Exploring the Transfer Properties of Language Models: A Language-agnostic Hypothesis

Anonymous ACL submission

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 of 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 nor language syntactical proximity, which strengthens our hypothesis that the model 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 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 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 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 by a model are less influenced by the linguistic surface form and more 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 involves both training a model from scratch and finetuning a pretrained model in a **source language**, employing datasets in a **target language** that span different orders of magnitude. To measure the transfer of knowledge from the pretrained model to the downstream task, we calculate the number of additional tokens required for the scratch-trained model to achieve comparable performance. We explain it in more depth in Section 3.3.

These experiments aim to quantify the transferability of pretraining data across distributions, specifically, between different languages. The following subsections highlight specific details of our methodology.

3.1 Evaluation Metric

We have chosen perplexity as our performance metric for all experiments. Since perplexity is based on the model's loss (e^{loss}), this choice facilitates future experiments by allowing the formulation of equations predicting the models' behavior in terms of transfer learning, as in Hernandez et al. (2021).

3.2 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.

3.3 Data Transfer Estimation

Building on Hernandez et al. (2021), we estimate the amount of data transferred from pretraining, measured in bytes. The Data Transfer coefficient (or D_T) is the amount of data in the target language that a model trained from scratch should see to achieve the same performance as a model that was initialized in a source language. The intuition behind this method is that this number of additional tokens a scratch-initialized model needs when compared to a pretrained one indicates how much of the

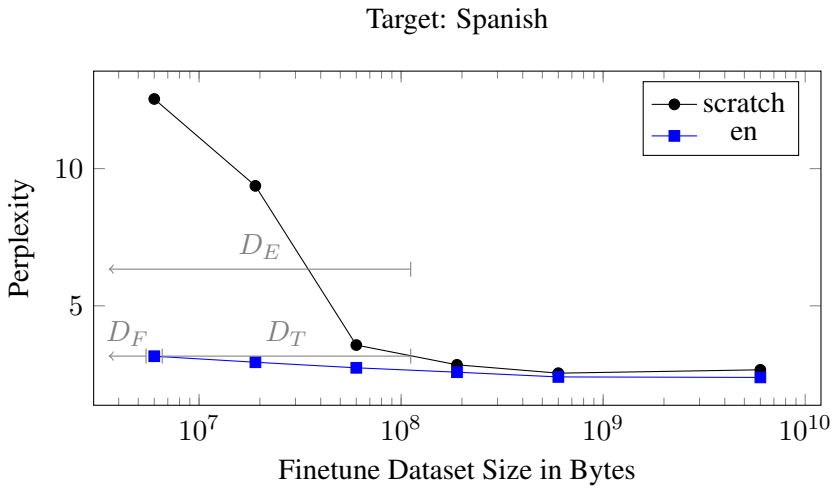

Figure 1: Example illustrating how the coeficients D_T, D_F and D_E are calculated. Each series represents a different initialization. D_T is the number of additional tokens in the target language that a from-scratch model would have needed to achieve the same perplexity of a model finetuned from English.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

language modeling pretraining the latter is using on the downstream task. Figure 1 illustrates this measurement in an example.

In contrast to the original work, we utilize linear interpolation based on experiment data points. We perform experiments with finetune datasets that span four orders of magnitude. These data points, both from scratch and pretrained models, are used to perform linear interpolations to estimate D_E, which is the effective amount of data, accounting for the finetune dataset size (D_F). By subtracting the latter, we get D_T, which accounts only for the usage of the model's pretraining:

$$D_T = D_E - D_F$$

Since we use a byte vocabulary, the amount of data transferred is measured in bytes.

3.4 Language Contamination

A potential reason for a pretrained model excelling in a cross-lingual task is the presence of a substantial amount of data in the target language in its pertaining dataset, referred to as language contamination. To measure this impact, following the exploration by Blevins and Zettlemoyer (2022), we analyze the rates of target language fragments in the source language dataset and vice versa. Correlating these rates with the model's data transfer indicator enables us to assess the contamination's impact.

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 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.

4.2 Datasets

For training and finetuning, language subsets from the mC4 dataset (Xue et al., 2021) for the selected languages were utilized.[1] Datasets were truncated to control token exposure. Pretraining datasets comprised approximately 6 billion tokens, while finetuning datasets ranged from 6 million to 6 billion tokens.

4.3 Model Architecture

Our model follows a decoder-only architecture, employing a byte vocabulary with an embedding dimension of 640. The model consists of 10 layers, each featuring 10 attention heads with dimensions of 64. The intermediate dimension of Multi-Layer Perceptron (MLP) has a dimension of 2560, resulting in a total of approximately 65 million parameters. The non-linearity function used throughout the model is GELU (Hendrycks and Gimpel, 2023).

4.4 Training details

Models were trained using 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. 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 300 hours of a TPU v2-8 (seven hours of pretraining per model, and fifteen hours for the largest finetune).

5 Results

In this section, we present experiment outcomes, concentrating on assessing the model's performance and cross-lingual transfer capabilities. Results are consolidated in Table 2, exclusively reporting instances where source and target languages differ.

Throughout this section, we emphasize findings from models finetuned on the 6 million token dataset unless specified otherwise. This represents an extreme scenario, testing models with minimal target language resources. Analyzing these results is crucial for investigating the hypothesis that models universally leverage knowledge in a language-agnostic manner.

5.1 Performance with different initializations

In this subsection, we highlight the results specific to three target languages—Spanish, Arabic, and Japanese—through the examination of perplexity values, as depicted in Figure 2.

A noteworthy observation is the convergence of results between models initialized from scratch and pretrained models, occurring approximately around 10^9 tokens of the target language. This intersection implies that a model pretrained on a foreign language remains beneficial, especially in scenarios with limited labeled data, a common characteristic of many low-resource languages. This finding underscores the practicality of leveraging pretrained models for effective cross-lingual transfer.

Our results collectively imply that the choice of the source language for pretraining plays a minor role in determining cross-lingual model performance. This phenomenon aligns with our hypothesis that, during pretraining, the models acquire highly generalized representations, facilitating transferability across multiple languages.

5.2 Data Transfer estimation for target languages

In our exploration of data transfer, we adopt a methodology inspired by Hernandez et al. (2021), estimating the data transferred from pretraining using Linear Interpolation[2] with our experiment data points. Notably, we express the data transfer in bytes, aligning with our reliance on a byte-level tokenizer. To visually guide our findings, we present a scatter plot in Figure 3, offering a representation of the data transfer across target languages and source language variations.

One intriguing outcome surfaces: the amount of data transfer remains remarkably consistent across all target languages, spanning at most one order of magnitude. The values are concentrated within the range of 50 to 100 megabytes. Figure 4 illustrates

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

[2] We use the Numpy package for estimating the Data Transfer. For more details, see https://numpy.org/doc/stable/reference/generated/numpy.interp.html

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.

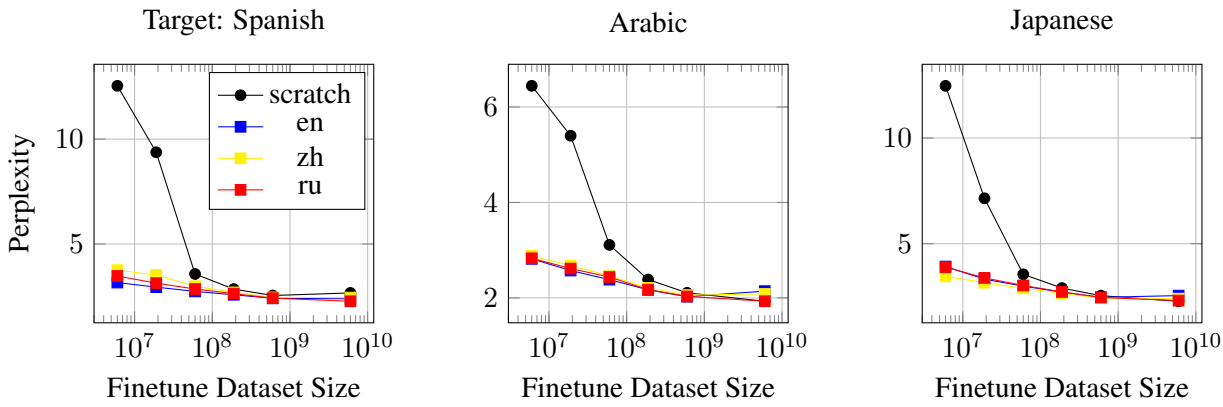

Figure 2: Results measured in Perplexity per token for three target languages. Each series represents a different initialization: train from scratch, finetune from an English, Chinese, or Russian model.

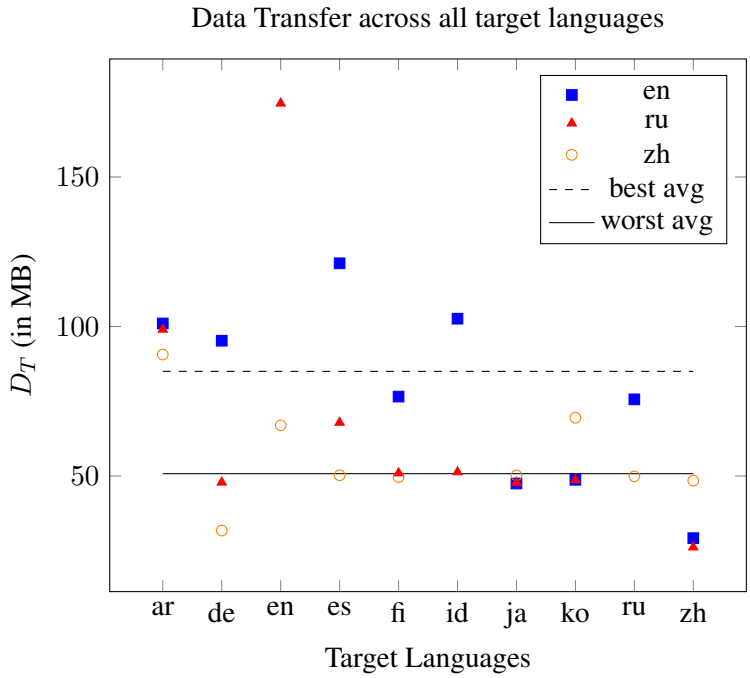

Figure 3: Dispersion chart for Data Transfer (D_T) across target languages. Each series corresponds to a distinct source language. The first dashed line (top-to-bottom) indicates the average of the best results (higher transfer), while the second one represents the average of the worst results (lower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

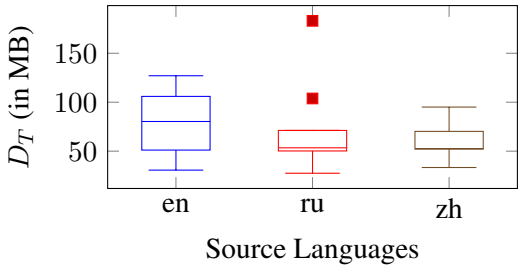

Figure 4: Boxplot with Data Transfer results for the 6 million tokens datasets in all target languages. Each series represents data from a different source language. D_T is expressed in megabytes.

the distribution of data transfer values for source languages, revealing a consistent and low variability pattern. Notably, both Chinese (zh) and Russian (ru) display strikingly similar distributions, emphasizing the uniformity of data transfer characteristics observed across different linguistic contexts, with subtle variations more pronounced in English data. Examination of the first quartile further underscores the remarkable resemblance among all languages.

This notable uniformity supports our language-agnostic hypothesis, implying that the knowledge acquired during pretraining and subsequently transferred to a downstream task exhibits striking similarity across diverse target languages.

An interesting observation emerges from the clustering of four target languages (Finnish, Indonesian, Japanese, and Korean), where two distinct initializations yield nearly identical data transfer amounts. This suggests that the "language-agnostic" component acquired during pretraining is consistent in these cases. A similar trend is observed for German, Spanish, and Russian, though with more noticeable variation between the two data points for each language.

Moreover, specific languages highlight distinct scenarios in data transfer. Arabic (ar) and Chinese (zh) represent edge scenarios, with all source languages performing optimally in Arabic and exhibiting low transfer for Chinese. We attribute the perplexities observed for a model trained from scratch in both languages (with 6 million tokens) to potential variations in the evaluation dataset examples, containing information or concepts that are either easily or challenging to transfer.

English (en) emerges as a standout performer in transferring knowledge to most of our selected target languages. One plausible explanation is the ubiquitous presence of English in corpora from other languages. We explore this hypothesis further in Section 5.3. Another consideration is that the English slice of mC4 used in pretraining may contain a wealth of knowledge transferrable to the evaluation sets of any target language, independent of the language itself.

Additionally, we observe that Chinese (zh) tends to transfer effectively to Japanese (ja) and Korean (ko), both of which are considered closer languages.

5.3 Language Contamination

We evaluate language contamination bidirectionally: measuring target language contamination in the pretraining dataset and source language contamination in the target dataset. Following the approach outlined by Blevins and Zettlemoyer (2022) and employing *fasttext* (Bojanowski et al., 2017) for language classification with a threshold of 0.6, we compute Spearman correlations between language ratios and data transfer coefficients. Table 3 summarizes the results.

Correlation	ρ	*p-value*
D_T and contamination on source	0.191	0.0157
D_T and contamination on target	0.265	0.0021

Table 3: Spearman Correlation (ρ) and p-value assessing the correlation of 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).

In this analysis, we exclude the 6 billion tokens finetune dataset size to mitigate the impact of the ossification effect, as observed in Hernandez et al. (2021). The ossification effect results in a performance drop for the pretrained model with larger finetune datasets, worsening its perplexity compared to a scratch-trained model. Given its potential to introduce noise and adversely affect the coefficient calculation, we exclude this data point, considering it is only one per source-target language pair. We also use the permutation test to calculate the *p-value* because of the size of our sample (< 500 observations).

The observed weak association between selected coefficients and language contamination suggests a negligible impact on cross-lingual performance. This contradicts the findings of Blevins and Zettlemoyer (2022), indicating a minor role for language contamination in the source dataset.

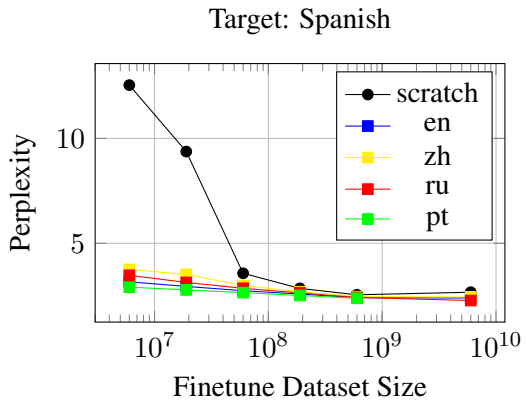

Figure 5: Results for Spanish, measured in Perplexity per token. Each series represents data from a different source language with Spanish as the finetuning (target) language.

Additionally, we explore language contamination in target datasets, positing that the presence of widespread languages, such as English, might influence the model's token predictions by providing contextual clues. The identified weak association does not support the language contamination hypothesis.

5.4 Language Distance and Data Transfer

Expanding our investigation, we explore the potential correlation between source and target language distances and their influence on Data Transfer during pretraining. Following the methodology outlined by Littell et al. (2017), we leverage syntactic language distances computed in advance.

Our findings reveal a very weak correlation (ρ = -0.220) between source-target language distances and Data Transfer, with a p-value exceeding 0.9, suggesting limited statistical significance and caution in drawing conclusions from the dataset.

To deepen our analysis, an additional controlled experiment is conducted by pretraining a language model in Portuguese and evaluating its performance against the Spanish target language. Portuguese is known to be similar to Spanish. Our results, depicted in Figure 5, are compared with various initializations, including more distant languages such as Chinese.

Notably, all initializations exhibit comparable performance, indicating that language distance has a minimal impact on the model's overall effectiveness.

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 4: 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.

5.5 Commutative property exploration

We examine the commutative property of data transfer between English (en), Russian (ru), and Chinese (zh) in our cross-lingual experiments (Table 4). 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 investigates 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.

Surprisingly, our findings indicate that the amount of data transferred across 10 distinct target languages, from a diverse set of script systems and linguistic families, remains consistently close. This supports our hypothesis, suggesting that models rely on a universal form of knowledge. The ability of models to achieve comparable performance, irrespective of linguistic dissimilarity between source

and target languages, is underscored by the uniformity of language-agnostic representations, as depicted in Figures 3 and 4.

Despite exposure to only a few tokens in the target language, our models consistently 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.

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 provides compelling evidence for the efficacy of language-agnostic representations in enabling cross-lingual transferability. The robustness of our models and the role of the byte-level tokenizer offer promising avenues 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. The inclusion of additional languages in the pretraining phase would enhance the robustness of our analysis by minimizing noise. 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 noise, particularly in analyses such as the commutative property assessment.

8 Conclusion and Future Work

Our study delves into the transferability of language-agnostic knowledge in cross-lingual scenarios, leveraging a byte-level tokenizer and an adapted methodology inspired by Hernandez et al. (2021). By measuring and gaining insights into 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 aligns with our hypothesis, emphasizing the significance of language-agnostic representations. 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 noise.

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

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.

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.

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.

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 pre-training. 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.