# Identifying the Correlation Between Language Distance and Cross-Lingual Transfer in a Multilingual Representation Space

**Fred Philippy**[1,2]* and **Siwen Guo**[1] and **Shohreh Haddadan**[1]

[1]Zortify Labs, Zortify S.A.
19, rue du Laboratoire L-1911 Luxembourg
[2]SnT, University of Luxembourg
29, Avenue J.F Kennedy L-1359 Luxembourg
{fred, siwen, shohreh}@zortify.com

## Abstract

Prior research has investigated the impact of various linguistic features on cross-lingual transfer performance. In this study, we investigate the manner in which this effect can be mapped onto the representation space. While past studies have focused on the impact on cross-lingual alignment in multilingual language models during fine-tuning, this study examines the absolute evolution of the respective language representation spaces produced by MLLMs. We place a specific emphasis on the role of linguistic characteristics and investigate their inter-correlation with the impact on representation spaces and cross-lingual transfer performance. Additionally, this paper provides preliminary evidence of how these findings can be leveraged to enhance transfer to linguistically distant languages.

## 1 Introduction

It has been shown that language models implicitly encode linguistic knowledge (Jawahar et al., 2019; Otmakhova et al., 2022). In the case of multilingual language models (MLLMs), previous research has also extensively investigated the influence of these linguistic features on cross-lingual transfer performance (Lauscher et al., 2020; Dolicki and Spanakis, 2021; de Vries et al., 2022). However, limited attention has been paid to the impact of these factors on the language representation spaces of MLLMs.

Despite the fact that state-of-the-art MLLMs such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), use a shared vocabulary and are intended to project text from any language into a language-agnostic embedding space, empirical evidence has demonstrated that these models encode language-specific information across all layers (Libovický et al., 2020; Gonen et al., 2020). This leads to the possibility of identifying distinct monolingual representation spaces within the shared multilingual representation space (Chang et al., 2022).

Past research has focused on the cross-linguality of MLLMs during fine-tuning, specifically looking at the alignment of representation spaces of different language pairs (Singh et al., 2019; Muller et al., 2021). Our focus, instead, is directed towards the absolute impact on the representation space of each language individually, rather than the relative impact on the representation space of a language compared to another one. Isolating the impact for each language enables a more in-depth study of the inner modifications that occur within MLLMs during fine-tuning. The main objective of our study is to examine the role of linguistic features in this context, as previous research has shown their impact on cross-lingual transfer performance. More specifically, we examine the relationship between the impact on the representation space of a target language after fine-tuning on a source language and five different language distance metrics. We have observed such relationships across all layers with a trend of stronger correlations in the deeper layers of the MLLM and significant differences between language distance metrics.

Additionally, we observe an inter-correlation among language distance, impact on the representation space and transfer performance. Based on this observation, we propose a hypothesis that may assist in enhancing cross-lingual transfer to linguistically distant languages and provide preliminary evidence to suggest that further investigation of our hypothesis is merited.

## 2 Related Work

In monolingual settings, Jawahar et al. (2019) found that, after pre-training, BERT encodes different linguistic features in different layers. Merchant et al. (2020) showed that language models do not forget these linguistic structures during fine-tuning on a downstream task. Conversely, Tanti et al.

*Research was conducted at Zortify.



*Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2023)*, pages 22–29

(2021) have shown that during fine-tuning in multilingual settings, mBERT forgets some language-specific information, resulting in a more cross-lingual model.

At the representation space level, Singh et al. (2019) and Muller et al. (2021) studied the impact of fine-tuning on mBERT's cross-linguality layer-wise. However, their research was limited to the evaluation of the impact on cross-lingual alignment comparing the representation space of one language to another, rather than assessing the evolution of a language's representation space in isolation.

## 3 Methodology

### 3.1 Experimental Setup

In this paper, we focus on the effect of fine-tuning on the representation space of the 12-layer multilingual BERT model (`bert-base-multilingual-cased`). We restrict our focus on the Natural Language Inference (NLI) task and fine-tune on all 15 languages of the XNLI dataset (Conneau et al., 2018) individually. We use the test set to evaluate the zero-shot cross-lingual transfer performance, measured as accuracy, and to generate embeddings that define the representation space of each language. More details on the training process and its reproducibility are provided in Appendix A.

### 3.2 Measuring the Impact on the Representation Space

We focus on measuring the impact on a language's representation space in a pre-trained MLLM during cross-lingual transfer. We accomplish this by measuring the similarity of hidden representations of samples from different target languages before and after fine-tuning in various source languages. For this purpose, we use the Centered Kernel Alignment (CKA) method (Kornblith et al., 2019)[1]. When using a linear kernel, the CKA score of two representation matrices $X \in \mathbb{R}^{N \times m}$ and $Y \in \mathbb{R}^{N \times m}$, where $N$ is the number of data points and $m$ is the representation dimension, is given by

$$CKA(X, Y) = 1 - \frac{\|XY^\intercal\|_F^2}{\|XX^\intercal\|_F \|YY^\intercal\|_F}$$

where $\|\cdot\|_F$ is the Frobenius norm.

[1]CKA is invariant to orthogonal transformations and thus allows to reliably compare isotropic but language-specific subspaces (Chang et al., 2022).

**Notation** We define $H^i_{S \to T} \in \mathbb{R}^{N \times m}$ as the hidden representation[2] of $N$ samples from a target language $T$ at the $i$-th attention layer of a model fine-tuned in the source language $S$, where $m$ is the hidden layer output dimension. Similarly, we denote the hidden representation of $N$ samples from language $L$ at the $i$-th attention layer of a pre-trained base model (i.e. before fine-tuning) as $H^i_L \in \mathbb{R}^{N \times m}$. More specifically, the representation space of each language will be represented by the stacked hidden states of its samples.

We define the impact on the representation space of a target language $T$ at the $i$-th attention layer when fine-tuning in a source language $S$ as follows:

$$\Phi^{(i)}(S, T) = 1 - CKA\left(H^i_T, H^i_{S \to T}\right)$$

### 3.3 Measuring Language Distance

In order to quantify the distance between languages we use three types of typological distances, namely the syntactic (SYN), geographic (GEO) and inventory (INV) distance, as well as the genetic (GEN) and phonological (PHON) distance between source and target language. These distances are pre-computed and are extracted from the URIEL Typological Database (Littell et al., 2017) using `lang2vec`[3]. For our study, such language distances based on aggregated linguistic features offer a more comprehensive representation of the relevant language distance characteristics. More information on these five metrics is provided in Appendix B.

## 4 Correlation Analysis

**Relationship Between the Impact on the Representation Space and Language Distance.** Given the layer-wise differences of mBERT's cross-linguality (Libovický et al., 2020; Gonen et al., 2020), we measure the correlation between the impact on the representation space and the language distances across all layers. Figure 1 shows almost no significant correlation between representation space impact and **inventory** or **phonological** distance. **Geographic** and **syntactic** distance mostly show significant correlation values at the last layers. Only the **genetic** distance correlates significantly across all layers with the impact on the representation space.

[2]We refer here to the hidden representation of the `[CLS]` token which is commonly used in BERT for classification tasks.

[3]https://github.com/antonisa/lang2vec

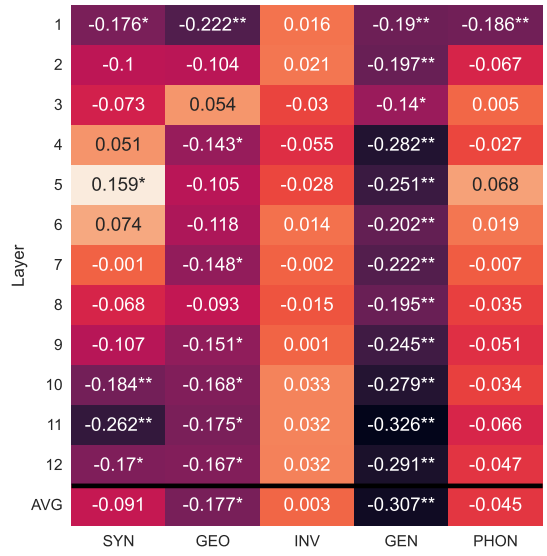


Figure 1: **Pearson correlation coefficient** between the **impact on a target language's representation space when fine-tuning in a source language** and different types of **linguistic distances between the source and target language** for each layer. Same source-target language pair data points were excluded in order to prevent an overestimation of effects. ($^{*}p < 0.05$, and $^{**}p < 0.01$, two-tailed).

**Relationship Between Language Distance and Cross-Lingual Transfer Performance.** Table 1 shows that all distance metrics correlate with cross-lingual transfer performance, which is consistent with the findings of Lauscher et al. (2020). Furthermore, we note that the correlation strengths align with the previously established relationship between language distance and representation space impact, with higher correlation values observed for syntactic, genetic, and geographic distance than for inventory and phonological distance. The exact zero-shot transfer results are provided in Figure 3 in Appendix C.

| | Pearson | Spearman |
|---|---|---|
| SYN | $-0.3193^{**}$ | $-0.4683^{**}$ |
| GEO | $-0.3178^{**}$ | $-0.3198^{**}$ |
| INV | $-0.1706^{*}$ | $-0.1329^{*}$ |
| GEN | $-0.3364^{**}$ | $-0.3935^{**}$ |
| PHON | $-0.2075^{**}$ | $-0.2659^{**}$ |

Table 1: Pearson and Spearman **correlation coefficients** quantifying the relationship between **zero-shot cross-lingual transfer performance** and different **language distance metrics**. ($^{*}p < 0.05$, and $^{**}p < 0.01$, two-tailed).

**Relationship Between the Impact on the Representation Space and Cross-Lingual Transfer Performance.** In general, cross-lingual transfer performance clearly correlates with impact on the representation space of the target language, but this correlation tends to be stronger in the deeper layers of the model (Table 2).

| Layer | Pearson | Spearman |
|---|---|---|
| 1 | $0.2779^{*}$ | $0.3233^{*}$ |
| 2 | $0.2456^{*}$ | $0.2639^{*}$ |
| 3 | $0.5277^{*}$ | $0.5926^{*}$ |
| 4 | $0.3585^{*}$ | $0.3411^{*}$ |
| 5 | $-0.009$ | $0.0669$ |
| 6 | $0.1033$ | $0.1969$ |
| 7 | $0.2945^{*}$ | $0.3500^{*}$ |
| 8 | $0.3004^{*}$ | $0.3517^{*}$ |
| 9 | $0.4209^{*}$ | $0.4583^{*}$ |
| 10 | $0.6088^{*}$ | $0.6532^{*}$ |
| 11 | $0.7110^{*}$ | $0.7525^{*}$ |
| 12 | $0.5731^{*}$ | $0.5901^{*}$ |
| All | $0.4343^{*}$ | $0.5026^{*}$ |

Table 2: **Pearson correlation** coefficients between **cross-lingual transfer performance** and the **impact on the representation space of the target language**. ($^{*}p < 0.01$, two-tailed).

## 5 Does Selective Layer Freezing Allow to Improve Transfer to Linguistically Distant Languages?

In the previous section we observed an inter-correlation between cross-lingual transfer performance, the linguistic distance between the target and source language, and the impact on the representation space. Given this observation, we investigate the possibility to use this information to improve transfer to linguistically distant languages. More specifically, we hypothesize that it may be possible to regulate cross-lingual transfer performance by selectively interfering with the previously observed correlations at specific layers. A straightforward strategy would be to selectively freeze layers, during the fine-tuning process, where a significant negative correlation between the impact on their representation space and the distance between source and target languages has been observed. By freezing a layer, we manually set the correlation between the impact on the representation space and language distance to zero, which may simultaneously reduce the significance of the

| Exp. | Frozen Layers | SYN | GEO | INV | GEN | PHON | CLTP |
|---|---|---|---|---|---|---|---|
| | | **-0.7354** | **-0.5109** | **-0.4907** | **-0.6116** | **-0.5776** | **66.70** |
| A | {2} | -0.7310 | -0.5109 | -0.4791 | -0.6009 | -0.5791 | 66.53 |
| B | {5} | -0.7438 | -0.5053 | -0.4897 | -0.6148 | -0.5896 | 66.77 |
| C | {1,2,6} | -0.7325 | -0.5000 | -0,4846 | -0.6065 | -0.5666 | 66.75 |

Table 3: Pearson **correlation coefficients** quantifying the relationship between **cross-lingual transfer performance** and different **language distance metrics** after freezing different layers during fine-tuning. The first row contains baseline values for full-model fine-tuning. The last column provides the average cross-lingual transfer performance (CLTP), measured as accuracy, across all target languages. English has been the only source language.

correlation between language distance and transfer performance.

Wu and Dredze (2019) already showed that freezing early layers of mBERT during fine-tuning may lead to increased cross-lingual transfer performance. With the same goal in mind, Xu et al. (2021) employ meta-learning to select layer-wise learning rates during fine-tuning. In what follows, we will, however, not focus on pure overall transfer performance. Our approach is to specifically target transfer performance improvements for target languages that are linguistically distant from the source language, rather than trying to achieve equal transfer performance increases for all target languages.

### 5.1 Experimental Setup

For our pilot experiments, we focus on English as the source language. Additionally, we choose to carry out our pilot experiments on layers 1, 2, 5, and 6, as the representation space impact at these layers exhibits low correlation values with transfer performance (Table 2) and high correlations with different language distances (Figure 2 in Appendix C). This decision is made to mitigate the potential impact on the overall transfer performance, which could obscure the primary effect of interest, and to simultaneously target layers which might be responsible for the transfer gap to distant languages. We conduct 3 different experiments aiming to regulate correlations between specific language distances and transfer performance. In an attempt to diversify our experiments, we aim to decrease the transfer performance gap for both a single language distance metric (Experiment **A**) and multiple distance metrics (Exp. **C**). Furthermore, in another experiment we aim at deliberately increasing the transfer gap (Exp. **B**).

### 5.2 Results

Table 3 provides results of all 3 experiments.

**Experiment A.** The $2^{nd}$ layer shows a strong negative correlation (-0.66) between representation space impact and inventory distance to English. Freezing the $2^{nd}$ layer during fine-tuning has led to a less significant correlation between inventory distance and transfer performance (+0.0116).

**Experiment B.** The $5^{th}$ layer shows a strong positive correlation (0.499) between representation space impact and phonological distance to English. Freezing the $5^{th}$ layer during fine-tuning has led to a more significant correlation between phonological distance and transfer performance (-0.012).

**Experiment C.** The $1^{st}$ layer, $2^{nd}$ layer and $6^{th}$ layer show a strong negative correlation between the impact on the representation space and the syntactic (-0.618), inventory (-0.66) and phonological (-0.543) distance to English, respectively. Freezing the $1^{st}$, $2^{nd}$ and $6^{th}$ layer during fine-tuning has led to a less significant correlation of transfer performance with syntactic (+0.0029) and phonological (+0.011) distance.

## 6 Conclusion

In previous research, the effect of fine-tuning on a language representation space was usually studied in relative terms, for instance by comparing the cross-lingual alignment between two monolingual representation spaces before and after fine-tuning. Our research, however, focused on the absolute impact on the language-specific representation spaces within the multilingual space and explored the relationship between this impact and language distance. Our findings suggest that there is an intercorrelation between language distance, impact on the representation space, and transfer performance

which varies across layers. Based on this finding, we hypothesize that selectively freezing layers during fine-tuning, at which specific inter-correlations are observed, may help to reduce the transfer performance gap to distant languages. Although our hypothesis is only supported by three pilot experiments, we anticipate that it may stimulate further research to include an assessment of our hypothesis.

## Limitations

It is important to note that the evidence presented in this paper is not meant to be exhaustive, but rather to serve as a starting point for future research. Our findings are based on a set of 15 languages and a single downstream task and may not generalize to other languages or settings. Additionally, the proposed hypothesis has been tested through a limited number of experiments, and more extensive studies are required to determine its practicality and effectiveness.

Furthermore, in our study, we limited ourselves to using traditional correlation coefficients, which are limited in terms of the relationships they can capture, and it is possible that there are additional correlations that could further strengthen our results and conclusions.

## Ethics Statement

This study was designed to minimize its environmental impact by reducing the amount of required computational resources to run our experiments. We are aware of the high energy consumption and carbon footprint associated with large-scale machine learning experiments and took steps to minimize these impacts.

Additionally, in this study, our objective was to address the performance gap in languages that are underrepresented in comparison to high-resource languages, rather than solely striving for performance enhancement.

## References


Tyler Chang, Zhuowen Tu, and Benjamin Bergen. 2022. The Geometry of Multilingual Language Model Representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chris Collins and Richard Kayne. 2011. *Syntactic Structures of the World's Languages*. New York University, New York.

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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the Best of Cross-lingual Transfer: Evidence from POS Tagging with over 100 Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. 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.

Błażej Dolicki and Gerasimos Spanakis. 2021. Analysing The Impact Of Linguistic Features On Cross-Lingual Transfer. ArXiv:2105.05975 [cs].

Matthew S. Dryer and Martin Haspelmath. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not Greek to mBERT: Inducing Word-Level Translations from Multilingual BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 45–56, Online. Association for Computational Linguistics.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2015. *Glottolog 2.6*. Max Planck Institute for the Science of Human History, Jena.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*,

pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529. PMLR.

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.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig. 2015. *Ethnologue: Languages of the World, Eighteenth edition*. SIL International, Dallas, Texas.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the Language Neutrality of Pre-trained Multilingual Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What Happens To BERT Embeddings During Fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Steven Moran, Daniel McCloy, and (eds.). 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First Align, then Predict: Understanding the Cross-Lingual Ability of Multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Yulia Otmakhova, Karin Verspoor, and Jey Han Lau. 2022. Cross-linguistic Comparison of Linguistic Feature Encoding in BERT Models for Typologically Different Languages. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 27–35, Seattle, Washington. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. BERT is Not an Interlingua and the Bias of Tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, Hong Kong, China. Association for Computational Linguistics.

Marc Tanti, Lonneke van der Plas, Claudia Borg, and Albert Gatt. 2021. On the Language-specificity of Multilingual BERT and the Impact of Fine-tuning. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 214–227, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 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.

Weijia Xu, Batool Haider, Jason Krone, and Saab Mansour. 2021. Soft Layer Selection with Meta-Learning for Zero-Shot Cross-Lingual Transfer. In *Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing*, pages 11–18, Online. Association for Computational Linguistics.

## A Technical Details

### A.1 Data

We perform our experiments on the XNLI (Conneau et al., 2018) dataset[4]. The dataset contains 392.702 train, 2.490 validation and 5.010 test samples, derived from the English-only MultiNLI (Williams et al., 2018), which have been translated to Arabic (ar), Bulgarian (bg), German (de), Greek (el), Spanish (es), French (fr), Hindi (hi), Russian (ru), Swahili (sw), Thai (th), Turkish (tr), Urdu (ur), Vietnamese (vi) and Chinese (zh). The objective of the dataset is to evaluate a model's capability of classifying the relationship between two sentences, namely a premise and a hypothesis, as entailment, contradiction, or neutral.

The dataset has been released under a *Creative Commons Attribution Non Commercial 4.0 International*[5] license (CC BY-NC 4.0).

### A.2 Model

We use the base cased multilingual BERT (Devlin et al., 2019) model, which has 12 attention heads and 12 transformer blocks with a hidden size of 768. The dropout probability is 0.1. The model has 110M parameters and covers 104 languages. Its vocabulary size is about 120k.

### A.3 Training

We fine-tune the models using the HuggingFace Transformers (Wolf et al., 2020) and PyTorch (Paszke et al., 2019) frameworks. We use AdamW (Loshchilov and Hutter, 2019) as an optimizer, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$. We train for 3 epochs with a batch size of 32 and an initial learning rate of $2e^{-5}$ with linear decay. Full model fine-tuning on a single language took about 2.5 hours on a single NVIDIA® V100 GPU. Total GPU hours for all 18 fine-tuned models (15 and 3 in Sections 4 and 5 respectively was about 45 hours.

In order to minimize computational costs and reduce our environmental impact, we chose not to conduct a full hyper-parameter search and instead used the fixed values reported in Section 3.1.

For reproducibility, our code is provided here: https://anonymous.4open.science/r/sigtyp2023_workshop_paper-223F.

[4] https://github.com/facebookresearch/XNLI

[5] https://creativecommons.org/licenses/by-nc/4.0/

## B Additional Information on Language Distance Metrics

We used the following lang2vec distances:

1. **Syntactic Distance** is the cosine distance between the syntax feature vectors of languages, sourced from the World Atlas of Language Structures.[6] (WALS) (Dryer and Haspelmath, 2013), Syntactic Structures of World Languages[7] (SSWL) (Collins and Kayne, 2011) and Ethnologue[8] (Lewis et al., 2015).

2. **Geographic Distance** refers to the shortest distance between two languages on the surface of the earth's sphere, also known as the orthodromic distance.

3. **Inventory Distance** is the cosine distance between the inventory feature vectors of languages, sourced from the PHOIBLE[9] database (Moran et al., 2019).

4. **Genetic Distance** is based on the Glottolog[10] (Hammarström et al., 2015) tree of language families and is obtained by computing the distance between two languages in the tree.

5. **Phonological Distance** is the cosine distance between the phonological feature vectors of languages, sourced from WALS and Ethnologue.

The values range from 0 to 1, where 0 indicates the minimum distance and 1 indicates the maximum distance.

## C Additional Figures

Figure 2 provides **Pearson correlation coefficients** between the **impact on the target language representation space** when fine-tuning in **English** and different types of **linguistic distances between English and the target language** for each layer. English-English data points were excluded in order to prevent an overestimation of effects.

Figure 3 contains the cross-lingual zero-shot transfer results. The numbers illustrated in the figure represent accuracies.

[6] https://wals.info

[7] http://sswl.railsplayground.net/

[8] https://www.ethnologue.com/

[9] https://phoible.org/

[10] https://glottolog.org

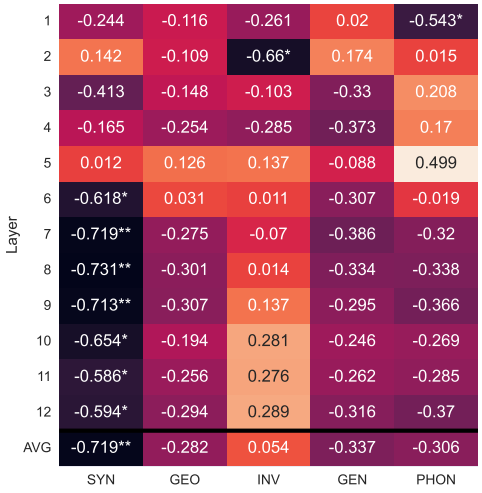


Figure 2: Pearson correlation coefficients between the impact on the representation space and different types of linguistic distances (with English as the only source language). ($^{*}p < 0.05$, and $^{**}p < 0.01$, two-tailed).

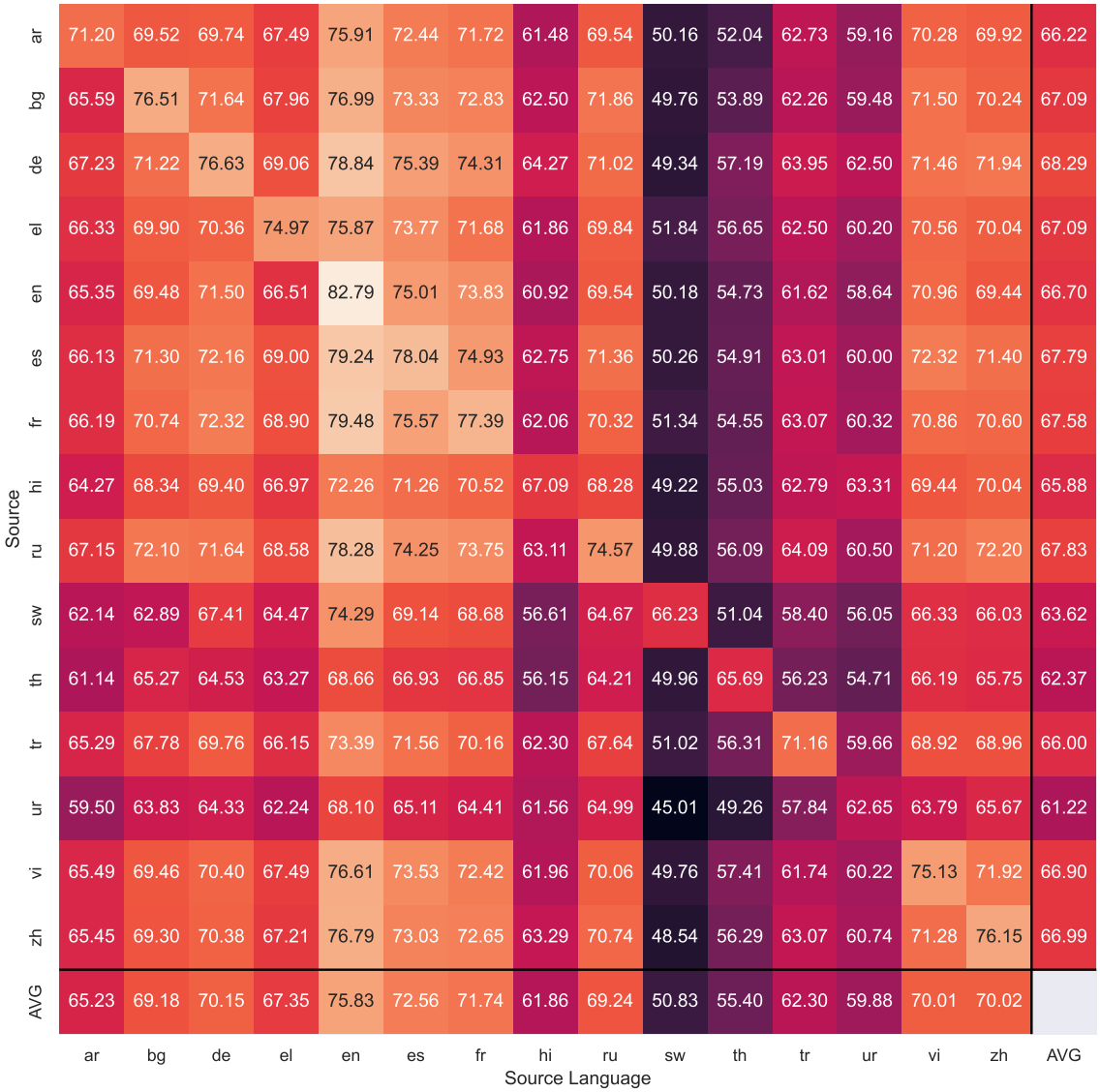


Figure 3: Cross-lingual zero-shot transfer results for XNLI