Measuring Cross-lingual Transfer in Bytes

Leandro Rodrigues de Souza
FEEC, UNICAMP, Brazil

Thales Sales Almeida
IC, UNICAMP, Brazil
Maritaca AI, Brazil

Roberto Lotufo
FEEC, UNICAMP, Brazil
NeuralMind, Brazil

Rodrigo Nogueira
FEEC, UNICAMP, Brazil
Maritaca AI, Brazil

Abstract

Multilingual pretraining has been a successful solution to the challenges posed by the lack of resources for languages. These models can transfer knowledge to target languages with minimal or no examples. Recent research suggests that monolingual models also have a similar capability, but the mechanisms behind this transfer remain unclear. Some studies have explored factors like language contamination and syntactic similarity. An emerging line of research suggests that the representations learned by language models contain two components: a language-specific and a language-agnostic component. The latter is responsible for transferring a more universal knowledge. However, there is a lack of comprehensive exploration of these properties across diverse target languages. To investigate this hypothesis, we conducted an experiment inspired by the work on the Scaling Laws for Transfer. We measured the amount of data transferred from a source language to a target language and found that models initialized from diverse languages perform similarly to a target language in a cross-lingual setting. This was surprising because the amount of data transferred to 10 diverse target languages, such as Spanish, Korean, and Finnish, was quite similar. We also found evidence that this transfer is not related to language contamination or language proximity, which strengthens the hypothesis that the model also relies on language-agnostic knowledge. Our experiments have opened up new possibilities for measuring how much data represents the language-agnostic representations learned during pretraining.[1]

1 Introduction

The emergence of self-supervised pretraining models such as BERT has revealed a notable phenomenon of cross-lingual transfer even when these models are trained on multilingual corpora devoid of paired translation examples. For example, LLAMA (Touvron et al., 2023), which was trained self-supervisedly on an English-centric corpus, exhibits surprising multilingual capabilities (Yuan et al., 2023; Ye et al., 2023). The underlying mechanisms driving this behavior remain unclear, with hypotheses ranging from the presence of shared "anchor" tokens (Pires et al., 2019) to language contamination (Blevins and Zettlemoyer, 2022), yet no scientific consensus has been reached.

Research in this area often involves the use of pre-existing language models (LMs), which are subsequently finetuned on supervised datasets in different languages (de Souza et al., 2021; Yuan et al., 2023). However, when evaluating multiple languages, conventional methodologies encounter two significant challenges: firstly, the dependence on supervised finetuning datasets, which often vary in size and quality, complicating cross-lingual comparisons; secondly, the use of subword tokenizers, which do not represent all languages equally.

In this work, we avoid these problems by working with a byte-level tokenizer and by using auto-regressive language models trained in self-supervised from scratch in one language and then finetuned on another. To measure the effect of transfer learning, we employ the concept of data transfer (Hernandez et al., 2021), which allows us to quantify how much each different source language contributes to the perplexity of the target language.

Our main contribution is providing a method that measures how much knowledge, in bytes, is transferred from one language to another. By applying it, our findings reveal a surprising trend: even when comparing linguistically distant languages, the data transfer metrics are of a comparable magnitude. This research contributes additional evidence supporting the language-agnostic hypothesis, which suggests that the internal representations developed

[1] The code used in our experiments is publicly available at `https://github.com/lersouza/language-transfer`. We rely on the mC4 dataset from Huggingface, available at `https://huggingface.co/datasets/mc4`

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7526–7537
June 16-21, 2024 ©2024 Association for Computational Linguistics

by a model are not only influenced by the linguistic surface form but also by the cultural and semantic content of the training data.

2 Related Work

Prior work attributed the success of multilingual models in cross-lingual transfer to “anchor” tokens (Pires et al., 2019). However, subsequent research demonstrated that models could perform well even without these tokens (Artetxe et al., 2020), highlighting the significance of shared parameters during training (Conneau et al., 2020). Competitive results were achieved by monolingual models with minimal or no adaptation (Artetxe et al., 2020; de Souza et al., 2021).

Investigations by Blevins and Zettlemoyer (2022) linked these findings to language contamination, where pretraining datasets contained target language data. Additional factors contributing to cross-lingual transfer success include dataset statistics, language attributes (Lin et al., 2019), language structure (Lin et al., 2019; Papadimitriou and Jurafsky, 2020; Chiang and yi Lee, 2020; Ri and Tsuruoka, 2022), and token overlap between training and target languages (Beukman and Fokam, 2023). The role of language script (Fujinuma et al., 2022) and model tokenizer (Rust et al., 2021) was also noted, prompting the use of a byte tokenizer to address these issues (Xue et al., 2022; Abonizio et al., 2022).

Recent research proposed a two-component model representation hypothesis—language agnostic and language specific (de Souza et al., 2021; Zeng et al., 2023; Wu et al., 2022). While promising, no study has measured how much of the language-agnostic component is used in settings with multiple source and target languages. Additionally, existing research still applies the source language vocabulary to the target language, potentially compromising input representations and affecting results.

To address these gaps, we draw on Hernandez et al. (2021) and employ a byte vocabulary in our experiments to overcome current literature limitations.

3 Methodology

Inspired by Hernandez et al. (2021), our methodology focuses on quantifying the transferability of pretraining data across distributions, particularly between different languages. We select a **target** language and finetune models initialized from various **source** languages onto it. Subsequently, we evaluate each model on a **target** language test set and compare their performance. We introduce the Data Transfer (D_T) metric to estimate knowledge transfer, explained in more depth in 3.1. This process is repeated across different **target** languages to observe effects across a broad linguistic spectrum.

These experiments aim to quantify and compare cross-lingual knowledge transfer from different source languages. This analysis seeks to uncover the extent to which transferability depends on specific languages and the importance of language-agnostic components in learned representations. The following sections provide further details.

3.1 Data Transfer Estimation

To quantify the knowledge transfer from pretrained models to a given target language, we utilize the Data Transfer (D_T) metric. This metric assesses the effectiveness of pretraining data by measuring the additional tokens required in the target language for a model initialized from scratch to match the performance of a model pretrained on a source language and finetuned on the target. Figure 1 illustrates this concept.

In our experiments, we utilize a set of M different dataset sizes, denoted as $\mathbf{D_F} = \{s_0, s_1, ..., s_m\}$, in the target language. Initially, we train a random-initialized model on these datasets, resulting in a set of perplexities $\mathbf{P_R} = \{p_{r,0}, p_{r,1}, ..., p_{r,m}\}$. Subsequently, we train from scratch another model on a fixed amount of tokens (e.g., 6B) in a source language ℓ, and then finetune it on the target language using the same dataset sizes $\mathbf{D_F}$, generating another set of perplexities $\mathbf{P}_\ell = \{p_{\ell,0}, p_{\ell,1}, ..., p_{\ell,m}\}$.

To estimate the Data Effective metric (D_E), which represents the amount of data needed to achieve a certain performance, we utilize a linear interpolation function $\gamma(y', \mathbf{X}, \mathbf{Y})$. This function interpolates between discrete data points $(x_j, y_j) \in X \times Y$, evaluated at y'. In our case, the calculation is expressed as:

$$D_{E,i} = \gamma(p_{\ell,i}, \mathbf{D_F}, \mathbf{P_R}) \tag{1}$$

Here, $D_{E,i}$ signifies the Data Effective metric for the i-th perplexity value in $\mathbf{P}_\ell$. $\mathbf{P_R}$ denotes the set of perplexity values derived from the random-initialized model, while $\mathbf{D_F}$ represents the dataset sizes employed during finetuning.

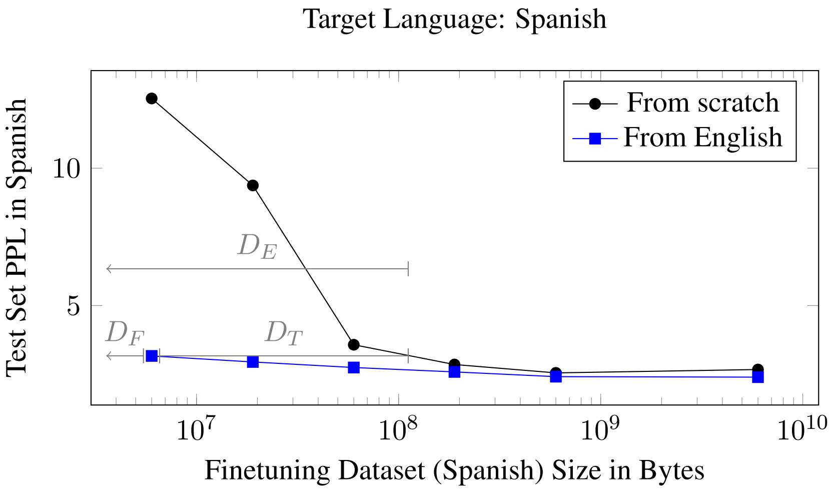

Figure 1: Example illustrating how the coeficients D_T, D_F and D_E are calculated. Each series represents a different initialization. D_T is the number of additional tokens in the target language that a from-scratch model would have needed to achieve the same perplexity of a model finetuned from English. D_F is the size of the dataset used for finetuning and D_E accounts for all data, both D_F and D_T.

We utilize the linear interpolation function provided by the NumPy library to approximate $D_{E,i}$. Further details can be found in the NumPy documentation.[2]

Finally, the Data Transfer metric is computed by subtracting the i-th dataset size s_i from $D_{E,i}$:

$$D_{T,i} = D_{E,i} - s_i \tag{2}$$

By subtracting the dataset size from $D_{E,i}$, we account only for the data coming from pretraining. Since a byte vocabulary is utilized, the amount of data transferred is measured in **bytes**.

3.2 Task and Evaluation Metric

We adopt Language Modeling as our main task with perplexity as the performance metric for all experiments. Perplexity, derived from the model's loss (e^{loss}), facilitates future predictions of model behavior in transfer learning scenarios, following the approach in Hernandez et al. (2021). This choice also allows extensive experiments across multiple languages, leveraging datasets like mC4 (Xue et al., 2021) to overcome size limitations inherent in supervised datasets.

3.3 Tokenization Impact

In cross-lingual setups, the choice of tokenization method holds considerable significance (Rust et al., 2021). While subword tokenizers are commonly employed in cross-lingual experiments, using a tokenizer trained in a source language on a distant target language may result in an increased number of tokens. This can lead to the utilization of undertrained embeddings in some instances, introducing challenges for effective sentence representation. Furthermore, dealing with different scripts introduces the issue of numerous "unknown" tokens, exacerbating the difficulty of obtaining suitable input representations for the model.

To address these challenges, we opt for a byte vocabulary based on the approach proposed by Xue et al. (2022), which allows us to standardize representations across all languages, ensuring that each model encounters the same quantity of UTF-8 bytes. By doing so, we mitigate the use of unknown tokens and undertrained embeddings, thereby minimizing the impact of tokenization issues on the performance of our experiments.

3.4 Language Contamination

A potential reason for a pretrained model's superior performance in cross-lingual tasks is the presence of a substantial amount of data in the target language in its pretraining dataset, a phenomenon referred to as language contamination. To quantify this impact, following the approach outlined by Blevins and Zettlemoyer (2022), we examine the rates of target language fragments in the source language dataset and vice versa.

Specifically, for a given source language ℓ and target language t, we calculate the ratio of all lines

[2] See https://numpy.org/doc/stable/reference/generated/numpy.interp.html

classified as ℓ in the target dataset (known as contamination on target) and as t in the pretraining dataset (known as contamination on source). We perform language detection using the *fasttext* tool (Bojanowski et al., 2017), employing a threshold of 0.6 for classification.

Next, we compute the Spearman correlation between the set of Data Transfer metrics $\mathbf{D_T}$ and those ratios obtained from the outcomes of our experiments.

Correlating these rates with the model's data transfer indicator allows us to evaluate the impact of language contamination on model performance.

3.5 Language Similarity

Language similarity is often cited as a crucial factor influencing cross-lingual transfer performance in natural language processing tasks. In this work, we aim to investigate the relationship between language similarity and cross-lingual transfer effectiveness based on the outcomes of our experiments.

To explore this relationship, we measure various distances between languages, including syntactic, geographic, and phonological distances. These distances are calculated based on the methodology proposed by Littell et al. (2017).

We aim to correlate these language distances with the data transfer metric (D_T). By employing Spearman correlation analysis, we seek to discern whether there exists a significant correlation between the observed D_T values and the measured language distances.

This analysis elucidates whether our experimental results can be attributed to the similarities between the languages involved in our cross-lingual experiments.

4 Experiments

This section presents the languages, datasets, model architecture, and training details for our experiments.

4.1 Languages

Source Languages Selection. We chose three diverse languages—English, Russian, and Chinese—for the source language during the pretraining phase. This selection ensures a broad linguistic spectrum while adhering to pretraining budget constraints.

Target Languages Selection. Ten target languages, spanning various language families and different scripts, were chosen to establish a diverse cross-lingual setting. Details, including language codes, are provided in Table 1.

Code	Language	Family	Script
ar	Arabic	Afro-Asiatic	Arabic
en	English	Indo-European	Latin
es	Spanish	Indo-European	Latin
zh	Chinese	Sino-Tibetan	Hanzi
fi	Finnish	Uralic	Latin
de	German	Indo-European	Latin
ko	Korean	Koreanic	Hangul
id	Indonesian	Austronesian	Latin
ja	Japanese	Japonic	Kanji, Hiragana, Katakana
ru	Russian	Indo-European	Cyrillic

Table 1: Characteristics of selected target languages.

4.2 Datasets

For training and finetuning, language subsets from the mC4 dataset (Xue et al., 2021) for the selected languages were utilized.[3] Pretraining datasets comprised approximately 6 billion tokens, while finetuning datasets ranged from 6 million to 6 billion tokens. Documents were sampled at random without replacement until the desired amount of tokens was reached.

4.3 Model Architecture

Our model is a decoder-only Transformer (Vaswani et al., 2017) that uses a byte vocabulary with 256 embeddings with a dimension of 640. The model follows closely the implementation provided in the T5x library. It consists of 10 layers, each having 10 attention heads with dimensions of 64. The intermediate dimension of Multi-Layer Perceptron (MLP) has a dimension of 2560 and GELU (Hendrycks and Gimpel, 2023) activations. The parameters of the embeddings matrix and the final dense layer are shared. We use relative positional embeddings (Shaw et al., 2018). The resulting model has approximately 65 million parameters.

4.4 Training details

Models were trained using a causal language modeling objective. Each batch has 512 sequences of 1024 tokens. We use the AdamW optimizer with an initial learning rate of 2e-4, which decayed to 2e-5 through cosine decay following Hoffmann et al. (2022). Finetuning employed a constant learning rate of 2e-5 over 10 epochs, except for the 6 billion dataset size where we limited it to 3 epochs.

[3] See `https://huggingface.co/datasets/mc4` for more details.

This adjustment was based on preliminary experiments indicating that the model tends to overfit beyond this epoch count in larger datasets. The best model was selected based on the lowest perplexity achieved on the development set. Warmup steps varied with finetuning dataset sizes (ranging from 0 for smaller datasets to 3000 for larger ones), aligning with findings that smaller datasets completed finetuning before warmup completion (Hernandez et al., 2021). We utilized the T5X framework (Roberts et al., 2022) for our experiments. We used a total of 600 hours of a TPU v2-8 (seven hours of pretraining per model, and fifteen hours for the largest finetuning).

5 Results

Results are compiled in Table 2, where we exclusively report instances involving different source and target languages.

Given our methodology, where we vary the source language while keeping the target language constant to assess the impact of pretraining language in cross-lingual scenarios, it is essential to read the table vertically unless stated otherwise. Each row represents the results obtained by finetuning the model from a specific source language to a given target language (indicated in the column). This vertical arrangement facilitates the comparison of model performance across different source languages for the same target language. Furthermore, test sets vary significantly for each language due to the nature of the mC4 dataset. Consequently, results across target languages are not comparable.

Throughout this section, we highlight findings from models finetuned on 6 million tokens (i.e., $\mathbf{D_F} = \{6\text{MB}\}$) unless otherwise specified.[4] This extreme scenario tests models with minimal target language resources.

5.1 Performance with different initializations

We delve into the results of three target languages: Spanish, Arabic, and Japanese. The perplexity scores across all dataset sizes for these languages are highlighted in Figure 2, offering a chance for in-depth analysis despite the constraints of space.

A key observation is the consistent proximity of perplexity values for all three source languages in every target language. For instance, while one might expect a significant advantage for Chinese as a source language when finetuned in Japanese, or for English when paired with Spanish, this is not the case. This suggests the model leverages source language representations even when it lacks significant similarity with the target language.

Taken together, our results indicate that the model can rely on representations beyond those capturing language structure. This observation supports the recent hypothesis that these representations encompass both language-specific and language-neutral components, strengthening the latter as an important aspect.

5.2 Data Transfer estimation for target languages

Analysis of the Data Transfer (D_T) metric in Table 2 reveals that values are consistently close across different source languages for a given target. Notable examples include **Arabic** (en: 101MB, ru: 99MB, zh: 90MB), **Japanese** (en: 47.5MB, ru: 47.8MB, zh: 69.48MB), and **Finnish** (en: 102.62MB, ru: 51.32MB, zh: 76.57MB).

A clear pattern emerges: one initialization stands out for most target languages, while the other two show similar D_T values. This pattern is visually represented in Figure 3, particularly evident for Finnish (fi), Indonesian (id), Japanese (ja), Korean (ko), and Chinese (zh). Additionally, closely clustered results are observed for Arabic (ar), German (de), Spanish (es), and Russian (ru).

With three diverse initializations, the fact that two consistently show similar D_T values, even when distant from the target language (e.g., Russian and Chinese for Finnish), suggests the models leverage language-agnostic representations. This observation aligns with the expected behavior, where D_T would be close to zero if language-specific knowledge were dominant.

English (en) demonstrates effective knowledge transfer to most target languages, potentially due to its widespread presence in corpora across languages. This hypothesis is explored further in Section 5.3. Additionally, Chinese (zh) appears to transfer effectively to Japanese (ja) and Korean (ko), likely due to their linguistic proximity.

We observe that Chinese (zh) tends to transfer effectively to Japanese (ja) and Korean (ko), both of which are considered closer languages. This, together with our other observations, indicates that, while not determinant, language-specific component also plays a role in cross-lingual transfer.

Finally, upon examination of Figure 4, we also

[4] This restriction applies only to the finetuned models. For the from-scratch ones, we need their perplexities on multiple target language dataset sizes to estimate D_T.

Source Lang.	Metric	ar	de	en	es	fi	id	ja	ko	ru	zh
Scratch init.	Perplexity	6.44	14.82	16.28	12.54	12.71	12.00	12.47	11.69	6.27	15.34
English	Perplexity	2.82	3.67	-	3.16	3.57	2.61	3.92	3.58	2.44	4.43
	D_T	**101.02**	**95.25**	-	**121.14**	**76.57**	**102.62**	47.50	48.74	**75.64**	**29.21**
Russian	Perplexity	2.83	3.98	3.66	3.47	3.80	2.84	3.89	3.58	-	4.52
	D_T	99.00	47.87	**174.63**	67.88	50.96	51.32	47.81	48.69	-	26.18
Chinese	Perplexity	2.88	4.26	3.89	3.75	3.98	2.98	3.46	3.48	2.72	-
	D_T	90.63	31.76	66.96	50.27	49.65	50.21	**69.48**	**49.88**	48.47	-

Table 2: Results for Perplexity and Data Transfer (in MB) for all target and source languages. All metrics are reported after finetuning the models in 6 million tokens of the target language.

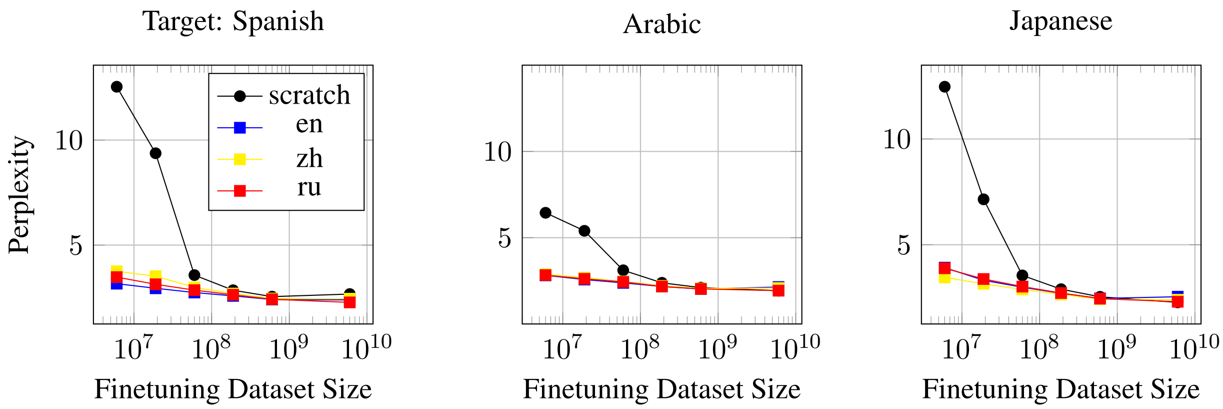

Figure 2: Results measured in Perplexity per token for three target languages. Each series represents a different initialization: train from scratch, finetune from an English, Chinese, or Russian model.

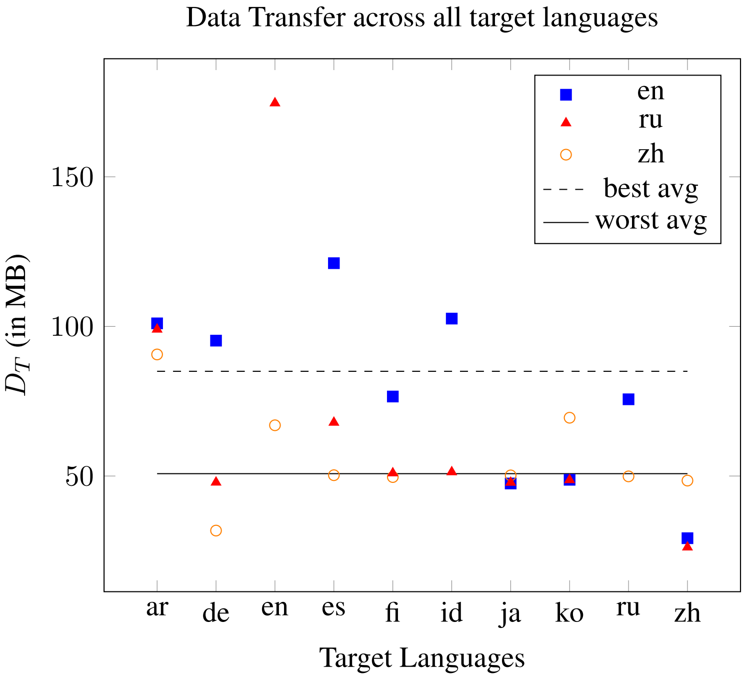

Figure 3: Dispersion chart for Data Transfer (D_T) across target languages. Each series corresponds to a distinct source language. The first dashed line (top-to-bottom) indicates the average of the best results (higher transfer), while the second one represents the average of the worst results (lower transfer).

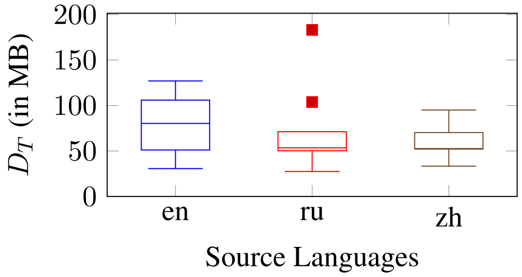

Figure 4: Boxplot with Data Transfer results for the 6 million tokens datasets in all target languages.

observe that most D_T values are clustered between 50MB and 100MB, with low variation within a source language across all target languages. While not ideal, this comparison across target languages indicates that the transfer from our source models is relatively consistent, not spanning more than one order of magnitude.

5.3 Language Contamination Impact

Table 3 summarizes the results of assessing the language contamination effect in our experiments.

Correlation	ρ	*p-value*
$\mathbf{D_T}$ and contamination on source	0.191	0.0157
$\mathbf{D_T}$ and contamination on target	0.265	0.0021

Table 3: Spearman Correlation (ρ) and p-value assessing the correlation of $\mathbf{D_T}$ with both the ratio of a target language in the source dataset (contamination on source) and with source language in the target dataset (contamination on target).

The analysis excludes the 6 billion tokens finetuning dataset size to mitigate the ossification effect, as observed in Hernandez et al. (2021). This effect leads to a performance drop for pretrained models with larger finetuning datasets, worsening perplexity compared to scratch-trained models. Excluding this data point helps avoid introducing noise and adverse effects on coefficient calculation, given its singular occurrence per source-target language pair. Additionally, due to the sample size (< 500 observations), the permutation test is utilized to calculate the *p-value*.

Although a correlation of 0.191 exists between D_T and contamination on the source dataset, this coefficient indicates a weak association, suggesting a minimal impact on cross-lingual performance, contradicting the findings by Blevins and Zettlemoyer (2022).

Exploration of language contamination in target datasets reveals a higher correlation of 0.265, particularly influenced by widespread languages like English. However, this coefficient still signifies a weak association between D_T and target contamination, thus not supporting the language contamination hypothesis.

5.4 Language Similarity and Data Transfer

This subsection outlines our analysis of the correlation between language distances and the data transfer metric (D_T), summarized in Table 4.

Measure to correlate with D_T	ρ	*p-value*
Syntactic distance	-0.147	> 0.7
Geographic distance	-0.110	
Phonological distance	-0.117	
Genetic distance	-0.150	
Inventory distance	-0.090	
Featural distance	-0.145	

Table 4: Spearman Correlation (ρ) and p-value assessing the correlation of $\mathbf{D_T}$ with a diverse set of language distance measurements.

We find a weak correlation between source-target language distances and Data Transfer. Since we are considering multiple language characteristics, such as syntax and phonology, the results suggest that language similarity has a minor role in knowledge transfer between distinct languages. Nonetheless, the small number of source languages necessitates a cautious interpretation of these results, especially since all obtained p-values exceed 0.7, indicating limited statistical significance.

Because of that, we conducted a controlled experiment, pretraining a language model in Portuguese and evaluating its performance on the Spanish target language — a language known for its similarity to Portuguese. Results, depicted in Table 5, were compared across various initializations, including more distant languages like Chinese.

	Spanish	
Source Lang.	D_T	Perplexity
Portuguese	**164.47**	**2.91**
English	121.14	3.16
Russian	67.88	3.47
Chinese	50.27	3.75

Table 5: Results for Data Transfer (in MB) and Perplexity in Spanish, highlighting Portuguese as a source language. All metrics are reported after finetuning the models in 6 million tokens of the target language. English, Russian, and Chinese results are the same as Table 2, added to facilitate comparison.

When initialized with Portuguese, the model achieves a lower Perplexity in Spanish compared to when initialized with other languages. Additionally, D_T peaks among all initializations, suggesting the influence of language proximity between Portuguese and Spanish. English initialization also yields comparable results, with a D_T difference of around 40 MB and a perplexity variation of 0.25. Chinese and Russian show the lowest, yet similar, scores.

One possible interpretation is that the language-agnostic component accounts for 50% of the transfer, with Russian and Chinese being more distant from Spanish, while the language-specific component contributes the remaining 50%, considering closer linguistic and script systems. However, further investigation with more language pairs is necessary to determine the actual factors influencing transfer performance, including linguistic structure, script, or shared cultural knowledge in pretraining datasets.

5.5 Commutative property exploration

Pair (L_1, L_2)	$L_1 \rightarrow L_2$	$L_2 \rightarrow L_1$	Δ
en, ru	75.64	174.63	98.99
en, zh	29.21	66.96	37.75
ru, zh	26.18	48.47	22.29

Table 6: Analysis of the Commutative Property in terms of Data Transfer D_T. We analyze pairs of languages (L_1, L_2), reporting the observed D_T from L_1 to L_2 and vice-versa. Values are reported in megabytes.

We examine the commutative property of data transfer between English (en), Russian (ru), and Chinese (zh) in our cross-lingual experiments (Table 6). Notably, the data transfer amounts exhibit non-commutative behavior, revealing variations in knowledge transfer efficiency across bidirectional language pairs.

In the English-to-Russian transfer (en, ru), data transfer is more efficient when directed from Russian to English (174.63) compared to the reverse direction (75.64), indicating an asymmetry in knowledge transfer. Similarly, in the English-to-Chinese transfer (en, zh), data transfer is more substantial from English to Chinese (66.96) than in the reverse direction (29.21).

The Russian to Chinese transfer (ru, zh) also demonstrates a non-commutative pattern, with higher data transfer from Russian to Chinese (48.47) than in the reverse direction (26.18).

The variance in mC4 subsets for each language introduces significant differences in both pretraining and evaluation datasets, potentially contributing to the absence of a commutative behavior. A more in-depth analysis would necessitate repeating experiments with equivalent datasets.

6 Discussion

Our study aims to measure how much knowledge, in bytes, is transferred from one language to another, enabling the investigation of the effectiveness of language-agnostic representations acquired during pretraining in cross-lingual scenarios. We hypothesize that these representations enable models to perform well on downstream tasks across diverse languages, which is observed in state-of-the-art multilingual models.

In our results, we consistently find that at least two source languages demonstrate very close D_T values when evaluated against a target language, despite the diverse set of script systems and linguistic characteristics involved. This observation suggests that the data transferred from these languages to the target language is not primarily related to language-specific components but rather to language-agnostic ones. For instance, as illustrated in Figure 3, both English and Russian, despite being known as distant languages, achieve nearly identical D_T values when evaluated on Korean, a language distinct from either. Moreover, all three source languages are remarkably close when evaluated on the Japanese test set.

Despite exposure to only a few tokens in the target language, our models demonstrate similar perplexity performance, indicating high adaptability and generalization across a broad range of languages. This reinforces the notion that the language-agnostic component plays a crucial, uniform role across source languages.

Notably, our results are not attributed to pretraining exposure to target languages, since there is a weak correlation of language contamination with the data transfer coefficient. Additionally, the observed performance is not solely dependent on language proximity, as suggested in other works.

While perplexity offers valuable insights, generalizable conclusions require evaluation in downstream tasks, especially in under-resourced languages. To explore this further, we conducted a small-scale experiment, detailed in Appendix A, finetuning our pretrained models for a low-resource language inference task. Surprisingly, we observe comparable accuracy scores between Portuguese and Russian, with good results also for English and Chinese, suggesting that our initial findings with perplexity may extend broadly. However, a more detailed investigation is required to understand cross-lingual transfer mechanisms fully.

The novelty of our approach is employing a byte-level tokenizer and adapting Hernandez et al. (2021) for a cross-lingual scenario. The byte-level approach facilitates consistent model embeddings across diverse scripts, enabling effective cross-lingual knowledge transfer without language-specific tokenization or preprocessing. This is supported by the strong performance of ByT5 compared to mT5 in Xue et al. (2022).

In conclusion, our study suggests the presence of language-agnostic representations contributing to cross-lingual transferability, while also laying the foundation to measure it through the Data Transfer metric. The observed consistency in model performance across diverse languages, facilitated by the byte-level tokenizer, indicates the potential for more efficient and generalizable natural language understanding across linguistic boundaries in computational linguistics and NLP.

7 Limitations

Our study has certain limitations that merit consideration. Firstly, our choice of initializing models with only three languages, while diverse, leaves room for improvement. Including additional languages in the pretraining phase would enhance the robustness of our analysis by minimizing possible bias towards the selected languages while providing more samples for our correlation analysis. However, this expansion would necessitate a more substantial computational budget.

Secondly, our reliance on small models, specifically a 65 million parameter model, limits the scope of our findings as larger models may exhibit different behavior. Additionally, the capacity of very large models for few-shot learning opens avenues for further exploration in the domain of transfer learning.

Lastly, the heterogeneity of the mC4 dataset across languages introduces a potential source of variability in the models' exposure to different knowledge. While the impact of this variation on data transfer remains unclear, conducting experiments with controlled datasets would offer valuable insights. Moreover, employing a more comparable test set could help mitigate statistical variance, particularly in analyses such as the commutative property assessment.

8 Conclusion and Future Work

Our study delves into the transferability of knowledge in cross-lingual scenarios, leveraging a byte-level tokenizer and an adapted methodology inspired by Hernandez et al. (2021). By measuring the models' reliance on pretraining when executing tasks in diverse languages, our approach offers an understanding of the cross-lingual capabilities of language models. The results provide evidence that language-agnostic representations also play an important role in downstream tasks. This not only contributes to the current understanding of cross-lingual transferability but also serves as a catalyst for further exploration into the properties of language-agnostic knowledge transfer. For future research directions, we envision key investigations that can build upon the insights presented in this paper:

1. **Expand Experiment Range:** Use more source languages so we can draw stronger conclusions.

2. **Controlled Datasets Usage:** Employ controlled datasets and comparable test sets to address mC4 dataset heterogeneity, offering clearer insights into varied knowledge exposure impact on cross-lingual transferability and mitigating variance.

3. **Explore Larger Models:** Investigate the use of larger models in few-shot learning downstream tasks as complementary evaluations to perplexity measurements.

4. **Measure D_T from Non-natural languages:** Perform experiments with non-natural language data, such as artificial languages with hierarchical structures. This exploration could shed light on whether knowledge transfer primarily occurs due to the content of pretraining or is largely influenced by the linguistic form.

Acknowledgments

We would like to thank Hugo Queiroz Abonizio, Andrea Roque, and Luciane Kawamata Notario for the useful discussions and insights. We would like to thank Google TPU Research Cloud for the computing. This research was partially funded by grant 2022/01640-2 from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

References

Hugo Abonizio, Leandro Rodrigues de Souza, Roberto Lotufo, and Rodrigo Nogueira. 2022. MonoByte: A pool of monolingual byte-level language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3506–3513, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Michael Beukman and Manuel Fokam. 2023. Analysing cross-lingual transfer in low-resourced african named entity recognition.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Cheng-Han Chiang and Hung yi Lee. 2020. Pre-training a language model without human language.

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.

Leandro Rodrigues de Souza, Rodrigo Nogueira, and Roberto Lotufo. 2021. On the ability of monolingual models to learn language-agnostic representations.

Yoshinari Fujinuma, Jordan Boyd-Graber, and Katharina Kann. 2022. Match the script, adapt if multilingual: Analyzing the effect of multilingual pretraining on cross-lingual transferability. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1500–1512, Dublin, Ireland. Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2023. Gaussian error linear units (gelus).

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. 2021. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. 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.

Isabel Papadimitriou and Dan Jurafsky. 2020. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6829–6839, 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.

Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. 2020. The assin 2 shared task: a quick overview. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 406–412. Springer.

Ryokan Ri and Yoshimasa Tsuruoka. 2022. Pretraining with artificial language: Studying transferable knowledge in language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7302–7315, Dublin, Ireland. Association for Computational Linguistics.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang

Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. 2022. Scaling up models and data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*.

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

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xianze Wu, Zaixiang Zheng, Hao Zhou, and Yong Yu. 2022. LAFT: Cross-lingual transfer for text generation by language-agnostic finetuning. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 260–266, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

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

Jiacheng Ye, Xijia Tao, and Lingpeng Kong. 2023. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability.

Fei Yuan, Shuai Yuan, Zhiyong Wu, and Lei Li. 2023. How multilingual is multilingual llm? *arXiv preprint arXiv:2311.09071*.

Jiali Zeng, Yufan Jiang, Yongjing Yin, Yi Jing, Fandong Meng, Binghuai Lin, Yunbo Cao, and Jie Zhou. 2023. Soft language clustering for multilingual model pretraining. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7021–7035, Toronto, Canada. Association for Computational Linguistics.

A Appendix: Downstream Task Experiment

To further explore the generalizability of our findings beyond casual language modeling with the perplexity metric, we conducted an additional experiment focusing on a distinct downstream task. This experiment involved finetuning our models pretrained on the selected source languages for a specific downstream task, targeting a different language.

A.1 Experiment Details

For this experiment, we selected Portuguese as the target language and the Recognizing Textual Entailment (RTE) task from the ASSIN2 (Real et al., 2020)[5] dataset. The dataset comprised 6,500 training examples and 500 validation instances.

We finetuned the source models for 10 epochs using a constant learning rate of $5e-5$ and a batch size of 128. The objective of the task was to predict whether one sentence entails another, with evaluation based on accuracy measured on the validation set.

A.2 Results

Our results, summarized in Table 7, reveal comparable performance between Portuguese (our baseline) and Russian, despite being considered a distant language. English lags by nearly 6 percentage points compared to our baseline, while Chinese exhibits the poorest performance, trailing our baseline by almost 20 percentage points.

Despite the limited scope of our experiment, focusing on only one target language and task, these findings suggest significant knowledge transfer across languages with varying degrees of similarity. For example, even with the lowest performance observed in Chinese, it still outperforms a

[5] https://sites.google.com/view/assin2/

Source Language	**Accuracy (%)**
Portuguese (baseline)	85.0
Russian	82.0
English	78.6
Chinese	65.6

Table 7: Results from finetuning our source models on the ASSIN2 Recognizing Text Entailment task. We report the accuracy obtained in our validation set.

random classifier by 15 percentage points (ASSIN2 contains two distinct classes).

Based on our findings, cross-lingual knowledge transfer appears to occur even with more distant languages. These results underscore the importance of further exploration in this area, indicating a promising potential for measuring knowledge transfer in cross-lingual scenarios and delineating the contributions of language-agnostic and language-specific components in the models' representations.