# Cross-lingual Similarity of Multilingual Representations Revisited

**Maksym Del** and **Mark Fishel**
Institute of Computer Science
University of Tartu, Estonia
{maksym,mark}@tartunlp.ai

## Abstract

Related works used indexes like CKA and variants of CCA to measure the similarity of cross-lingual representations in multilingual language models. In this paper, we argue that assumptions of CKA/CCA align poorly with one of the motivating goals of cross-lingual learning analysis, i.e., explaining zero-shot cross-lingual transfer. We highlight what valuable aspects of cross-lingual similarity these indexes fail to capture and provide a motivating case study *demonstrating the problem empirically*. Then, we introduce *Average Neuron-Wise Correlation (ANC)* as a straightforward alternative that is exempt from the difficulties of CKA/CCA and is good specifically in a cross-lingual context. Finally, we use ANC to construct evidence that the previously introduced "first align, then predict" pattern takes place not only in masked language models (MLMs) but also in multilingual models with *causal language modeling* objectives (CLMs). Moreover, we show that the pattern extends to the *scaled versions* of the MLMs and CLMs (up to 85x original mBERT).[1]

## 1 Introduction

Similarity indexes like Canonical Correlation Analysis (CCA, Hotelling, 1936) or Centered Kernel Alignment (CKA, Kornblith et al., 2019) aim to find a similarity between parallel sets of different representations of the same data. The deep learning community adapted these indexes to measure similarity between representations that *come from different models* (Raghu et al., 2017; Morcos et al., 2018; Kornblith et al., 2019). Another line of work used the same methods to measure similarity *between different languages* which come from a *single* multilingual model (Kudugunta et al., 2019; Singh et al., 2019a; Conneau et al., 2020; Muller et al., 2021).

In this paper, we argue that while CCA/CKA methods are a good fit for the first case, they are a suboptimal choice for the second scenario.

First, we employ a real-world motivating example to demonstrate that CKA can fail to capture the notion of similarity that we consider helpful in a cross-lingual context. We also discuss the general problems of CKA/CCA indexes and conclude that they are not well aligned with some of the goals of cross-lingual analysis (**Section 4**).

Next, we propose and verify an Averaged Neuron-Wise Correlation (ANC) as a straightforward alternative. It exploits the fact that representations from the same model have apriori-aligned neurons, which is the desired property in a cross-lingual setup (**Section 5**).

Finally, Muller et al. (2021) demonstrated the so-called "first align, then predict" representational pattern in a multilingual model: the model first aligns representations of different languages together, and then (starting from the middle layers) makes them more language-specific again (to accompany the language-specific training objective). The finding is insightful but only considers mBERT (Wu and Dredze, 2019) which is a masked language model (MLM) with 110M parameters. Thus, it is unclear if the "first align, then predict" pattern is specific to this model or more general. In this study, we use ANC to show that the pattern generalizes to the GPT-style (Brown et al., 2020) causal language models (CLMs, Lin et al., 2021) and extends to *large-scale* MLMs and CLMs (**Section 6**).

In this paper we are interested specifically in the scenario of measuring the strength of cross-lingual similarity of representations that come from a single multilingual language model. This scenario is very common in the field as it is often not feasable to train a separate models for each language and we present a method that allows for better representational similarity analysis then CKA/CCA.

---

[1] Our code is publicly available at https://github.com/TartuNLP/xsim

In summary, our contributions are three-fold:

- conceptual and *empirical* critique of CKA/CCA for cross-lingual similarity analysis (**Section 4**);

- *Average Neuron-Wise Correlation* as a simple alternative method designed specifically for cross-lingual similarity (**Section 5**);

- *scaling laws* of cross-lingual similarity in both multilingual MLMs and CLMs (**Section 6**).

## 2   Related work

Hotelling (1936) introduced CCA as a method for measuring canonical correlations between two sets of random variables. Raghu et al. (2017) proposed a variant of the CCA called SVCCA and used it to analyze representations *between different neural networks*. Morcos et al. (2018) proposed PWCCA, another improvement to CCA for the network analysis, and Kornblith et al. (2019) analyzed CCA, SVCCA, PWCCA, and other methods concluding that CKA is superior to them.

In a cross-lingual setting, we have a single network, and we compare representations that come from different languages. Following the introduction of SVCCA, Kudugunta et al. (2019) used it to compare language representations (at different layers) in a multilingual neural machine translation system. The method we present in this work applies to the seq2seq models, but in this work, we focus on models trained with CLM and MLM objectives while leaving seq2seq for future work. Singh et al. (2019a) performed a similar study where they focused on the multilingual BERT model[2] and employed PWCCA as a similarity index. The conclusion was that language representations diverge with network depth.

On the other hand, Conneau et al. (2020) and Muller et al. (2021) used CKA and behavior analysis to show that the opposite pattern takes place: language representations align with the network depth and only moderately decrease towards the end. In other words, representations first converge towards language neutrality and then recover some language-specificity. The alignment makes zero-shot cross-lingual transfer possible, and slight divergence accompanies language-specific training objectives (such as English downstream prediction

task or predicting words in the particular language as in masked language modeling objective). Following Muller et al. 2021, we call this phenomenon the *"first align, then predict"* pattern.

Eventually, Del and Fishel (2021) showed that the similarity analysis was different because Singh et al. (2019a) used CLS-pooling while Muller et al. (2021) used mean-pooling to convert token embeddings into a sentence representation. They also showed that mean-pooling is a better option.

Finally, Li et al. (2015) aligned most correlated neurons between layers of two different networks and then computed similarity from the recovered correspondence. The method we propose in this paper is similar in spirit to this one, except we focus on the cross-lingual analysis of multilingual models and thus have no need to find the alignment between neurons.

In this work, we build on these studies in three ways: we demonstrate that even CKA can fail to provide relevant cross-lingual similarity, we propose another method to compare multilingual representations, and we reveal that the "first align, then predict" pattern generalizes across training objectives and holds for models of large sizes.

## 3   Similarity Indexes Background

In this section, we provide some background on CKA and CCA, SVCCA, and PWCCA similarity indexes[3]. We focus on the parts of the methods most relevant to the key points we make in this work. For the full mathematical description refer to Kornblith et al. (2019).

**Neuron**   Following related works, we define a neuron as a vector of values it takes over a dataset (Li et al., 2015; Raghu et al., 2017; Morcos et al., 2018; Kornblith et al., 2019). Formally, let $D$ be a dataset consisting of data examples $\vec{d}$:

$$D = \{\vec{d_1}, \cdots \vec{d_m}\}$$

Let $\varphi_i$ be a function that returns a neuron activation value for the training example at the $i$-th unit of the $l$-th layer of the network. The *neuron* $\vec{z}_i$ is the *vector* of network activations recorded by applying $\varphi_i$ over the elements of $D$, i.e.

$$\vec{z^i} = [\varphi_i(\vec{d_1}), \cdots, \varphi_i(\vec{d_m})]$$

---

[3]In the paper, we refer to both SVCCA and PWCCA simply as CCA unless otherwise specified.

In practice, we pass a set of data examples to the network and record activations for each unit at every layer. The vector of these activations is what we consider a representation of a neuron $\vec{z}$.

**Layer** The frequent goal of representational similarity analysis is to compare layers of neural networks. Under our definition, the layer $L$ is the list of vectors (matrix) that consists of the *neurons* at a particular depth, i.e.

$$L = [\vec{z^i}, \cdots, \vec{z^n}]$$

where $n$ is the number of neurons at layer $L$. Alternatively, we can think of layer $L$ as the subspace of $R^m$ spanned by its neurons $(\vec{z^i}, \cdots, \vec{z^n})$, where $m$ is the number of examples in the dataset.

CCA/CKA indexes rely on the idea of subspaces spanned by the neurons, making them powerful when comparing representations across *different networks*. There can be more neurons in the first layer than in the second; the neurons also do not need to be aligned. CCA/CKA uses neurons only to describe the vector subspaces and then compare the subspaces as opposite to the neurons themselves.

That is why methods like CKA and CCA try to find some second-order descriptions of representational spaces (e.g., gram matrices/canonical vectors) and compare these. The decisions on what second-order information to consider and what comparison technique to use define the differences between the indexes.

**Dominant Correlations** The first step for all methods is to center each neuron in the layer representations:

$$X := L_1 - mean(L_1)$$
$$Y := L_2 - mean(L_2)$$

Let X and Y have $p_1$ and $p_2$ neurons (columns). Consider gram matrix $XX^\mathrm{T}$. Because neurons in X are centered, $XX^\mathrm{T}$ is proportional to covariance matrix of $X$. Therefore, the elements in $XX^\mathrm{T}$ correspond to all pairwise covariance similarities data points in X (the same holds for $YY^\mathrm{T}$).

Now consider doing eigendecomposition of $X^\mathrm{T}X$. Eigenvectors $\vec{u}_X^i | i \in \{1, ..., m\}, \vec{u}_X^i \in R^m$ will represent directions of the most dominant correlations of data points in X. Also, we can think about vectors $\vec{u}_X^i$ as of *eigenneurons*, the ones that explain the most variance in the representational space of other neurons. $\lambda_X^i$ is then the $i^{\mathrm{th}}$ eigenvalue of $XX^\mathrm{T}$ (the strengths of the eigenneurons).

**CCA** The directions $\vec{u}_X$ and $\vec{u}_Y$ are orthogonal by the definition of the eigendecomposition. The pair of vectors with the maximum dot product $\langle \vec{u}_X, \vec{u}_Y \rangle$ is called the first pair of canonical directions. The value of their dot product is the first CCA coefficient. Then the second pair produces the second canonical coefficient, and so on.

The formula for the CCA similarity index is then as follows (from Kornblith et al., 2019):

$$CCA(XX^\mathrm{T}, YY^\mathrm{T}) = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \langle \vec{u_X^i}, \vec{u_Y^j} \rangle^2 / p_1.$$

(1)

**CKA** We might also consider weighting the CCA correlations by their eigenvalues. This results in Linear CKA (from Kornblith et al., 2019):

$$\mathrm{CKA}(XX^\mathrm{T}, YY^\mathrm{T}) =$$

$$= \frac{\sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \lambda_X^i \lambda_Y^j \langle \vec{u_X^i}, \vec{u_Y^j} \rangle^2}{\sqrt{\sum_{i=1}^{p_1} (\lambda_X^i)^2} \sqrt{\sum_{j=1}^{p_2} (\lambda_Y^j)^2}}$$

(2)

In this work, we focus on Linear CKA because related works such as Muller et al. (2021) and Conneau et al. (2020) use it.

**SVCCA** If we also decide to apply SVD as the preprocessing step after centering, we get SVCCA. CCA then computes correlation coefficients only for top K components from SVD transformed data (right singular values) and thus can be better averaged (see Equation 1).

**PWCCA** Finally, instead of taking a simple average of CCA coefficients or weighting them by singular values (as in CKA), we might weight them weights (loosely speaking) related to the CCA directions that encapsulate the most data when projected.

In summary, all these methods are related and based on the idea that we can deduce some dominant correlation directions in $X$ and $Y$ and then compare these. Another way to look at it is that if CCA/CKA can represent neurons in $Y$ as linear combinations of neurons in $X$, these correlation methods will generally respond with high scores.

The differences between methods make them invariant to the data scaling, centering, and orthogonal transformations. At the same time, CCA and SVCCA will not change their scores under any invertible linear transformations of either $X$ or $Y$ (see Kornblith et al., 2019 for more details).

# 4 Problems With CKA/CCA

By performing an illustrative experiment, let us introduce problems with CKA and CCA indexes.

Specifically, we want to check if different normalization choices of the Transformer (Vaswani et al., 2017) layers impact the zero-shot cross-lingual transfer capabilities of the model and the similarity of cross-lingual representations it learns.

This section presents a two-fold case against CKA/CCA for cross-lingual similarity analysis:

- empirical: CKA fails to uncover relationships between similarity after the architectural change that does not hurt the performance of the model;

- conceptual: lack of interpretability and unsatisfying underlying assumptions in CCA/CKA.

## 4.1 Experiments Setup

**Models** We train the following three XLM-Roberta (Conneau and Lample, 2019) language models (`base` size versions) from scratch (each with a different normalization schema):

- Post-LN (`scale_post`): normalization block is placed *after* the residual connections in the transformer block (part of the original Transfomer);

- Pre-LN (`scale_pre`): normalization block is placed *before* the residuals (this was shown to improve training by Xiong et al., 2020);

- Normformer (`scale_normformer`): normalization block is placed *before* the residuals *and* FeedForward, Residual, and Self-Attention layers are also normalized (Shleifer et al., 2021).

**Pre-Training** We pre-train a model based on XLM-R Base using 50M sentences uniformly sampled from four languages: English, French, Estonian, and Bulgarian. We chose the languages to be reasonably diverse: French is the most similar to English in both grammar and alphabet, Bulgarian is from a different language group (Slavic), and Estonian is from a completely different language family (Finno-Ugric). We train the model for 1M batches of 512 sentences from the *CC100* dataset using two Nvidia A100 GPUs. The only architectural difference from the original XLM-Roberta is that we change normalization types to Pre-LN and Normformer; other setup details are painstakingly identical.

**Experiment 1: XNLI Fine-Tuning** After having three models pretrained, we fine-tune each of them on XNLI sentence classification task (Conneau et al., 2018). We use only English data for training but evaluate on English and other language evaluation sets (we only skip Estonian since it is not a part of XNLI). This setup, where we tune on one language but use another at test time, is called *zero-shot cross-lingual transfer*.

**Experiment 2: CKA Similarirty** After having the XNLI zero-shot cross-lingual transfer scores, we extract sentence representations from all layers of each model and compare layers using the CKA similarity index.

The parallel corpus is composed of Singh et al. (2019b)'s extension of the XNLI dataset (10k examples for each pair)[4].

We embed the source and target sentences with the models and perform mean-pooling over tokens at each layer for each language pair (as suggested by Del and Fishel, 2021). Next, we compare two parallel sets of sentence representations using the CKA similarity index to get a similarity score for each layer.

**Experiment 3: Per-Layer Matching Accuracy** Lastly, to get insight into some cross-lingual behavioral capabilities of representations at each layer, we analyze them with a sentence-matching probing task.

We use the same data and pooling strategy as in the CKA analysis. For each English sentence, we find the closest target sentence in the opposite language (out of all 10k targets) by cosine similarity. If this sentence is the actual parallel counterpart (translation) of the English sentence, we say the model got this English example correct. Then we compute the accuracy of this sentence matching as the ratio between correctly labeled English examples and the total number (10k) of English examples.

Throughout this work, we conduct experiments across languages sampled from the four language families: Germanic, Romance, Slavic, Baltic, and Finno-Ugric. While the results hold across the complete set of languages from our work, we showcase different subsets of languages from language families in different experiments to introduce more diversity while keeping the plots concise.

---

[4] Using XNLI for both fine-tuning and CKA analysis allows us to avoid domain mismatch scenarios entirely

## 4.2 Experiments Results

**Experiment 1: XNLI Fine-Tuning** See Table 1 for our models' zero-shot cross-lingual transfer performance on the XNLI validation set.

| Normalization | en | fr | bg |
|---|---|---|---|
| scale_post | 0.79 | 0.72 | 0.70 |
| scale_pre | 0.81 | 0.72 | 0.72 |
| scale_normformer | 0.79 | 0.72 | 0.71 |

Table 1: Accuracy of XLM-Roberta Base Transformers pre-trained with different normalization schemes and fine-tuned on the English portion of the XNLI sentence classification task. The models show similar zero-shot cross-lingual transfer performance.

The Table shows that all three models achieve solid zero-shot transfer performance with a cross-lingual transfer gap of 7-9%. We see no significant gains from the scale_pre or scale_normformer, but crucially we see no significant losses either.

**Experiment 2: CKA Similarirty** We present per-layer CKA similarity results for the pre-trained (untuned) models in Figure 1.

Figure 1 reveals that while for scale_post and scale_pre CKA show fairly high cross-lingual performance at all layers, the Normformer results are drastically different. While the similarity for the first half of the layers increases (layers 0-5), the CKA score drops dramatically at the middle layer of the network and continues to hang around zero for all remaining layers (layers 6-12).

This result is especially surprising because CKA confidently gives similarity scores that are almost zero, while Table 1 shows no substantial difference in the zero-shot cross-lingual transfer results between English and other languages. For tuned models the CKA also fails to reveal similarity for layers 6-11 (Figure 8 in Appendix A).

In this example, CKA is not capturing the notion of similarity that would coincide with zero-shot cross-lingual transfer performance for XLM-Normformer. Zero-shot transfer (say) from English requires language representations that *converge* to English values so the other languages can re-use the linear prediction head (calibrated for English).

To double-check the result we also retrain the scale_normformer the second time with a different random restart and get the same CKA results (see Figure 7 in Appendix A).
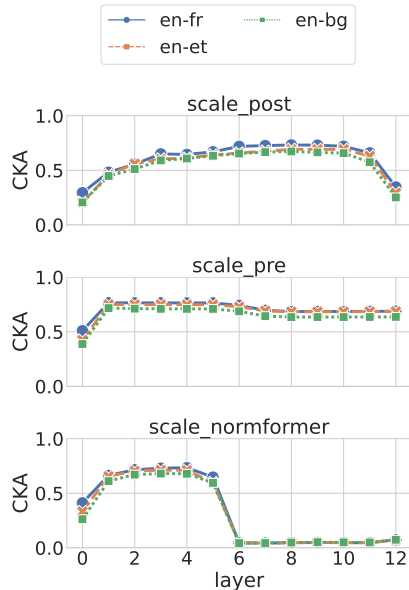


Figure 1: Motivating example 1: counter-intuitive CKA (dis)similarity of XLM-Normformer layers. CKA index shows drastic dissimilarity for layers 6-12 despite remarkable zero-shot cross-lingual transfer performance of the model.
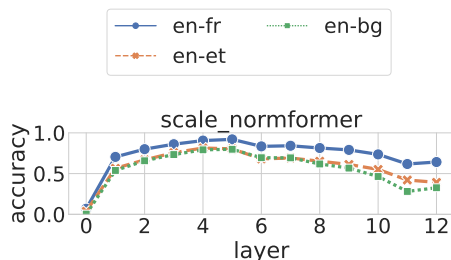


Figure 2: Per-layer sentence matching accuracy for the XLM-Normformer. The result again shows relatively high matching scores for the deeper layers in contrast to the CKA result from Figure 1. There is some decline, but nothing like zero similarity of CKA.

**Experiment 3: Per-Layer Matching Accuracy** However, let us also see the results of our sentence matching task to verify whether these deep representations in Normformer are useful. Figure 2 shows the resulting per-layer accuracy.

The pattern shows that layers 6-12 show some significant cross-lingual matching scores (>50% for French) with only a slightly decreasing trend. This experiment confirms that there are aspects of cross-lingual similarity in these multilingual representations that CKA failed to reveal.

### 4.3 Downsides of CCA

This section shows that the family of CCA-like similarity indexes suffers from similar issues as

CKA. The first downside is that CCA is hard to interpret. CCA is a second-order similarity index (similarly to CKA), which makes it hard to trace the reasons for high/low CCA scores to specific neurons or give any other fine-grained explanation. The second downside is that it is also not robust and has led to the misleading conclusion in the related literature (as demonstrated in Del and Fishel 2021). We discuss these downsides in more detail below.

**Interpretability** Another interesting aspect of our Normformer case is that PWCCA and SVCCA similarity indexes show correlations of about 0.5-0.8 for the layers 6-12 (see Figure 9 in Appendix A for verification). It indicates something special about CKA eigenvalue weighting, normalization (the denominator in Equation 2), or both. One possibility is that dominant eigenneurons (the ones that also have high eigenvalues) in *monolingual* representational spaces are unproportionally similar to each other (and this causes a high denominator and thus the low CKA scores).

In any case, even if we recover what eigenvalues/normalization components cause these extremely low values, it would be even harder to track down which individual neurons cause the problem and to what extent (CCA/CKA methods essentially find linear combinations of the neurons and so mix them up). It highlights the interpretability issue with CKA/CCA indexes that arises when these indexes disagree with our sanity check and with others.

**Conflicting Literature** The disagreement between CCA/CKA also caused a problem of conflicting evidence in the literature. Namely, Singh et al. (2019a) used PWCCA to conclude that mBERT representations diverge starting from the early layers. However, this contradicts the evidence from the multiple behavior studies of mBERT that argue that the opposite is true (Wu and Dredze, 2019; Pires et al., 2019; Liu et al., 2020; Libovický et al., 2020; Conneau et al., 2020; Muller et al., 2021). Del and Fishel (2021) find that merely changing the index from PWCCA to SVCCA or CKA in (Singh et al., 2019a) produces results consistent with related works. It highlights the reliability issue with CKA/CCA.

In summary, similarity indexes value different aspects of representations and correspond to different concepts of similarity. It is, therefore, necessary to consult the specific analysis goal to define what

we want the similarity to capture. It brings us to Section 5 where we propose a simple alternative method that aligns well with the goals of cross-lingual similarity analysis.

## 5 Method: Average Neuron-Wise Correlation (ANC)

In Section 4 we demonstrated multiple drawbacks that CCA/CKA similarity indexes have in the cross-lingual context.

### 5.1 Definition

**Assumption** In this section, we propose a straightforward alternative method that builds on the assumption that neurons in representations for different languages are aligned one-to-one a priori. We find this assumption reasonable to make for several reasons.

First, it aligns well with the goal that motivated most cross-lingual similarity analysis works: zero-shot cross-lingual transfer learning. Zero-shot transfer is possible because a linear prediction head fine-tuned (usually) for English can exploit **direct** linear relationships between English and (say) French representations. Indeed, the linear prediction head calibrates each weight to work with the specific English neuron. Having that specific neuron similar to the French neuron allows the linear head to work on French.

Second, it allows us to decompose the similarity index into correlations of individual neurons, thus facilitating interpretability. We can explicitly see which neurons contribute to the similarity the most/the least, and these neurons have an interpretation of being the most language-specific/language-natural.

Third, it captures the most natural objectives that many cross-lingual alignment literature consider (Wu and Dredze, 2020): representations of the same sentences should have the exact representations (in case the network is aligned). Residual connections strengthen this assumption for hidden layers.

**Description** The solution is straightforward: we compute individual correlations between pairs of English and (say French) neurons and calculate an average score. We also take absolute values of the correlations because the network can swap a negative correlation into a positive with a simple negative weight at the next layer.

Thus, we define Average Neuron-Wise Correlation (ANC) as follows.

Let the centered (by neurons) layer representations be

$$X := L_1 - mean(L_1)$$
$$Y := L_2 - mean(L_2)$$

The (Pearson) correlation $corr$ between two neurons $\vec{z_x}$ and $\vec{z_y}$ form $X$ and $Y$ is defined as:

$$corr(\vec{z_x}, \vec{z_y}) = \frac{\langle \vec{z_x}, \vec{z_y} \rangle}{\|\vec{z_x}\|\|\vec{z_y}\|} \qquad (3)$$

We thus define The ANC similarity between two layers $L_1$ and $L_2$ as:

$$ANC(X, Y) = \frac{\sum_{i=1}^{n} abs(corr(\vec{z_x^i}, \vec{z_y^i}))}{n} \qquad (4)$$

It is only possible for us to construct such an index because the neurons come from a single network where we already know what alignment between neurons is (and ought to be). The method will not work if neurons come from layers of two different networks, for example. In these cases, CCA-like indexes are likely the best fit.

## 5.2 Sanity Checks

In this subsection, we verify that our method gives plausible predictions in the cases where we already know what the result should be.

**Based on the Insight From the Literature** We based this sanity check on the known insight from the literature. The multilingual BERT model (`bert-base-multilingual-cased`) is widely studied in the literature (Wu and Dredze, 2019; Pires et al., 2019; Liu et al., 2020; Conneau et al., 2020). Muller et al. (2021) provided direct behavioral evidence that representations in mBERT (`bert-base-multilingual-cased`) should follow the "first align, then predict" pattern: they first converge towards each other and diverge slightly only at deep layers.

Libovický et al. (2020) and Del and Fishel (2021) demonstrated that the said pattern generalizes to the XLM-Roberta (`xlm-roberta-base`) model (Conneau and Lample, 2019), which is similar in size and training objective to mBERT with the main differences being the removal of the next sentence prediction loss and training on the segments of texts (irrespectively to sentence boundaries)
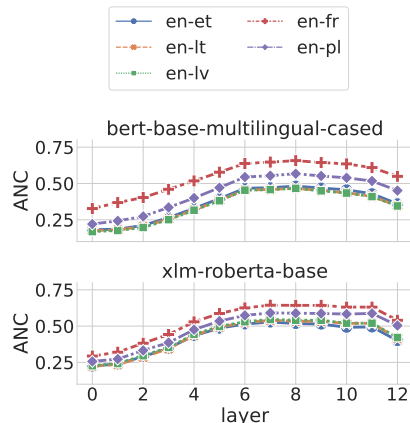


Figure 3: ANC result for the mBERT and XLM-R models. Our method captures the "first align, then translate" pattern presented in Muller et al. (2021) and Del and Fishel (2021).

So our method should reveal the "first align, then predict" pattern in these two cases. Otherwise, we conclude that it fails to capture the relevant properties of similarity we desire.

Figure 3 shows the resulting ANC scores for mBERT and XLM-R `base` models.

The result demonstrates that our method passes the proposed sanity check by being able to reveal the "first align, then predict" pattern. Also, the correlation at the most language natural layers is about 0.7, which indicates that the ANC's *strong assumption* of one-to-one aligned neurons is informative. Lastly, we can see that the ANC distance between English and other languages is more considerable for mBERT than for XLM-R, which corresponds to how these models perform in a cross-lingual transfer (Conneau and Lample, 2019).

**Based on the Experiment in Section 4** We base this sanity check on the same XLM-Roberta Normformer experiment that we used to present the CKA failure case in Section 4. Our method should be able to reveal that representations at deeper layers in `scale_normformer` are somehow cross-lingually similar. Moreover, it should also keep the results for the analogous `scale_post` and `scale_pre` models models in agreement.

We present ANC results for the Section 4 experiment in Figure 4.

The figure shows that unlike CKA (Figure 1), the ANC is able to reveal the "first align, then predict" pattern for the `scale_normformer` and better explains the evidence we provided in Table 1 and Figure 2.
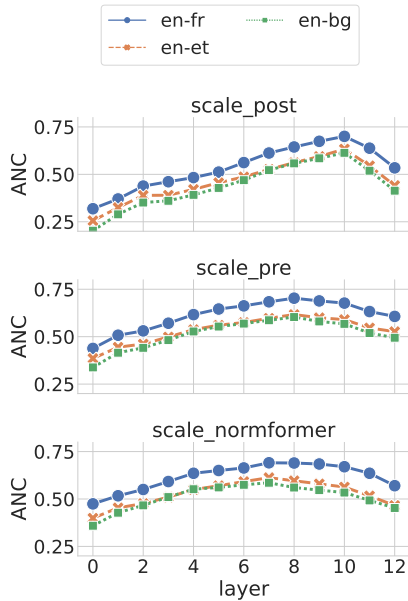
191

Figure 4: ANC result for the three models we presented in Section 4. Our method, unlike CKA (Figure 1), does capture the cross-lingual similarity existing in the deeper layers of XLM-Roberta Normformer (*scale_normformer*).

In summary, this section demonstrated that our method passes the sanity checks of both related literature and the Section 4 experiment (that made CKA fail). In addition, considering how simple it is to interpret ANC scores (the score is a simple average of neuron-wise correlations), the method is a beneficial tool for comparing representation between languages in a single multilingual model.

## 6 Scaling Laws of Cross-lingual Representational Similarity in Multilingual Models

In previous sections, we justified our claim that ANC is better suited for cross-lingual analysis than CCA/CKA methods. In this section, we present an application of ANC to the analysis of representational similarity scaling in cross-lingual language models.

Most related works that analyzed representational patterns in multilingual language models focused on a single model, such as `base` version of mBERT or XLM-R. In Section 5.2 we covered these models showing that ANC accompanies our representational similarity index demands from these models. However, as the model scaling brings significant improvements in downstream tasks performance, we must focus our analysis efforts on the large models and scaling laws (Bowman, 2022).

| Name | type | #params | l | n | #lgs |
|---|---|---|---|---|---|
| xlm-roberta-base | MLM | 270M | 12 | 758 | 100 |
| xlm-roberta-large | MLM | 550M | 24 | 1024 | 100 |
| xlm-roberta-xl | MLM | 3.5B | 36 | 2560 | 100 |
| xlm-roberta-xxl | MLM | 10.7B | 48 | 4096 | 100 |
| xglm-564M | CLM | 564M | 24 | 1024 | 30 |
| xglm-1.7B | CLM | 1.7B | 24 | 2048 | 30 |
| xglm-2.9B | CLM | 2.9B | 48 | 2048 | 30 |
| xglm-4.5B | CLM | 4.5B | 48 | 4096 | 134 |
| xglm-7.5B | CLM | 7.5B | 32 | 4096 | 30 |

Table 2: Model details for XLM-R and XGLM models we study. *type*: training objective of the model, *#params*: number of parameters, $l$: number of layers, $n$: number of hidden units (neurons at each layer), *#lgs*: number of languages used in pertaining.

In this section, we use ANC to explore if the "first align, then predict" pattern generalizes to CLMs and if it preserves in the large-scale versions of multilingual MLMs and CLMs.

**Model Details** We describe the models we study in Table 2. The Table shows that there are two groups of models: MLMs (encoder only) and CLMs (decoder only). Models in each group notably vary in a number of parameters and neurons at each layer.

**Results** Figures 5 and 6 reveal that the cross-lingual similarity of multilingual representations in all the networks we study follows the same "first align, then translate" pattern. It happens despite differences in training objectives, number of languages, and sizes. Therefore, this result provides evidence that multilingual models rely on the exact mechanism described in (Muller et al., 2021), independently of the size or the MLM/CLM training objective.
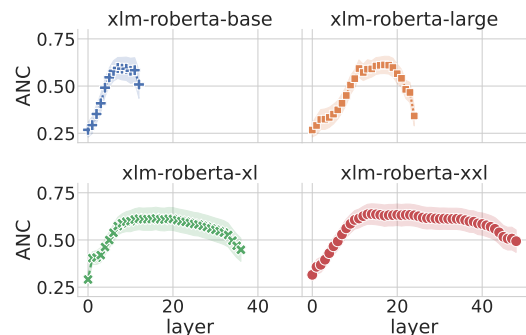


Figure 5: ANC cross-lingual representational similarity for the XLM-R MLM-style models of different sizes. All models follow a similar "first align, then predict" pattern. We aggregate among en-fr, en-de, en-ru, and en-et pairs and show similarity average and spread.
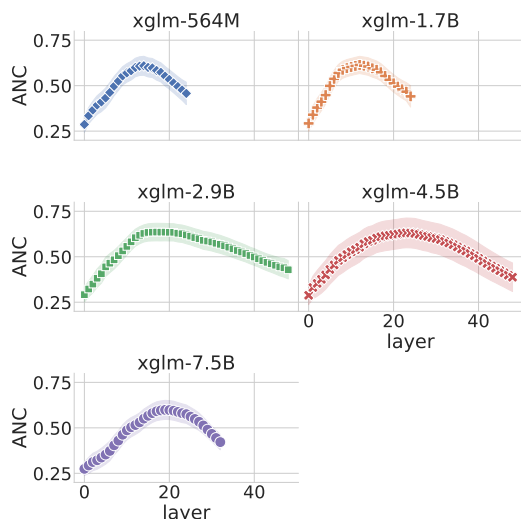
Figure 6: ANC cross-lingual representational similarity for the XGLM CLM-style models of different sizes. All models follow a similar "first align, then predict" pattern. We aggregate among en-fr, en-de, en-ru, and en-et pairs and show similarity average and spread.

## 7 Conclusion

In this study, we introduced an example where *CKA* drastically fails to reveal the cross-lingual similarity between language representations across the deeper layers of the multilingual model. We also highlighted that *CCA* methods suffer from related problems as well (despite passing that concrete sanity check that CKA failed).

Then, we proposed a new approach: Average Neuron-Wise Correlation (ANC), which builds on the assumption of neuron alignment in cross-lingual representations. We verified that our method passes the sanity check at which CKA fails and produces results harmonious with the evidence from related work.

Finally, we used ANC to show that the "first align, then translate" pattern of cross-lingual representations generalizes to CLMs and the larger scales of MLMs and CLMs.

## Acknowledgements

---

[5] https://browser.mt/

## Ethical Considerations

Our work aims to improve the methodology used to perform cross-lingual similarity analysis in multilingual models. We showed that our method outperforms the previous tooling across languages sampled from the four language families: Germanic, Romance, Slavic, Baltic, and Finno-Ugric. To introduce more diversity, we sample different languages from these language families in different experiments. However, we did not experiment with other language families and extremely low-resource languages. These languages might be underrepresented in pretrained LMs and require different analysis tooling. At the time being, we recommend using our method *together* with the previous methods for more reliable results in these cases.

## References

Samuel Bowman. 2022. The dangers of underclaiming: Reasons for caution when reporting how NLP systems fail. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7484–7499, Dublin, Ireland. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexis Conneau and Guillaume Lample. 2019. *Cross-Lingual Language Model Pretraining*, chapter 33. Curran Associates Inc., Red Hook, NY, USA.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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*. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–

6034, Online. Association for Computational Linguistics.

Maksym Del and Mark Fishel. 2021. Similarity of Sentence Representations in Multilingual LMs: Resolving Conflicting Literature and Case Study of Baltic Languages. arXiv.

Harold Hotelling. 1936. Relations Between Two Sets Of Variates*. *Biometrika*, 28(3-4):321–377.

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.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual nmt representations at scale. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John E. Hopcroft. 2015. Convergent learning: Do different neural networks learn the same representations? In *FE@NIPS*.

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.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2021. Few-shot learning with multilingual language models. arXiv.

Chunxi Liu, Qiaochu Zhang, Xiaohui Zhang, Kritika Singh, Yatharth Saraf, and Geoffrey Zweig. 2020. Multilingual graphemic hybrid ASR with massive data augmentation. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 46–52, Marseille, France. European Language Resources association.

Ari Morcos, Maithra Raghu, and Samy Bengio. 2018. Insights on representational similarity in neural networks with canonical correlation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5732–5741. 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.

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.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6076–6085. Curran Associates, Inc.

Sam Shleifer, Jason Weston, and Myle Ott. 2021. Normformer: Improved transformer pretraining with extra normalization. arXiv.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019a. Bert is not an interlingua and the bias of tokenization. In *EMNLP*.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019b. 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.

University of Tartu. 2018. Ut rocket.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.

Shijie Wu and Mark Dredze. 2020. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, Online. Association for Computational Linguistics.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

## A Appendix

This appendix contains supplementary figures that support some auxiliary claims throughout the paper.
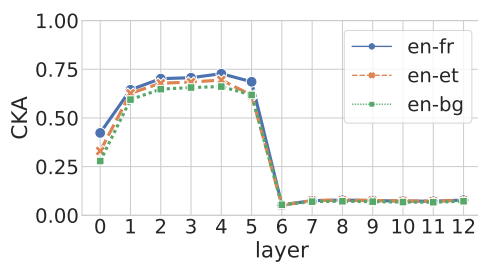


Figure 7: The CKA score for another Normformer (*scale normformer*) model that we pre-trained from the different initialization. The cross-lingual similarity of deeper layers is about zero according to CKA despite evidence of the opposite from Section 4.2
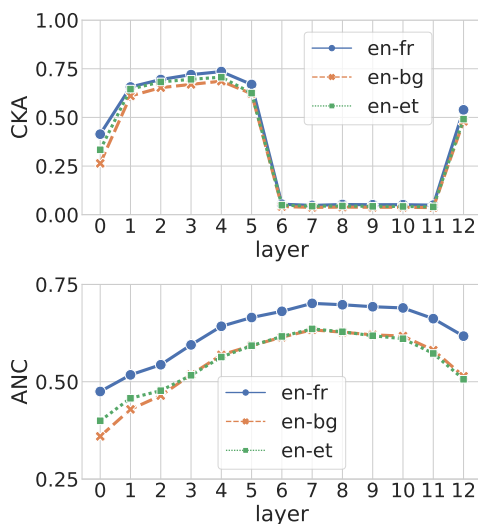


Figure 8: CKA and ANC results for the XLM-Normformer tuned on XNLI. The last layer is a CLS-pooled embedding (the one we tune for XNLI), while others are mean-poolings. CKA captures the similarity between CLS representations at the last layer but fails to capture it at layers 6-11. ANC captures the similarity across all layers.
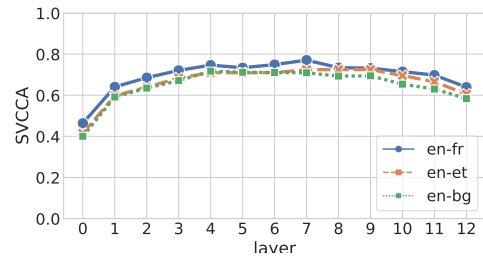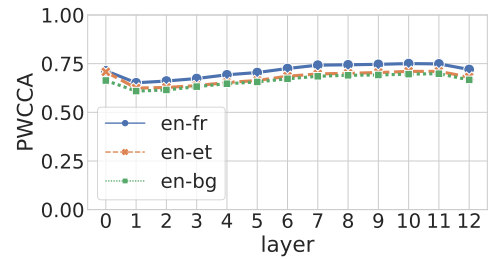


Figure 9: PWCCA and SVCCA results for the XLM-Normformer. These results are more intuitive to our notion of similarity for this particular case but struggle in other scenarios.