Unknown Script: Impact of Script on Cross-Lingual Transfer

Wondimagegnhue Tsegaye Tufa, Ilia Markov, Piek Vossen

Vrije Universiteit Amsterdam

De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

{w.t.tufa, i.markov, p.t.j.m.vossen}@vu.nl

Abstract

Cross-lingual transfer has become an effective way of transferring knowledge between languages. In this paper, we explore an often-overlooked aspect in this domain: the influence of the source language of a language model on language transfer performance. We consider a case where the target language and its script are not part of the pre-trained model. We conduct a series of experiments on monolingual and multilingual models that are pre-trained on different tokenization methods to determine factors that affect cross-lingual transfer to a new language with a unique script. Our findings reveal the importance of the tokenizer as a stronger factor than the shared script, language similarity, and model size.

1 Introduction

The dominant Natural Language Processing (NLP) approach nowadays involves cross-lingual transfer using pre-trained monolingual and multilingual language models. In line with this trend, numerous monolingual models have been released for various languages (Devlin et al., 2019; Cañete et al., 2023; Antoun et al., 2020). Multilingual models, which are trained on 100 or more languages, have also been introduced, such as XLM-R (Conneau et al., 2020) and m-BERT (Devlin et al., 2019).

Despite the advancements in the development of language models for high-resourced languages, the vast majority of the world's languages remain excluded from these models. Out of over 6,500 spoken languages globally, less than 2% are represented in the current models, leaving the rest unseen and unaccounted for in the current language processing technology (Muller et al., 2021; Hammarström, 2016; Joshi et al., 2020). Training a model for each of these languages is impractical due to substantial data and computational resource requirements. Several alternative approaches have been proposed, such as zero-shot
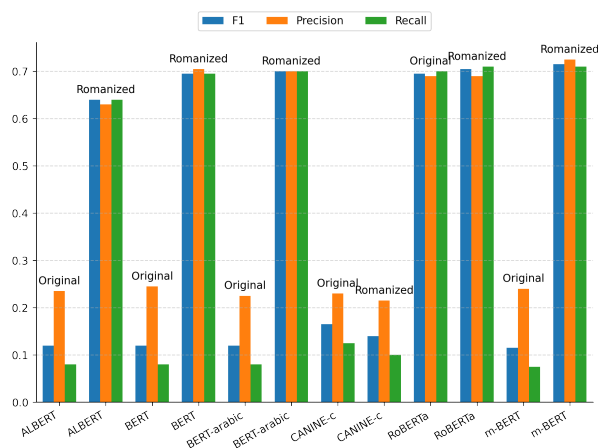

Figure 1: We analyze the effect of script and tokenizer on cross-lingual transfer on a target language with a new script. We select six monolingual and multilingual models pre-trained using sub-word tokenizers and character tokenizers. We fine-tune these models on the NER and POS tasks in the original script (FIDEL) and the romanized version (Latin). We observe that RoBERTa has better cross-lingual transfer in both the original script and the romanized version. We also find that romanization is strongly beneficial in all cases of subword-based models (ALBERT, BERT, m-BERT). Additionally, fine-tuning Arabic-BERT, which is typologically similar to our target language (Amharic), provides no advantage. We employ the base version of the models across all cases to ensure a fair comparison. The reported F1-score is averaged over five runs, with a standard deviation ranging between -0.003 and 0.009.

transfer (Pires et al., 2019; Conneau et al., 2020), language adapters (Pfeiffer et al., 2020), and extending multilingual models (Conneau et al., 2020; Devlin et al., 2019).

While such methods require fewer resources in the target language, they often yield sub-optimal results (Lauscher et al., 2020; Pfeiffer et al., 2021). For example, using even a small amount of labeled data in the target language has been shown to be more effective than a zero-shot transfer (Lauscher et al., 2020).

The effectiveness of cross-lingual transfer is influenced by several factors, such as language similarity between the source and target languages, the size of the pre-trained model, and the quality and amount of the pre-training and fine-tuning data (Muller et al., 2021; Pires et al., 2019; Cao et al., 2020; Wu and Dredze, 2019).

In this paper, we explore which factors determine cross-lingual transfer performance for a new language with a unique script. We consider a challenging case where the target language is not part of the pre-trained model, and the script has also not been seen in pre-training. Our analysis targets three main factors: language similarity, tokenization methods, and script. We design an experiment by varying these factors. We evaluate a range of existing monolingual and multilingual models, specifically choosing those trained on typologically related or typologically distant languages. Furthermore, we select models trained using various tokenizers, allowing us to assess how these choices impact cross-lingual performance for a language with a unique script. We focus on two main research questions:

1. To what extent does the script of a source language influence cross-lingual transfer to a new language in monolingual and multilingual models?

2. To what extent does the tokenizer influence cross-lingual transfer to a new language in monolingual and multilingual models?

Figure 1 shows the results of our experiment. Our analysis shows that RoBERTa has better cross-lingual transfer irrespective of the script. However, romanization strongly affects cross-lingual transfer for models pre-trained using sub-word tokenizers in monolingual and multilingual settings. We make our code available.[1]

2   Related Work

Under-resourced languages display considerable variation in several aspects (Joshi et al., 2020). First, the amount of data for these languages varies greatly. Secondly, many of these languages use scripts different from the Latin script (Muller et al., 2021; Joshi et al., 2020). Finally, regarding linguistic characteristics, these languages often have

---

[1] https://github.com/cltl/unkown_script/tree/main

distinct morphological and syntactic properties, especially when compared to high-resourced Indo-European languages.

Recently, cross-lingual transfer has become an effective method for extending the capabilities of pre-trained monolingual and multilingual models for various languages. In this section, we present an overview of studies exploring cross-lingual transfer.

Multilingual models and language adapters Multilingual models enable transfer between high-resource and low-resource languages (Conneau et al., 2020; Devlin et al., 2019). However, they suffer from the 'curse of multilingualism' and interference between languages (Conneau et al., 2020; Wang et al., 2020), where the model's effectiveness decreases as the number of languages increases due to the parameter limit of the model. Language adapters address these challenges by storing language-specific knowledge of each language in dedicated parameters (Pfeiffer et al., 2020). This increases the capacity of a multilingual model without introducing interference between languages. These methods are, however, not directly applicable to languages that use scripts not covered in the training data of these models (Pfeiffer et al., 2021).

Zero-shot and few-shot transfer In zero-shot transfer, a fine-tuned model on a resource-rich source language is directly applied to a resource-poor target language (Pires et al., 2019; Conneau et al., 2020). While this method is appealing, it often yields sub-optimal results (Lauscher et al., 2020; Pfeiffer et al., 2021). Alternatively, using even a small amount of labeled data in the target language (few-shot transfer) has shown to be more effective (Lauscher et al., 2020). It remains unclear what factors determine this effect and to what extent.

Factor analysis The success of cross-lingual transfer is impacted by various factors (Muller et al., 2021; Pires et al., 2019; Cao et al., 2020; Wu and Dredze, 2019). Muller et al. (2021) demonstrated that the performance of transfer can significantly differ based on factors such as the script of the language, the amount of available data, and the relationship between source and the target language. However, the literature concerning the effect of script and tokenizer is mixed. For example, Muller et al. (2021) identifies script as the most crucial factor affecting transfer performance and

shows that transliterating a script to the Latin script enhances the effectiveness of cross-lingual transfer. Contrary to this, Artetxe et al. (2019) and Karthikeyan et al. (2020) show that script and lexical overlap are less relevant and that large monolingual models learn semantic abstractions that are generalizable to other languages. A similar mixed result has been reported when examining the effect of tokenizers in cross-lingual transfer. Rust et al. (2021) show that tokenizers are a crucial factor in the success of cross-lingual transfer for multilingual models, while Wu et al. (2022) report it as less important. The analysis of (Muller et al., 2021) covers multilingual models, while Wu et al. (2022) focuses only on an English model.

While similar to the approach of Muller et al. (2021), our study takes a different direction. Instead of analyzing the performance of multiple languages with a single multilingual model, we focus on one language that is unique in its script and not covered by existing models. We select Amharic as our target language. Amharic is a Semitic language and is classified as morphologically complex. It has a unique script, distinct from Latin and Arabic alphabets, with no shared characters. Additionally, it is categorized as a Class 2 under-resourced language, indicating a significant lack of data and tools for language processing (Jo et al., 2021; Adelani et al., 2021). We evaluate a range of existing monolingual and multilingual models, specifically choosing those trained on languages either closely related to or distinct from Amharic. Furthermore, we select models trained using various tokenizers, allowing us to assess how these choices impact model performance for a language with a unique script. In this way, we measure the impact of different factors on cross-lingual transfer.

3 Methodology

We experiment with a few-shot setting in which we fine-tune pre-trained models on downstream tasks. The target language and script are not part of the pre-trained models we explore. Our few-shot setup follows a standard setting: we take an existing base model, fine-tune it, and test it on the original target language.

Language We experiment with Amharic as our target language. According to the language classification system presented in (Joshi et al., 2020), Amharic is categorized as a Class 2 under-resourced language. It possesses a unique script

and is characterized by its morphological complexity. We select two source languages for our analysis: English, which is typologically distant from Amharic, and Arabic, which is typologically related. Both English and Arabic of these languages do not share a script with Amharic. We use the original Fidel script and the romanized version for our fine-tuning and evaluation.

3.1 Task and model

Table 1 shows a summary of the pre-trained models we use and their corresponding tokenizers. The selection includes monolingual models trained in English and Arabic and various multilingual models.

Task and Datasets We experiment with two tasks: Named Entity Recognition (NER) and part-of-speech (POS) tagging. For the NER task, we use MasakhaNER (Adelani et al., 2021), and for POS, we use the Amharic-POS dataset from (Gezmu et al., 2021). The MasakhaNER dataset is a publicly accessible resource for the NER task in ten African languages. This dataset has four types of entity labels: PER (Person), ORG (Organization), LOC (Location), and DATE (Date). It has 1,750 training, 250 validation, and 500 test instances. The POS dataset contains 218K sentences with 18 POS tags. We sample 2.5K examples, using 1,750 sentences for training, 250 for validation, and 500 for testing.

Tokenizers Our model selection encompasses the most widely used tokenizers: SentencePiece (Kudo and Richardson, 2018), BPE (Sennrich et al., 2016), character-based tokenizer (Clark et al., 2022) and WordPiece (Schuster and Nakajima, 2012).

Fine-tuning We fine-tune all models on two tasks. Our fine-tuning is challenging because our target language is not seen during pre-training and has a unique non-Latin script. The experiment is designed to test two capabilities. First, we evaluate whether the models we are testing enable cross-lingual transfer to a new language and script unseen in pre-training. This experiment is intended to shed light on cross-lingual transfer to a target language that does not share a script with the source language. Second, we explore how different tokenization methods might facilitate cross-lingual transfer.

Model	Tokenizer	Model-Type	Language	Model Size
BERT (Devlin et al., 2019)	WordPiece	Monolingual	English	110M
RoBERTa (Liu et al., 2019)	BPE	Monolingual	English	125M
ALBERT (Devlin et al., 2019)	SentencePiece	Monolingual	English	12M
BERT-base-arabic (Antoun et al., 2020)	WordPiece	Monolingual	Arabic	110M
CANINE-c (Clark et al., 2022)	Character	Multilingual	Multiple	121M
m-BERT (Devlin et al., 2019)	WordPiece	Multilingual	Multiple	110M

Table 1: Overview of the models we use, tokenizers, model types, languages, and the model size.

Model	Fidel-NER			Latin-NER			Fidel-POS			Latin-POS		
	F1	P	R	F1	P	R	F1	P	R	F1	P	R
RoBERTa-base	**0.57**	0.59	0.55	0.55	0.53	0.56	**0.82**	0.79	0.85	**0.86**	0.85	0.86
CANINE-c	0.12	0.13	0.1	0.09	0.12	0.07	0.21	0.33	0.15	0.19	0.31	0.13
BERT-base	-	-	-	0.57	0.6	0.55	0.24	0.49	0.16	0.82	0.81	0.84
m-BERT	-	-	-	**0.59**	0.62	0.57	0.23	0.48	0.15	0.84	0.83	0.85
BERT-arabic	-	-	-	**0.59**	0.6	0.58	0.24	0.45	0.16	0.81	0.8	0.82
ALBERT-base	-	-	-	0.5	0.5	0.49	0.24	0.47	0.16	0.76	0.76	0.79

Table 2: Results of few-shot experiments where we fine-tune different models on NER and POS tasks with Fidel and romanized script. The empty cells show that we do not observe a decrease in the loss. We fine-tune all our models for 25 epochs. The F1 score is averaged over five runs with a standard deviation between 0.003 and 0.009. The highest F1 score for each script is highlighted in bold.

4 Results and Analysis

Table 2 shows the result for the few-shot setting on NER and POS fine-tuned on the original Amharic script (Fidel) and its romanized version. The RoBERTa-base model stands out, showing robust performance across both scripts, with a marginal preference for the Fidel script for the NER task and the Latin script for the POS task. Models such as BERT-base, m-BERT, BERT-Arabic, and ALBERT-base fail to recognize entities entirely in the Fidel script but show some capabilities with the Latin script. This pattern persists across both tasks, although the POS task has less variation.

Effect of the script The difference in the obtained results for Fidel and romanized versions highlights the script's effect on model performance. Models pre-trained on data predominantly in the Latin script struggle significantly with tasks in the Fidel script, as shown by the drastically lower F1 scores for most models trained on the Fidel script compared to the Latin script. This suggests a strong bias towards the script used during the training phase, with models performing better on scripts they have seen before. This is in line with the result reported by Purkayastha et al. (2023); Muller et al. (2021), which shows that the romanization of unknown script boosts transfer performance in multilingual models. However, we also observe this effect in all of the monolingual models we test.

The RoBERTa model seems to be an exception, showing a good performance before romanization, though romanization also slightly improves its performance.

Language relatedness English BERT-base and Arabic BERT can be compared directly since they are trained with similar training objectives, model sizes, and tokenizers. We observe a mixed result, with BERT-Arabic performing slightly better on the NER task but showing a lower score on the POS task.

Model size and tokenizers A plausible explanation for the performance variation between RoBERTa and the other models could be attributed to the model size and tokenizer. RoBERTa-base is the largest model with 125 million parameters. However, the performance does not consistently correlate with the model size. The other possible explanation is the tokenizer used. RoBERTa is trained using BPE over raw bytes instead of Unicode characters. The results show that the BPE representation enables the model to leverage knowledge that benefits the downstream tasks, even if the script is not included in the model's vocabulary.

5 Conclusion

In this paper, we explore cross-lingual transfer in less explored but challenging settings where the target language is not seen during pre-training and

has a unique non-Latin script. Our analysis shows considerable differences in cross-lingual transfer performance among various models, possibly attributable to two key factors: the size of the pre-trained model and the specific tokenizer used during pre-training. The model's size could impact its ability to capture and generalize across multiple languages, including a language distant from the pre-trained model's language. In light of recent studies that show the importance of tokenizers in cross-lingual transfer for under-resourced languages, we show that a choice of tokenizer plays a role in facilitating cross-lingual transfer.

6 Limitations

In our analysis, we intend to control for various factors that could influence the comparative performance of different models. However, residual differences in model parameters and the extent of pre-training data may have contributed to the observed disparities in the obtained results. Furthermore, our study did not involve training a model from scratch with fixed architecture, parameters, data, and domain while varying only the tokenizer or the model size. This limitation precludes a definitive conclusion about the isolated effect of the tokenizer on model performance. Hence, we identify the need for additional research where these elements are carefully controlled. Such an experiment would enable a more robust understanding of the tokenizer's role and interaction with other model characteristics in cross-lingual transfer learning.

References

David Ifeoluwa Adelani, Jade Z. Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba O. Alabi, Seid Muhie Yimam, Tajuddeen Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin P. Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane Mboup, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing K. Sibanda,

Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima Diop, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. Masakhaner: Named entity recognition for african languages. In *2nd AfricaNLP Workshop Proceedings, AfricaNLP@EACL 2021, Virtual Event, April 19, 2021*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. *CoRR*, abs/2002.03518.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2023. Spanish pre-trained bert model and evaluation data.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. <scp>canine</scp>: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

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.

Andargachew Mekonnen Gezmu, Binyam Ephrem Seyoum, Michael Gasser, and Andreas Nürnberger. 2021. Contemporary amharic corpus: Automatically morpho-syntactically tagged amharic corpus. *CoRR*, abs/2106.07241.

Harald Hammarström. 2016. Linguistic diversity and language evolution. *Journal of Language Evolution*, 1(1):19–29.

Yohan Jo, Seojin Bang, Chris Reed, and Eduard Hovy. 2021. Classifying argumentative relations using logical mechanisms and argumentation schemes. *Transactions of the Association for Computational Linguistics*, 9:721–739.

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, Online. Association for Computational Linguistics.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

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.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Sukannya Purkayastha, Sebastian Ruder, Jonas Pfeiffer, Iryna Gurevych, and Ivan Vulić. 2023. Romanization-based large-scale adaptation of multilingual language models.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Zhengxuan Wu, Isabel Papadimitriou, and Alex Tamkin. 2022. Oolong: Investigating what makes crosslingual transfer hard with controlled studies.