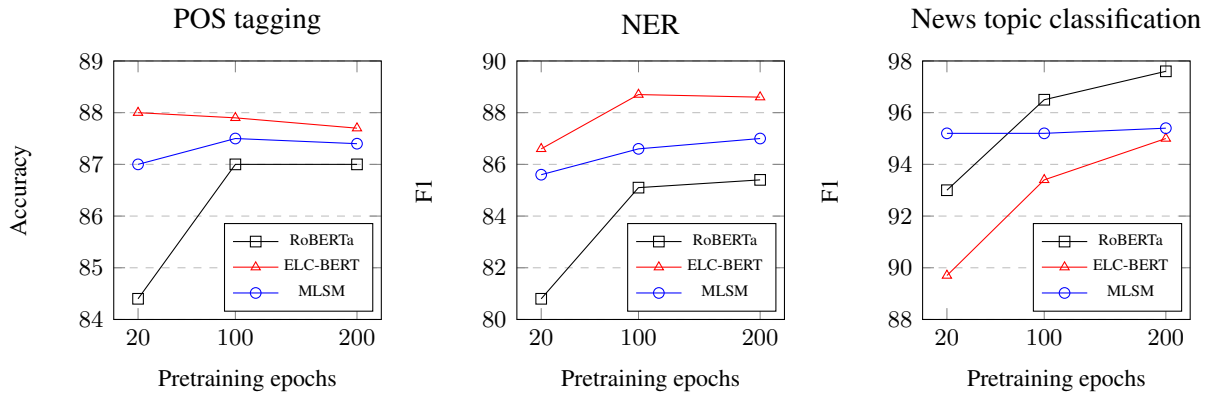


Figure 1: Downstream task performance for model checkpoints at different stages of pretraining.

2019). We pretrain two of the top BabyLM submissions, Every Layer Counts BERT (ELC-BERT) (Georges Gabriel Charpentier and Samuel, 2023) and Masked Latent Semantic Modeling (MLSM) (Berend, 2023b), for isiXhosa. We evaluate on isiXhosa NLU tasks and compare performance to a baseline RoBERTa architecture (Liu et al., 2019) pretrained on the same isiXhosa corpus.

Our results confirm the potential of data-efficient architectures for low-resource languages, with both BabyLMs obtaining performance gains over the RoBERTa baseline on POS tagging and NER. ELC-BERT proves especially effective, even rivalling one of our *skylines* (large-scale existing PLMs for isiXhosa). Unlike in the BabyLM challenge, our models do not outperform the best skylines, which we attribute to a lack of developmentally plausible data for isiXhosa. In summary, while our results indicate that low-resource gains are available from architectural innovations, they also highlight the continued need to develop higher-quality datasets for low-resource languages.

2 Background

2.1 PLMs for isiXhosa

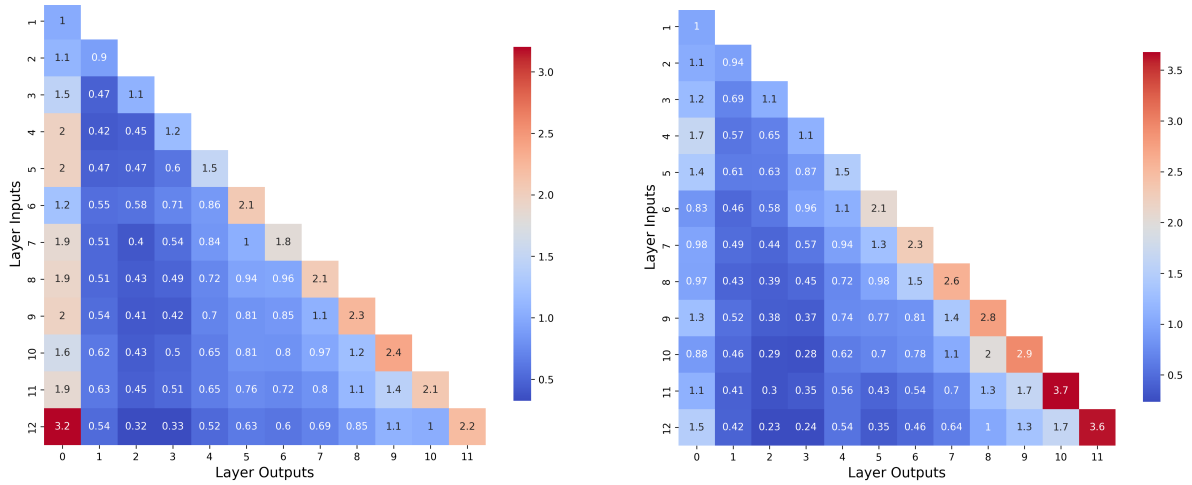
Pretraining corpora for isiXhosa are limited to 20m words (Xue et al., 2021). This is greater availability than most languages, but still two orders of magnitude less than even early PLMs (Devlin et al., 2019). As for other low-resource languages, multilingual modelling has improved performance for isiXhosa NLU. IsiXhosa is included in **XLM-R** (Conneau et al., 2020), a masked language model (MLM) pretrained on 100 languages. Two previous works improved performance for isiXhosa by adapting XLM-R through continued pretraining. **Afro-XLMR** (Alabi et al., 2022) adapts XLM-R for 23 African languages, including isiXhosa. **Nguni-XLMR** (Meyer et al., 2024) narrows the linguistic scope by adapting XLM-R for the four Nguni languages (isiXhosa, isiZulu, isiNdebele, Siswati), the closest linguistic relatives of isiXhosa.

2.2 BabyLM Architectures

The BabyLM challenge hosted three competition tracks, corresponding to different data restrictions. The *Small* and *Strict-Small* tracks were respectively limited to 100m and 10m words for pretraining, while the *Loose* track allowed non-linguistic data. ELC-BERT (Georges Gabriel Charpentier and Samuel, 2023) won both the *Small* and *Strict-Small* tracks, outperforming skyline models Llama2 (Touvron et al., 2023) and RoBERTa-base (Liu et al., 2019). MLSM (Berend, 2023a) was runner-up in the *Strict-Small* track. The *Strict-Small* data restriction (10m words) most closely aligns with the size of publicly available corpora for isiXhosa, which is why we chose the top models from this category.

2.2.1 Every Layer Counts BERT (ELC-BERT)

ELC-BERT (Georges Gabriel Charpentier and Samuel, 2023) adapts LTG-BERT (Samuel et al., 2023), an architecture designed to optimise pretraining on small corpora. ELC-BERT modifies residual connections to selectively weigh outputs from previous layers. Each layer's input is a combination of outputs from previous layers, weighted by learnable layer-specific weights. This is in contrast to standard residual connections, where the input is an equally weighted sum of all preceding outputs. The added expressivity of ELC-BERT, which allows the model to dynamically weigh how preceding layers are incorporated into computations, enables more sample-efficient learning.



(a) After 10 epochs of training.

(b) After 200 epochs of training.

Figure 2: Layer contribution heatmaps of isiXhosa ELC-BERT at different stages of pretraining.

2.2.2 Masked Latent Semantic Modeling (MLSM)

MLSM (Berend, 2023b,a) is an alternative to standard masked language modeling. Instead of tasking the model with predicting specific tokens, which can be challenging given limited training data, the model is trained to predict broader semantic categories. For example, if the model is tasked to predict the masked word "barbecue", it would generate predictions towards the semantic attributes associated with the word (e.g. "food", "outdoors", "fire"). MLSM uses a teacher model to determine latent semantic distributions for masked tokens, via sparse coding of their hidden representations. The final model is then a student model, trained to predict these latent semantic distributions rather than the exact identities of masked tokens.

3 Experimental Setup

Pretraining Our BabyLMs and baseline are pretrained on the WURA isiXhosa corpus (Oladipo et al., 2023), which is a compiled by filtering mC4 (Xue et al., 2021) to remove noise. The isiXhosa dataset contains 13m words, similar in size to BabyLM *Strict-Small*. Our models are trained for 200 epochs on a Tesla V100 GPU. We detail our training process in Appendix A.

Evaluation We evaluate on three isiXhosa datasets – MasakhaPOS (Dione et al., 2023) for POS tagging, MasakhaNER (Adelani et al., 2022) for NER, and MasakhaNEWS (Adelani et al., 2023) for news topic classification (NTC). Test set results are averaged across 5 finetuning runs.

4 Results

Table 1 presents our results. Both BabyLMs outperform the baseline on POS and NER, achieving large gains for NER (+3.2 F1 for ELC-BERT and +1.6 F1 for MLSM). As in the original shared task, ELC-BERT is the top-performing BabyLM. ELC-BERT demonstrates superior efficiency in both data utilisation (shown in Figure 1) and computate requirements (its pretraining time is 70% faster than MLSM). The BabyLMs fail to outperform RoBERTa on NTC. We attribute this to topic classification being an easier task than POS and NER, so data-efficiency is less critical. In fact, our RoBERTa baseline even outperforms two skylines on NTC, reaffirming previous findings that pretraining from scratch is sufficient for the simpler task of NTC (Ogueji et al., 2021; Dossou et al., 2022). We also posit that the architectures of ELC-BERT and MLSM are more suitable for word-level tasks than sequence-level tasks (discussed in section 5).

ELC-BERT outperforms one skyline, XLM-R, on two tasks. Unlike in the shared task, our models do not outperform the top skylines. We attribute this to an important difference between our setup and the shared task – the quality of pretraining data. The WURA corpus does not match the quality of the BabyLM data, which was curated to include developmentally plausible text (e.g. child-directed speech, educational content). The previous success of these models in English is due to a combination of modelling innovations and extremely high-quality pretraining data, which is lacking for low-resource languages like isiXhosa.

Figure 3: Top 10 semantic categories predicted by isiXhosa MLSM for named entities (sampled from MasakhaNER).

5 Analysis

The BabyLMs studied in this paper achieve data-efficiency by augmenting the standard MLM architecture. We now analyse how their unique architectural innovations encode the isiXhosa language.

ELC-BERT The residual connections of ELC-BERT learn to selectively weigh the output of previous layers. We visualise learned weights in Figure 2, comparing early pretraining to complete pretraining (intermediary stages are visualised in Figure 4 in the appendix). The weighting exhibits significant deviations from a standard Transformer layer (which assigns equal weight to all preceding outputs). In early stages of pretraining, the model is biased to the embedding layer and immediately preceding layers. As pretraining progresses, the model reduces its reliance on embeddings in favour of immediately preceding layers, but still assigns more weight to the embedding layer than the BabyLM ELC-BERT submission. We posit that this emphasis on the embedding layer underlies ELC-BERT's performance gains on POS tagging and NER, since embeddings encode information about word-level syntactic roles (Tenney et al., 2019).

MLSM During pretraining, MLSM predicts the latent semantic categories of masked tokens. To inspect the semantic distribution learned by the model, we extract the predictions for masked named entities in sentences sampled from MasakhaNER. Figure 3 shows the top 10 semantic categories (each corresponding to an index) assigned to four named entities. For each target word, we also plot the probabilities produced for the other words to compare distributions. In our sampled sentences, two of the words (*Justin* and *Augustine*) are names of persons, while the other two (*Morocco* and *Soweto*) are names of locations. The plots demonstrate more semantic overlap between the same types of named entities. The names of persons have seven overlapping semantic categories, while the names of locations have nine overlapping categories. Between the two named entity types, only two semantic categories overlap. This pattern indicates that MLSM effectively encodes the semantic properties of these named entities, to which we attribute its NER performance gains. We present a similar analysis for target words across POS tags in Appendix B.

6 Conclusion

This study explored the potential of two architectures from the BabyLM challenge, ELC-BERT and MLSM, to benefit low-resource languages. Comparing our findings to those of the BabyLM challenge, we draw three main conclusions. Firstly, the

gains obtained by isiXhosa BabyLMs show that the sample-efficiency sought by the BabyLM challenge can prove effective in real low-resource settings. Secondly, ELC-BERT once again emerges as the most data-efficient solution, even outperforming massively multilingual PLMs. Lastly, the fact that our BabyLMs do not outperform all skylines shows that the absence of high-quality corpora for isiXhosa poses a barrier to further gains. The findings of the BabyLM challenge can be attributed to both architectural innovations and specifically curated pretraining data. The BabyLM pretraining corpus includes child-directed speech, educational video subtitles, and articles from Simple Wikipedia (an edition of Wikipedia written in simplified English, using shorter sentences and common words). Such high-quality, developmentally plausible data is not publicly available for isiXhosa. Our results show that this limits the potential of BabyLMs for low-resource languages.

More generally, this work unites two directions of research – cognitively plausible modelling and NLP for low-resource languages. We hope more researchers pursue work at the intersection of these two subfields, since they share the goal of improving data-efficiency in the era of scaling.

7 Limitations

Our study focussed on a single language, isiXhosa, so our findings might not generalise to other low-resource languages. We chose isiXhosa because its data availability was well suited to our study. Publicly available pretraining corpora for isiXhosa are similar in size to the BabyLM *Strict-Small* corpus. In terms of downstream evaluation data, isiXhosa also has sufficient NLU datasets available to allow evaluation across sequence labelling and sequence classification tasks. The BabyLM challenge evaluated submissions across many more tasks than we did, some of which are much more challenging than our isiXhosa evaluation tasks. Ideally, one would evaluate our isiXhosa BabyLMs on datasets that test more aspects of language competence. This would reveal further insights into the value of BabyLM architectures compared to standard baselines and/or skylines, which might not align with our current findings. We hypothesise that more complex evaluation tasks would further highlight the value of BabyLMs over standard Transformer baselines, but due to the lack of additional isiXhosa evaluation datasets we are unable to test this.

References

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, and Joyce Nakatumba-Nabende. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic

of Korea. International Committee on Computational Linguistics.

Gábor Berend. 2023a. Better together: Jointly using masked latent semantic modeling and masked language modeling for sample efficient pre-training. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 298–307, Singapore. Association for Computational Linguistics.

Gábor Berend. 2023b. Masked latent semantic modeling: an efficient pre-training alternative to masked language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13949–13962, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, , and Charles D. Fenning. 2019. *Ethnologue: Languages of the World*, 22 edition. SIL International.

Lucas Georges Gabriel Charpentier and David Samuel. 2023. Not all layers are equally as important: Every layer counts BERT. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 238–252, Singapore. Association for Computational Linguistics.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Francois Meyer, Haiyue Song, Abhisek Chakrabarty, Jan Buys, Raj Dabre, and Hideki Tanaka. 2024. NGLUEni: Benchmarking and adapting pretrained language models for nguni languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12247–12258, Torino, Italia. ELRA and ICCL.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Owodunni, Odunayo Ogundepo, David Adelani, and Jimmy Lin. 2023. Better quality pre-training data and t5 models for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168, Singapore. Association for Computational Linguistics.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

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. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. 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*, pages 1–34, Singapore. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Training Details

For pretraining, we use the training scripts accompanying the BabyLM submissions, and use their hyperparameter settings for the *Strict-Small* track as a starting point. We pretrain our BabyLMs and RoBERTa baseline for 200 epochs of the isiXhosa WURA corpus. Our hyperparameter settings are listed in Table 2.

Model	LR	SL	H	BS
RoBERTa	$5e^{-5}$	512	12	8
ELC-BERT	$5e^{-4}$	128	12	128
MLSM (teacher)	$1e^{-4}$	128	12	64
MLSM (student)	$1e^{-4}$	128	12	64

Table 2: Pretraining hyperparameters (Learning Rate, Sequence Length, Hidden layers, Batch Size)

ELC-BERT pretraining Due to computational constraints, we trained our ELC-BERT model for 200 epochs, instead of the 2000 epochs of the BabyLM submission. Regardless, downstream performance for POS tagging and NER does plateau by 200 epochs (Figure 1). Besides the number of epochs, we made two changes to the hyperparameter settings of the ELC-BERT submission (Georges Gabriel Charpentier and Samuel, 2023). Firstly, we used a batch size of 128 (instead of 256) due to computational constraints. Secondly, the original learning rate ($1e^{-2}$) produced an unstable training loss, so after some experimentation we settled on a learning rate of $5e^{-4}$.

MLSM pretraining We trained the teacher and student model from scratch on the WURA dataset, keeping the same hyperparameters as the MLSM submission (Berend, 2023a). Our teacher model is based on the BERT-base-cased architecture[1] and is trained using a standard masked language modelling objective. We used the teacher model hidden layers to create a semantic dictionary for the student model. The student model is also based on the BERT-base-cased architecture, but is trained to predict semantic categories instead of masked tokens.

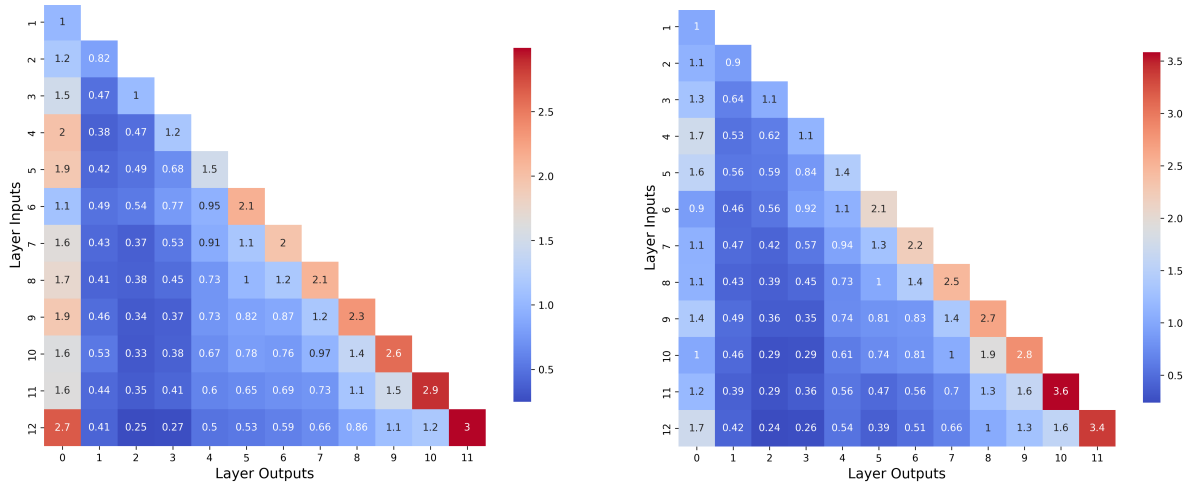
Finetuning We use the finetuning scripts provided by the MasakhaPOS (Dione et al., 2023), MasakhaNER (Adelani et al., 2022), and MasakhaNEWS (Adelani et al., 2023) datasets where possible, and adapt them for ELC-BERT. Each model is fine-tuned for 20 epochs per task, using the default hyperparameters provided in the respective dataset fine-tuning scripts. For each task, we perform 5 finetuning runs using different ran-

[1] https://huggingface.co/google-bert/bert-base-cased

dom seeds. We report the averages and standard deviations over these runs in Table 1.

B MLSM Analysis

The predictions shown in Figure 3 are obtained by masking the target words in sentences sampled from MasakhaNER. We conduct a similar analysis for target words with different POS tags, sampling sentences from MasakhaPOS. Figure 5 shows the top 10 semantic categories assigned to four words with different parts of speech. Two of the words (*phambi* and *emva*) are adpositions, while the other two (*kwaye* and *ukaba*) are conjunctions. The plots demonstrate less semantic overlap between same POS tags than named entity types. The adpositions have three overlapping semantic categories, while the conjunctions share four overlapping categories. Between the different POS tags, there is still minimal overlap: *phambi* shares one category with *kwaye* and three categories with *ukaba*, while *emva* shows no overlap with either conjunction. We attribute this pattern to the broader and less interchangeable nature of POS tags compared to named entities, making them less suited to MLSM's strengths. The reduced semantic overlap, compared to named entities, might be why MLSM's effectiveness varies across linguistic tasks. This aligns with the results shown in Table 1, where MLSM's performance gains for POS tagging show a narrower margin over the baseline compared to the improvements in NER.



(a) After 20 epochs of training.

(b) After 100 epochs of training.

Figure 4: Layer contribution heatmaps of isiXhosa ELC-BERT at different stages of pretraining.

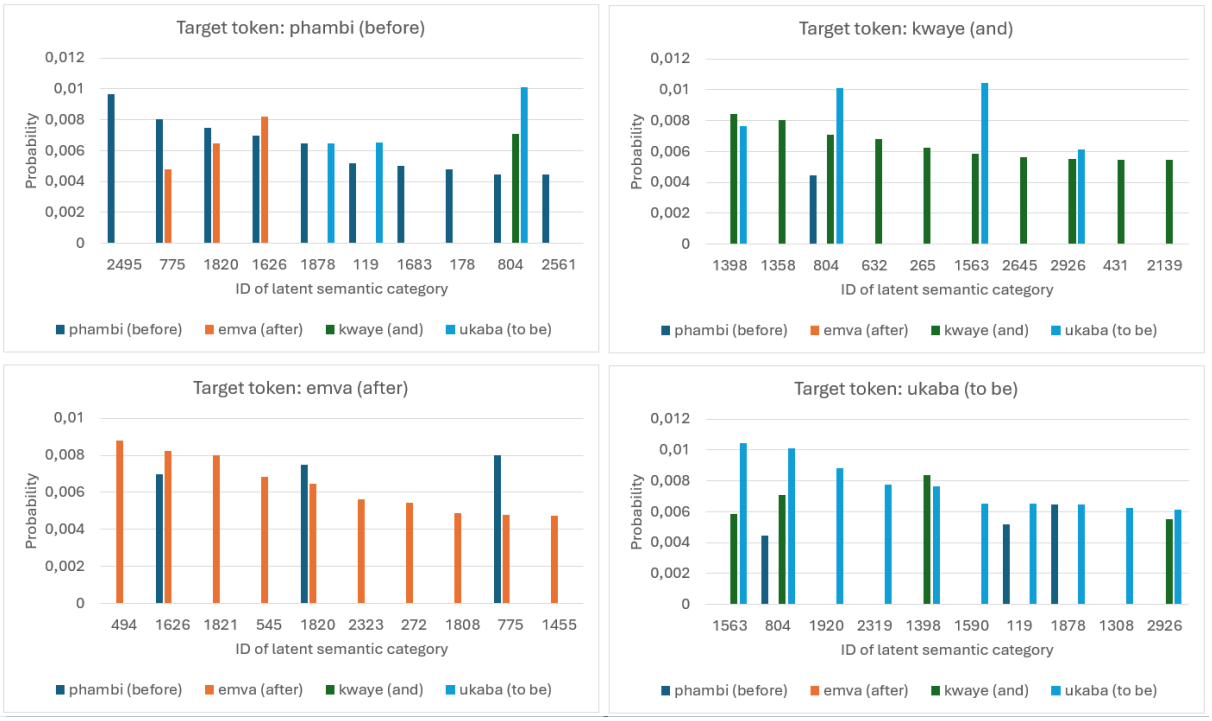


Figure 5: Top 10 semantic categories predicted by isiXhosa MLSM for target words (sampled from MasakhaPOS).