Language Complexity in Multilingual Language Models

Anonymous ACL submission

Abstract

Understanding the behavior of multilingual language models across different languages remains a significant challenge. Recent research has demonstrated that the Average Neuron-Wise Correlation (ANC) offers a comprehensive analysis of activation similarities in multilingual models. This study proposes the use of Average Wasserstein Distance (AWD) between activation value distributions and compares it to ANC across three datasets: XNLI, ReadMe++, and Vikidia. By applying these metrics, we aim to elucidate the underlying processes within large language models, thereby enhancing our understanding of cross-linguistic transfer and model accuracy.

1 Introduction

This paper explores methods to investigate cross-linguistic transfer, focusing on differences in activation patterns of Pre-trained Language Models (PLMs) like BERT (Devlin et al., 2019a) and RoBERTa (Conneau et al., 2019) when classifiers are transferred to a new target language. Specifically, we address the issue of language complexity, aiming to assess whether a text is complex.

Better understanding of the cross-lingual transfer capabilities of PLMs is essential for improving multilingual NLP, particularly in tasks requiring language adaptation, such as readability assessment and complexity classification. A key challenge lies in determining how well PLMs generalize linguistic structures across diverse languages, especially when dealing with underrepresented or typologically distant languages. Investigating activation patterns through ANC and AWD provides insights into how PLMs encode and transfer linguistic complexity, allowing us to identify factors that facilitate or hinder successful cross-lingual adaptation.

Previous research has primarily focused on the advantages of studying the transfer gap using the Average Neuron-Wise Correlation (ANC) metric over other methods, such as Centered Kernel Alignment (Del and Fishel, 2022), as well as by using layer ablation (Muller et al., 2021).

Our main contributions are:

1. Application of the Average Wasserstein Distance (AWD) to study latent spaces of multilingual PLMs;
2. Examination of the behavior of ANC on language complexity tasks;
3. Combining ANC and AWD for a more detailed explanation of hidden processes in PLM activations.

The code and data will be available in the accepted version of the paper.

2 Related Work

2.1 Analysis of Pretrained Language Models

Recent research has explored the internal mechanics of PLMs, particularly the behavior of hidden layers. Muller et al. (2021) introduced the ”first align, then predict” approach, where activation vectors are aligned before predictions are made. To analyze these hidden spaces, methods like Centered Kernel Alignment (CKA, (Kornblith et al., 2019)) and Canonical Correlation Analysis (CCA, (Hotelling, 1936)) have been used.

Del and Fishel (2022) analysed limitations of such methods as CCA and CKA and introduced Average Neuron-Wise Correlation (ANC), which has been shown to offer significant advantages over CKA and CCA, providing a more interpretable framework for analyzing the alignment of activation vectors across layers, particularly in the context of the XNLI task.

The idea of measuring cross-lingual transfer across the layers is also important for decoding-only language models such as Llama (Liu et al., 2025). However, as our concern here is with classification tasks, we keep investigating smaller and more efficient encoder-only PLMs.

2.2 Wasserstein Distance

The Wasserstein distance, or Earth Mover's Distance, measures the cost of transporting mass between two probability distributions, offering a comprehensive view of their similarity (Khamis et al., 2024). It is particularly useful for detecting shifts in word embedding distributions at the coordinate level (Ramírez et al., 2020). By using Wasserstein distance alongside ANC, we aim to gain a fuller understanding of neural network activation distributions.

3 Methodology

To investigate the process of language transfer, we compare the use of ANC and AWD over the values of activations in the PLM layers comparing predictions made for the original language (L_1) and for the target language (L_2) for semantically equivalent text segments:

$$X = L_1 - \text{mean}(L_1)$$

$$Y = L_2 - \text{mean}(L_2)$$

3.1 ANC definition

ANC calculates the correlation between activation vectors averaged by neurons:

$$ANC(X,Y) = \frac{\sum\limits_{i=1}^{n} abs(corr(X[i], Y[i]))}{n}, \quad (1)$$

where *corr* is Pearson's correlation.

3.2 Average Wasserstein definition

Given that ANC focuses only on the angle between vectors (as it is equivalent to the cosine distance), this does not fully capture the variation of the coordinates. This is the rationale for our proposal of the Average Wasserstein Distance (AWD), which is defined by mirroring ANC:

$$AWD(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left[W(X[i], Y[i])\right], \quad (2)$$

where W(X[i], Y[i]) is the Wasserstein Distance between i-th neuron of centered layer of L_1 and i-th neuron of centered layer of L_2:

$$W(a,b) = \sum_{j=0}^{k} |F_a(t_j) - F_b(t_j)|\, \Delta t \quad (3)$$

with Δt and F_a defined as:

$$\Delta t = \left[\max(a \cup b) - \min(a \cup b)\right]/k \quad (4)$$

$$t_j = \min(a \cup b) + j\Delta t \quad (5)$$

$$F_a(t) = \sum_{u \in a} I(u < t) / \sum_{u \in a} 1 \quad (6)$$

3.3 Complexity datasets

In addition to the XMLI task as used in (Del and Fishel, 2022), we examined two scenarios of complexity prediction: one with six classes and another one with two. For the six classes, we used ReadMe++ (Naous et al., 2023), which contains texts in five languages (English, French, Russian, Arabic, and Hindi), categorized into six complexity levels according to CEFR Council of Europe (2001). For the second classification task, we collected data from Vikidia[1], a website which maintains Wikipedia-style content aimed at “children and anyone seeking easy-to-read content”, and the corresponding entries from Wikipedia, aiming to predict whether a test text is suitable for Vikidia or not.

Initially, neither dataset included segments of the same meaning across languages. For the XNLI task, such texts were prepared using machine translation and human quality control (Conneau et al., 2018). To create the necessary parallel corpus for the complexity task, we translated the available English texts into Russian (ru), French (fr), Hindi (hi), Arabic (ar), Romanian (ro), Greek (el), Spanish (es), Hebrew (he), Turkish (tr), Belarusian (be), Ukrainian (uk), Bulgarian (bg), Chinese (zh), Japanese (ja), Irish (ga), German (de), Italian (it) and Welsh (cy) using Google Translate.

To ensure that the language complexity does not change through machine translation, we verified a small sample of translated texts by human annotation with no detectable changes in complexity. On the full dataset, we compared transfer learning on the original texts and on the translations. Table 1 shows that complexity predictions are preserved, with MAE scores even improving on the translated texts. This improvement may result from reduced annotation noise, as different annotators originally

[1] https://www.vikidia.org/

	Original ReadMe++	Translated ReadMe++
en	0.418	0.418
ar	0.669	0.716
fr	0.535	0.619
hi	0.552	0.643
ru	0.533	0.668

Table 1: Comparing mean absolute error (MAE) of predictions on the original ReadMe++ and on the translated English (hold out) part. To improve the stability of the results, the value in the table is obtained as an average over 10 batches of 100 elements selected randomly.

	XNLI	ReadMe++	Vikidia vs Wikipedia
en	0.536	0.602	0.851
ar	0.517	0.507	0.748
fr	0.532	0.542	0.785
hi	0.522	0.343	0.622
ru	0.531	0.544	0.707

Table 2: Average F1 weighted measure for BERT trained on English (averaging over 10 batches containing 100 texts each. All sampled from the hold out dataset).

marked the languages, potentially leading to varying interpretations of CEFR criteria. The same labels were applied to both the original and translated texts.

The use of the MAE metric allows us to quantitatively evaluate the limits of the model errors (in particular, to say that the model is critically wrong or wrong within the neighboring complexity levels), in contrast to the F1 metric, which only allows us to detect the presence of an error, but not to evaluate its criticality.

3.4 Transfer experiments

To investigate the transfer process, we trained our model on English texts using multilingual BERT and XLM-Roberta on different datasets and then made predictions on unrelated test texts in other languages (300 samples).

To understand the multilingual limits of this transfer, we have also conducted similar experiments to show transfer from other languages using the ReadMe dataset. For all experiments with XNLI and Viki use F1 also for Readme, but Readme — only experiments always use MAE. This is due to the fact that in the ReadMe dataset, an error of ±1 class is not as significant as an error of ±2 or more. This feature is not taken into account by the F1 metric, unlike MAE.

We measured ANC and AWD across the layers on these test texts and reported the interquartile range of their values as a robust measure of their variation.

For the sake of brevity, we only present the results for BERT, with the XLM-Roberta results included in the Appendix. These experiments were conducted in parallel for the XNLI task and for two text complexity prediction tasks, the primary focus of this study.

4 Experimental results

Table 2 presents the validation metrics for the trained model. On the task of classification of texts in ReadMe, it is clear how much the quality of transfer from English differs. Considering that BERT was trained on Wikipedia, and the task of HNLI is similar to the one on which BERT was trained, we can conclude that fine-tuning plays a greater role on the ReadMe task, rather than pre-training.

The first row of Figure 1 shows that ANC follows the pattern observed by Del and Fishel (2022) for XNLI, extending it to two additional tasks: ANC increases until the middle layers, then it plateaus.

The phenomenon of “first align, then predict” was first observed by Muller et al. (2021) using a layer ablation technique. Del and Fishel hypothesized that at layers where ANC increases, activation vectors “level out”, while prediction occurs at subsequent layers. Unlike ANC (a similarity measure), AWD is a distance. So it follows a reverse pattern, see the second row of Fig. 1: from layer 1 to 7, where the directions of activation vectors change significantly, the distribution of their coordinate values is close to the source language. However, instead of “leveling out”, as detected by ANC, a change occurs in the distribution of values. In complexity classification problems, the effect is more pronounced than in XNLI. Thus, the alignment phase consists of bringing the direction of the activation vectors closer to what was in the source language (thus, the model brings the unknown to the known). Later, the prediction stage consists of changing the distributions of activation values (the model highlights the differences). We hypothesize that the plateau observed through ANC is associated with its limited ”horizon” (as it only looks at the angles between the vectors), rather than with the absence of processes that change activations.

Pictures with aggregated data do not fully re-

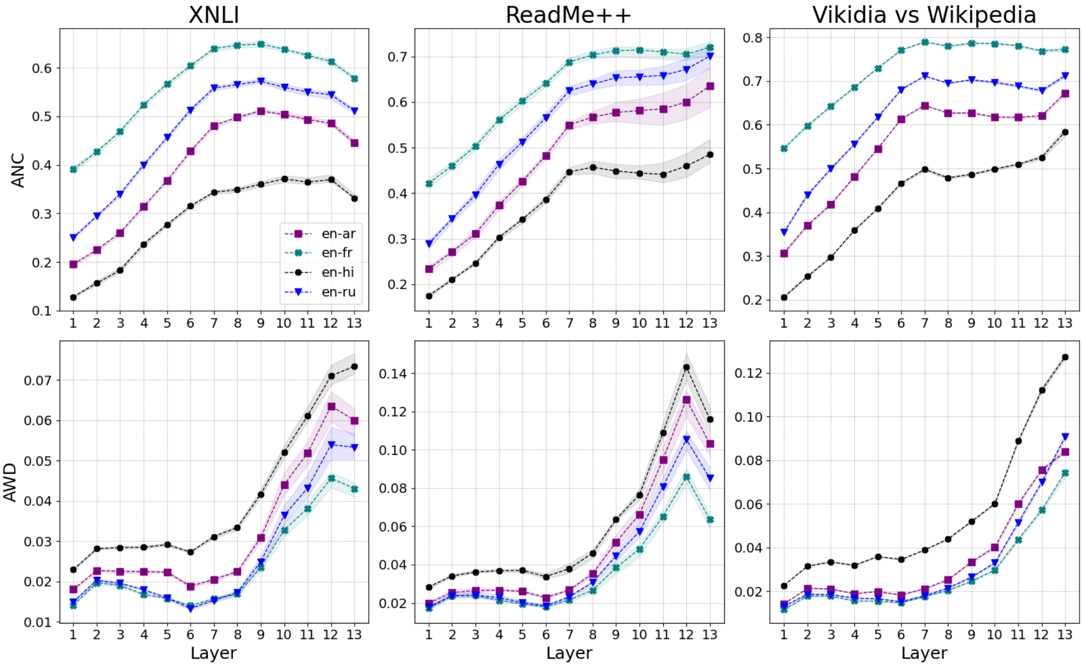

Figure 1: ANC and AWD for BERT. Activations were counted on 300 samples. The dot is the average value on the sample, the pale colored area shows interquartile range on activations in each layer.

flect how the quality metric depends on the distance measures. Figure 6 shows the relationship between weighted F1 and ANC (or AWD) on the last layer. The arrangement of the distributions relative to each other is different in XNLI vs our complexity tasks. Despite that, in all tasks the pattern is maintained: languages more similar to English have higher values on the ANC scale. For AWD the order is reversed, as it is a distance measure. For example, either ANC or AWD can be used to predict the transfer of complexity predictions to Hindi is likely to be less successful than transfer to French, see Table 1.

We can also use these measures to detect deviations from this pattern for individual examples. For example, the following French sentence from Readme++ *Aujourd'hui, la situation n'a fait qu'empirer.* "Today, the situation has only worsened." differs in syntax from English, while its translations into other languages in the Readme++ set are similar to English. We quantified translation similarity using the metric $(v - b)/b$ where b is the mean AWD across layers 1-7 (baseline) and v is the mean AWD across layers 7-13 (value). The results for this sentence span from 0.65 for Arabic and 1.17 for Russian to 4.05 for French and 5.98 for Hindi. In contrast, the average values across the entire sample range from 1.41 for French and 1.82 for Arabic to 1.89 for Russian and 3.42 for Hindi. In all cases, class 1 was predicted (which corresponds to the correct one). The abnormally high value for Hindi is likely to be due to worse representation quality from BERT's initial training sample.

5 Differences between ANC and AWD

To better understand how the linguistic properties of information representation are captured by these metrics, we calculated the syntactic distances (Belov et al., 2020) between the original sentences of the dataset and their translations. The edit distance metric was calculated on syntactic trees of sentences, where each node is a part of speech of a word from the sentence, and an edge is a type of connection between words[2]. The edit distance metric is directly proportional to the number of insertions/replacements/removals of a node or edge in one tree to obtain the original. Therefore, languages with syntactic constructions less similar to those found in English sentences have a higher value.

Figure 2 shows that in the last layer the AWD metric shows better correlation with syntactic edit distance in comparison to f1 and ANC.

[2] Using pretrained models from Stanza https://stanfordnlp.github.io/stanza/depparse.html

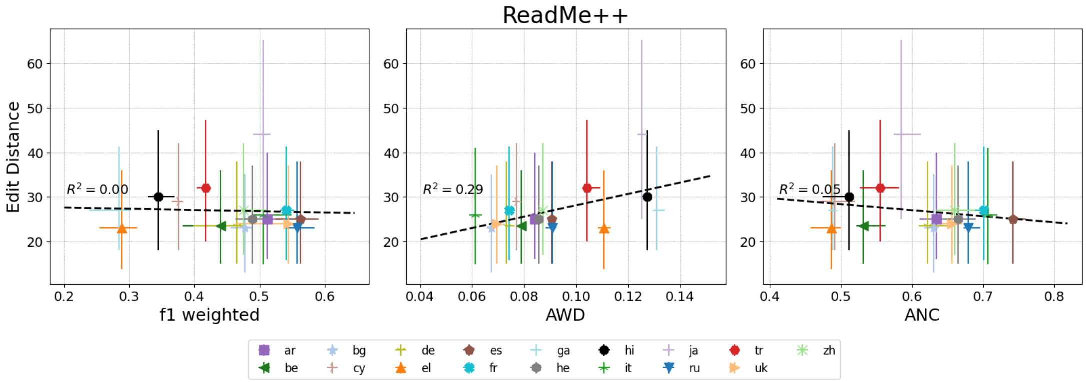

Figure 2: Dependence of syntactic edit distance on ANC and AWD at the last layer of the BERT model (trained on the English dataset) and the accuracy of prediction on the translated dataset.

6 Extended experiment

Experiments in which the model is trained in English and then makes predictions in other languages have shown significant correlations between F1 and ANC/AWD.

Figure 3 shows the performance of the model trained on the English dataset in all languages under consideration. It is noticeable that in all cases the correlation coefficients are high and there is an obvious dependence in the data. This behavior was observed in the original ReadMe datasets (5) and then generalized by translating its English part.

However, this effect is observed only for models trained on the English sample. For models trained on other languages (fig 4), the transfer effect suffers significantly. Moreover, in languages with a large resource base (those that were represented in large quantities in the BERT training sample), the disorder is noticeable, but not as strong as in Arabic and Hindi, where the dots represent a random cloud.

However, the model trained on other languages shows significantly more chaotic results 4.

7 Conclusions

In this study, we explored the cross-linguistic transfer of language complexity using PLMs, with a particular emphasis on the ANC metric and the introduction of the Wasserstein distance as a complementary measure. Our experiments, conducted with both BERT and xlm-roBERTa models, yielded several insights:

- **Metric Comparison**: the ANC metric alone may not fully capture the subtleties of activation similarities across languages. By incorporating the Wasserstein distance, we offered a more comprehensive perspective on these similarities, particularly in capturing distributional shifts that the ANC metric might overlook.

- **Language Complexity Transfer**: Our results indicate that language complexity, as measured by F1 scores, is largely preserved during machine translation, validating the use of translated texts in cross-linguistic studies.

In conclusion, combining ANC with AWD provides a more nuanced understanding of activation similarities in multilingual models. These insights not only enhance our understanding of cross-linguistic transfer but also have practical implications for developing more effective and accurate multilingual NLP systems.

Future Research Directions:

- Our combined approach highlighted a correlation between these metrics and model performance, suggesting that higher ANC and lower AWD are associated with better model accuracy, so AWD can can be used to estimate accuracy for new languages even without test data. This initial finding needs to be explored with more language pairs.

- Our analysis has shown that AWD is linked to syntactic differences between languages. However, the specific linguistic or computational properties that determine ANC remain to be systematically identified. Future research should aim to establish a precise link between ANC and relevant linguistic factors, which will contribute to a more refined understanding of multilingual model adaptation.

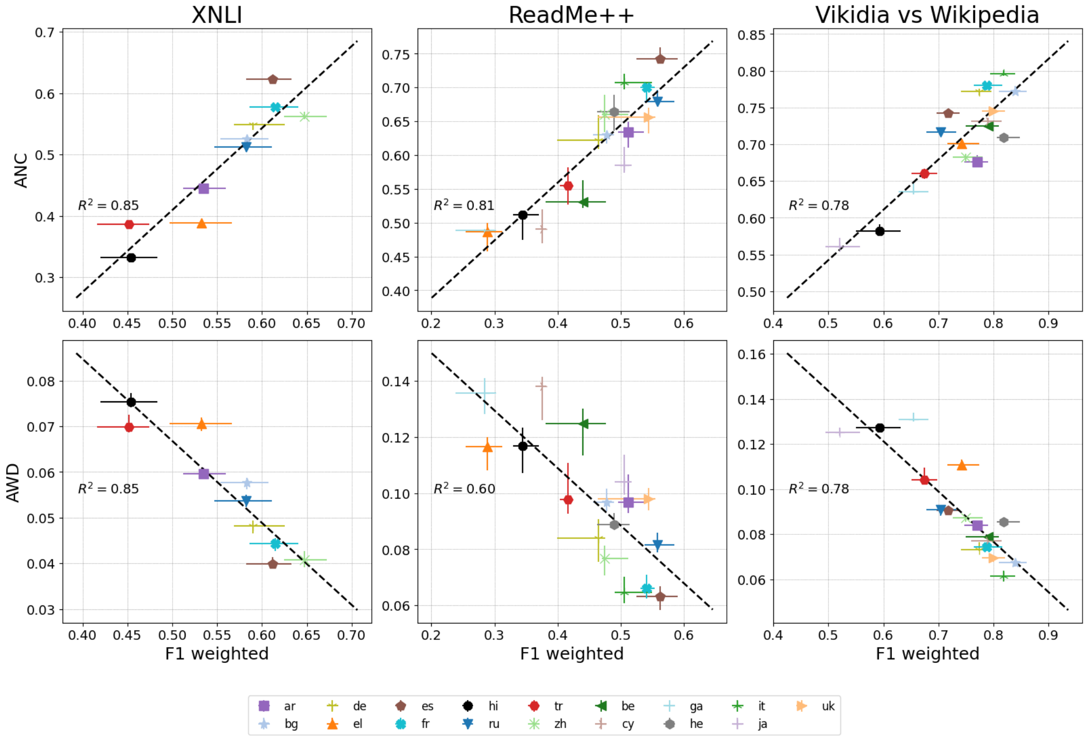

Figure 3: ANC and AWD at the last layer of the BERT model vs F1 on extended translated dataset.

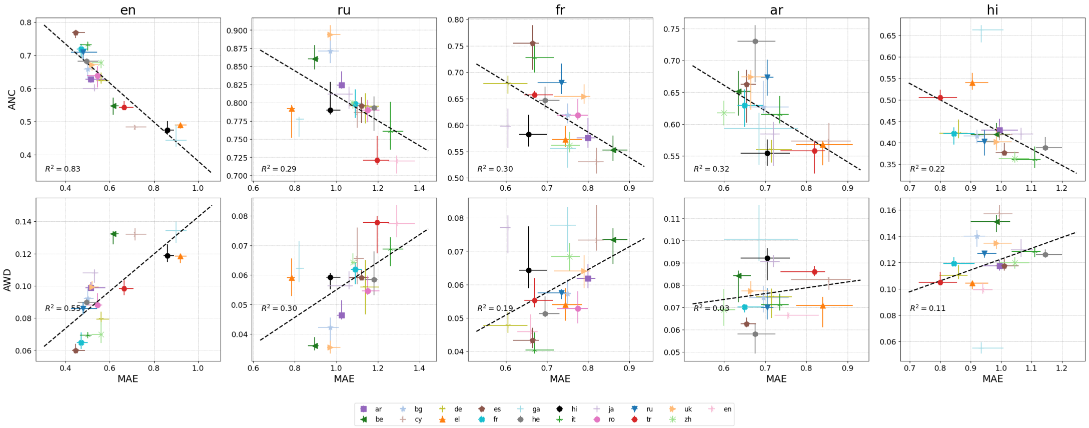

Figure 4: ANC and AWD at the last layer of the BERT model vs MAE on extended translated dataset. The columns show the results of inference of models that were trained on the original ReadMe++ dataset in English, French and Hindi, respectively.

8 Limitations

The obtained regularities are applicable only when transferring training from English to other languages.

9 Ethical statement

We are not aware of potential ethical risks in the study discussed in the paper. If anything it helps with understanding the process of cross-lingual transfer, thus potentially helping lesser-resourced languages.

In conducting the study we have been careful with the environmental impact of NLP research. Large Language Models are more computationally expensive, while they have been shown to be not better than PLMs in several text classification tasks. For each of the methods we provide estimates the computational costs of running the models (table 3).

	Time per one loop	Number of loops	Number of models	Total computer time
Train	0.5	11	5	27.5
Inference	0.1	30	5	15

Table 3: Computational costs of running the models on Google Colab's L4 GPU (in hours).

10 Bibliographical References

References

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255.

Sergey Belov, Daria Zrelova, Petr Zrelov, and Vladimir Korenkov. 2020. Overview of methods for automatic natural language text processing. *System Analysis in Science and Education*, pages 8–22.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*.

Maksym Del and Mark Fishel. 2022. Cross-lingual similarity of multilingual representations revisited. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 185–195, Online only. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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*, pages 4171–4186, Minneapolis, Minnesota.

Harold Hotelling. 1936. RELATIONS BETWEEN TWO SETS OF VARIATES*. *Biometrika*, 28(3-4):321–377.

Abdelwahed Khamis, Russell Tsuchida, Mohamed Tarek, Vivien Rolland, and Lars Petersson. 2024. Scalable optimal transport methods in machine learning: A contemporary survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

Weihao Liu, Ning Wu, Wenbiao Ding, Shining Liang, Ming Gong, and Dongmei Zhang. 2025. Selected languages are all you need for cross-lingual truthfulness transfer. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8963–8978, Abu Dhabi, UAE. Association for Computational Linguistics.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. 2018. Insights on representational similarity in neural networks with canonical correlation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 5732–5741, Red Hook, NY, USA. Curran Associates Inc.

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.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2023. Readme++: Benchmarking multilingual language models for multi-domain readability assessment. *arXiv preprint arXiv:2305.14463*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8).

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017a. Svcca: singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6078–6087, Red Hook, NY, USA. Curran Associates Inc.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017b. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pages 6078–6087.

Guillem Ramírez, Rumen Dangovski, Preslav Nakov, and Marin Soljačić. 2020. On a novel application of wasserstein-procrustes for unsupervised cross-lingual learning. *arXiv preprint arXiv:2007.09456*.

Serge Sharoff, Reinhard Rapp, and Pierre Zweigenbaum. 2023. *Building and Using Comparable Corpora for Multilingual Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Springer Nature.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

A Transferring language complexity with BERT

The figure 5 shows the aggregated values for the accuracy and proximity metrics of multilingual learning transfer.

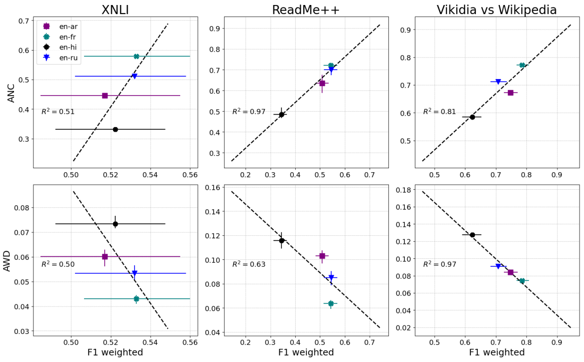

Figure 5: Mean value of ANC and AWD in the last layer of BERT. As the error bars, the first and third quartiles as minimum and maximum respectively.

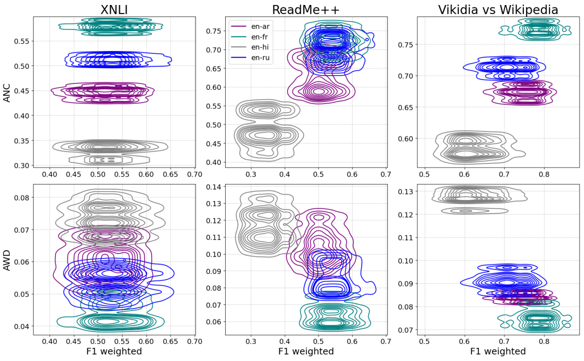

Figure 6: Distributions of proximity and accuracy matrices for BERT. The top row: distribution of ANC × weighted F1 (Cartesian product), the bottom row: AWD × weighted F1. Both ANC and AWD were taken from the last layer.

The difference in the behavior of the model in XNLI and complexity tasks is clearly visible. In the XNLI there is practically no correlation between F1 in ANC (AWD). While in both complexity problems, values are well described by a linear dependence. It follows that it is possible to use proximity metrics to predict the quality of a model in a language in which it has not been fine-tuned.

B Transferring language complexity with xlm-roBERTa

The same experiment was conducted with xlm-roBERTa. Since BERT and roBERTa were trained on different language domains (Conneau et al., 2019) they have different starting packs of language knowledge. On validation after fine-tuning, roBERTa showed similar results with BERT.

	XNLI	**ReadMe++**	**Vikidia vs Wikipedia**
ar	0.692	0.472	0.870
en	0.790	0.573	0.904
fr	0.746	0.518	0.871
hi	0.601	0.509	0.845
ru	0.716	0.465	0.858

Table 4: Average f1 weighted measure for xlm-roBERTa (averaging over 10 batches containing 100 texts each).

The expectation was that both models would show close results in the meaning of ANC and AWD metrics. However, the experiment showed the following pictures.

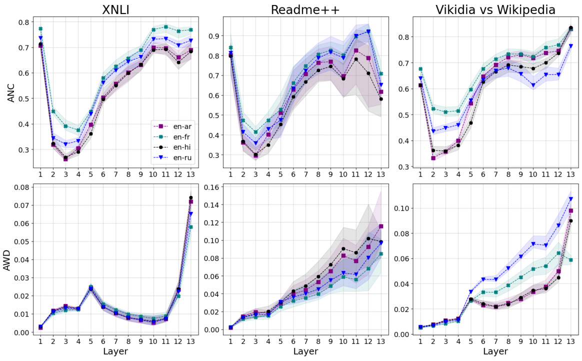

Figure 7: ANC and AWD for xlm-roBERTa obtained by sampling on parallel texts of the datasets. Activations were counted on 300 samples. The dot is the average value on the sample, the pale colored area shows interquartile range on all subsamples for one layer.

The behavior of the metrics in the languages studied is almost indistinguishable. This is probably due to the fact that the xlm-roBERTa saw a more balanced dataset across languages t the main training stage.

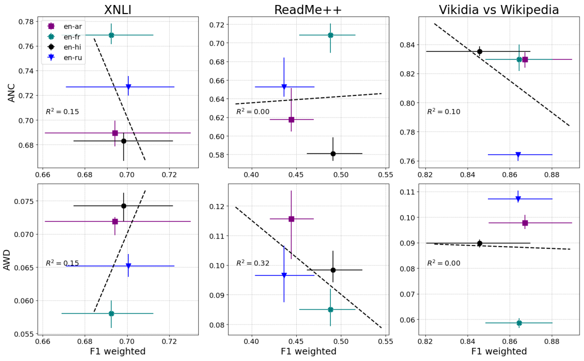

Figure 8: Mean value of ANC and AWD in the last layer of xlm-roBERTa. As the error bars, the first and third quartiles as minimum and maximum respectively.

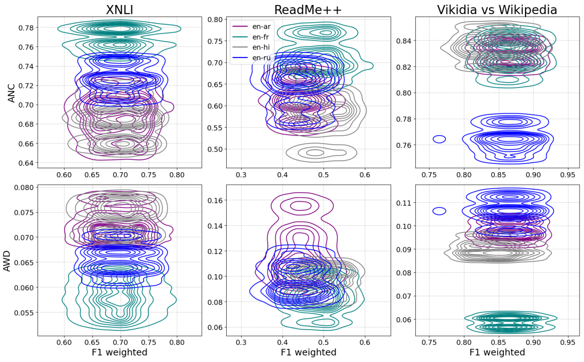

Figure 9: Distributions of proximity and accuracy matrices for xlm-roBERTa. The top row: distribution of ANC × f1 weighted (Cartesian product), the bottom row: AWD × f1 weighted. Both ANC and AWD were taken from the last layer.

C Appendix statistics of used datasets

Dataset	**Type**	**Size**	**Median length in words**	**Interquartal distance in words**
ReadMe	train en	2526	17	14
	test en	296	17	16
	test ar	296	15	13
	test fr	296	19	19
	test hi	296	19	18
	test ru	296	14	14
Vikidia	train en	2000	40	7
	test en	640	40	7
	test ar	640	32	9
	test fr	640	41	11
	test hi	640	41	10
	test ru	640	32	8
Wikipedia	train en	2000	40	0
	test en	640	40	0
	test ar	640	35	6
	test fr	640	44	6
	test hi	640	43	6
	test ru	640	34	5

Table 5: Statistics by words in datasets, used for training and analysis.