A Balanced Data Approach for Evaluating Cross-Lingual Transfer: Mapping the Linguistic Blood Bank

Dan Malkin[◇] **Tomasz Limisiewicz**[♠*] **Gabriel Stanovsky**[◇]
[◇] School of Computer Science, The Hebrew University of Jerusalem
[♠] Faculty of Mathematics and Physics, Charles University in Prague
{dan.malkinhueb,gabriel.stanovsky}@mail.huji.ac.il
limisiewicz@ufal.mff.cuni.cz

Abstract

We show that the choice of pretraining languages affects downstream cross-lingual transfer for BERT-based models. We inspect zero-shot performance in balanced data conditions to mitigate data size confounds, classifying pretraining languages that improve downstream performance as *donors*, and languages that are improved in zero-shot performance as *recipients*. We develop a method of quadratic time complexity in the number of languages to estimate these relations, instead of an exponential exhaustive computation of all possible combinations. We find that our method is effective on a diverse set of languages spanning different linguistic features and two downstream tasks. Our findings can inform developers of large-scale multilingual language models in choosing better pretraining configurations.[1]

1 Introduction

Pretrained language models are setting state-of-the-art results by leveraging raw texts during pretraining (PLMs; Peters et al., 2018; Devlin et al., 2019, inter alia). Interestingly, when pretraining on multilingual corpora, PLMs seem to exhibit *zero-shot* cross-lingual abilities, achieving non-trivial performance on downstream examples in languages seen only during pretraining. For example, in the bottom of Figure 1, a named entity recognition model finetuned on Russian is capable of predicting correctly name entity tags for texts in English, seen only during pretraining (Pires et al., 2019; Conneau et al., 2020b; K et al., 2020; Conneau et al., 2020a; Lazar et al., 2021; Turc et al., 2021).

Previous analyses examined how several factors contribute to this emerging behavior. For example, parameter sharing and model depth are important in certain configurations (K et al., 2020; Conneau et al., 2020b), as well as typological similarities

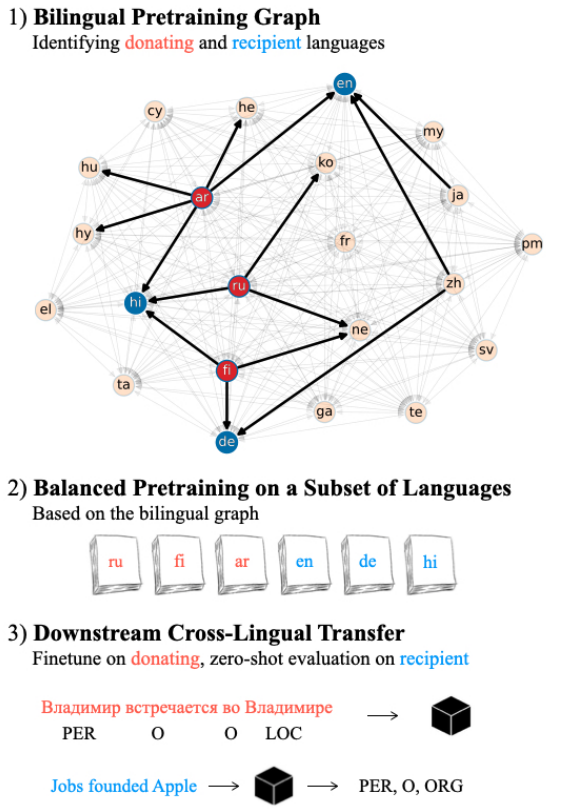


Figure 1: We build a complete, directed graph over a diverse set of 22 languages. Weighted edges show the improvement of bilingual LM over monolingual performance (bold edges represent larger weights). Languages which consistently improve performance are termed "donors" and marked in red, while languages which benefit most are termed "recipients" (marked in blue). We show that our observations hold in several configurations on two downstream tasks.

between languages (Pires et al., 2019), and the choice of specific finetune languages (Turc et al., 2021).

In this work, we focus on an important factor that we find missing in prior work, namely the effect that *pretraining* languages have on downstream zero-shot performance. In particular, we ask three major research questions: (1) Does the choice

[*] Work done while visiting the Hebrew University.
[1] Code and models are publicly available at: github.com/SLAB-NLP/linguistic-blood-bank

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4903 - 4915
July 10-15, 2022 ©2022 Association for Computational Linguistics

of pretraining languages affect downstream cross-lingual transfer, and if so, to what extent? (2) Is English the optimal pretraining language, when controlling for confounding factors such as data size and domain? And finally, (3) Can we choose pretraining languages to improve downstream zero-shot performance?

In addressing these research questions, we aim to decouple the *language* from its corresponding *dataset*. To the best of our knowledge, prior work has conflated pretrain corpus size and its domain with other examined factors, thus skewing results towards over-represented languages, such as English or German (Joshi et al., 2020).[2] To achieve this, we first construct a *linguistically-balanced* pretraining corpus based on Wikipedia, composed of a diverse set of 22 languages. We carefully control for the amount of data and domain distribution in each of the languages (Section 3).

Next, since the number of pretraining configurations grows exponentially with the number of languages n represented in the dataset, it is infeasible to exhaustively test all possible configurations, much less extend it for more languages.[3] In Section 4 we propose a novel pretraining-based approach that is quadratic in the number of languages. This is achieved by training all $\binom{n}{2}$ combinations of bilingual masked language models over our corpus, thus yielding a complete directed graph (Figure 1), where an edge $l_1 \rightarrow l_2$ estimates how much a language l_1 contributes to zero-shot performance in language l_2, based only on language modeling performance.

In Section 5, we use the graph to identify languages which generally contribute as pretraining languages (termed "donors"), and languages which often benefit from training with other languages (termed "recipients"). Further, we use the graph to make observations regarding the effect of typological features on bilingual language modeling, and make available an interactive graph explorer.

Finally, our evaluations on two multilingual downstream tasks (part of speech tagging and named entity recognition) lead to three main conclusions (Section 6): (1) the choice of pretraining languages indeed leads to differences in zero-shot performance; (2) controlling for the amount of data allotted for each language during pretraining questions the primacy of English as the main *pretraining* language; and (3) our hypotheses regarding donors and recipient language hold in both downstream tasks, and against two additional control groups.

[2] For example, English was X100 more likely to be sampled in mBERT's pretraining data than Icelandic.

[3] There are 2^n possible pretraining configurations taking into account inclusion and omission of every language.

2 Metrics for Pretraining-Aware Cross-Lingual Transfer

In this section, we extend existing metrics for zero-shot cross-lingual transfer to account for *pretraining* languages. Intuitively, our metrics for a model M and a given downstream task take into account three factors: (1) P, the set of languages seen during pretraining, (2) $s \in P$, the *source* language used for finetuning, and (3) $t \in P$, the *target* language, seen during inference.

Formally, we adapt the formulation of Hu et al. (2020) to define a *pretraining-aware* bilingual zero-shot transfer score $\mathcal{Z}$ as:[4]

$$\mathcal{Z}_P(s \rightarrow t) := \varepsilon(M^{P,s}, t) \quad (1)$$

Where $M^{P,l}$ is a model pretrained on the set of languages P and finetuned on downstream task instances in the language $l \in P$, and $\varepsilon(M, l)$ is an evaluation of model M on instances in language l in terms of the downstream metric, e.g., word-label accuracy for part of speech tagging.

Following, we extend the definition of zero-shot transfer score to a set of downstream test languages $D \subseteq P$ to measure P's aggregated effect on zero-shot performance, by averaging over all bilingual transfer combinations in D:

$$\mathcal{Z}_P(D) = \frac{1}{|D|^2 - |D|} \cdot \sum_{\substack{l_1, l_2 \in D \\ l_1 \neq l_2}} \mathcal{Z}_P(l_1 \rightarrow l_2) \quad (2)$$

In the following sections, we will use these metrics to evaluate how different choices for pretraining languages influence downstream performance.

3 Data Selection

We collect a pretraining dataset to test the effect of pretraining languages on cross-lingual transfer.

First, we choose a set of 22 languages from 9 language families, as listed in Table 1. These represent a wide variety of scripts, as well as typological

[4] We opt not to normalize the score by the monolingual performance as done in Turc et al. (2021), as we do not want it to affect the score.

and morphological features. We note that our approach can be readily extended to other languages beyond those selected in this study.

Second, we aim to balance the amount of data and control for its domain across languages, to mitigate possible confounders in our evaluations. Below we outline design choices we make toward this goal.

3.1 Data Balancing

To achieve a balanced dataset across our languages, we sample consecutive sentences from every language's Wikipedia dump from November 2021, such that each language is represented by 10 million characters.[5] This amount was chosen to align all languages to the lower-resource ones (e.g., Piedmontese or Irish) which comprise approximately of 10mb. We choose to sample texts from Wikipedia as it consists of roughly similar encyclopedic domain across languages, and is widely used for training PLMs (Devlin et al., 2019).

Can we balance the amount of information across languages? We note that a possible confound in our study is that languages may encode different amounts of *information* in texts of similar character count. This may happen due to differences in the underlying texts or in inherent language properties.[6] To estimate the amount of information in each of our 10^7 character partitions, we tokenize each language partition l with the same word-piece tokenizer, and look at the ratio between the total number of tokens in l and the number of unique tokens in l, finding a good correlation across all our languages ($r = 0.73$), which may indicate that our dataset is indeed balanced in terms of information. Our intuition is that an imbalanced amount of information would lead the tokenizer to "invest" more tokens in some of the languages while neglecting the less informative ones.

Is our sample representative of the full Wikipedia corpus in each language? Another concern may be that our sampled corpus per language is not indicative of the full corpus for that language, which may be much larger (see Table 1). To test this, we create three discrete length distributions. Two length distributions for sentences (in terms of words and tokens), and word length distribution in terms of characters. We then compare those three distributions between our sample and the full data using Earth Movers Distance. All means and standard deviations score below 0.001, indicating that indeed all samples are similarly distributed to their respective full corpus in terms of these metrics.

Language	Code	Family	Size [M chars] Wiki	Sample
Piedmontese	pms	Indoeuropean	14	10
Irish	ga	Indoeuropean	38	10
Nepali	ne	Indoeuropean	78	10
Welsh	cy	Indoeuropean	85	10
Finnish	fi	Uralic	131	10
Armenian	hy	Indoeuropean	174	10
Burmese	my	Sino-Tibetian	229	10
Hindi	hi	Indoeuropean	473	10
Telugu	te	Dravidian	533	10
Tamil	ta	Dravidian	573	10
Korean	ko	Korean	756	10
Greek	el	Indoeuropean	906	10
Hungarian	hu	Uralic	962	10
Hebrew	he	Afroasiatic	1,261	10
Chinese	zh	Sino-Tibetian	1,546	10
Arabic	ar	Afroasiatic	1,695	10
Swedish	sv	Indoeuropean	1,744	10
Japanese	ja	Japonese	3,288	10
French	fr	Indoeuropean	4,958	10
German	de	Indoeuropean	6,141	10
Russian	ru	Indoeuropean	6,467	10
English	en	Indoeuropean	14,433	10

Table 1: The size of the full Wikipedia dump for the languages in our study (in millions of characters) versus our fixed sized sampling of it. This exemplifies both the linguistic diversity as well as the variance in data sizes in the original Wikipedia corpus, often used for pretraining PLMs. In contrast, we create a balanced pretraining dataset by sampling 10M characters from all languages such that they conform to the smallest language portion in our set (Piedmontese).

4 Bilingual Pretraining Graph

In this section, we describe a method for estimating the effect that different pretrain language combinations would have on downstream zero-shot performance. This is achieved by evaluating bilingual performance on the pretraining masked language modeling (MLM) task.

We begin by describing our experimental setup, hyperparameters and hardware configuration (Section 4.1). In Section 4.2, we outline our estimation method, yielding a complete graph structure over our languages, amenable for future exploration and analyses (Figures 1, 2). In the following sections,

[5] Wikipedia dump was obtained and cleaned using wikiextractor (Attardi, 2015).

[6] For example logographic or abjad writing systems may be more condensed than other scripts (Perfetti and Liu, 2005).

	ar	cy	de	el	en	fi	fr	ga	he	hi	hu	hy	ja	ko	my	ne	pms	ru	sv	ta	te	zh	Don.
ar	0	12	6	-3	38	0	-2	1	26	66	44	35	6	11	-3	4	4	-18	11	-16	11	-0	233
cy	-5	0	42	-5	37	-2	28	17	25	50	39	28	6	9	36	66	3	4	18	2	9	10	417
de	-28	-0	0	-28	-4	-25	-4	-1	7	-7	0	-7	1	7	-2	1	1	-17	12	-4	-1	-0	-100
el	-5	12	41	0	38	-4	-5	20	4	-7	0	-7	13	8	-0	78	-1	3	13	-0	-1	0	199
en	-28	0	0	-31	0	-7	-5	12	1	-7	6	18	0	1	-3	47	0	-8	11	-11	-1	-0	-5
fi	-28	0	51	-7	-2	0	-2	-1	29	60	0	33	6	8	41	71	4	-17	-0	2	10	12	272
fr	-27	0	31	0	-3	-1	0	15	22	48	1	25	2	7	-4	2	3	16	11	-2	7	12	166
ga	-29	-0	6	-28	-2	-4	-4	0	7	-7	43	-18	4	8	35	3	-0	3	16	0	8	10	50
he	-5	10	38	-6	30	-25	-2	16	0	51	37	27	5	10	-3	62	3	-18	14	-16	-1	9	234
hi	-28	-0	-1	-31	-4	-25	15	-1	1	0	26	-7	-2	2	-3	1	0	-18	-0	-16	5	-0	-86
hu	-12	0	39	-28	-2	-25	-2	13	19	39	0	15	-0	6	-3	1	-1	-2	-0	-15	5	10	57
hy	-29	-4	-1	-15	-2	-10	14	9	1	-7	29	0	4	2	-1	1	1	-8	6	-8	-2	11	-9
ja	2	14	48	1	40	2	32	23	29	70	48	40	0	16	43	83	5	7	20	6	11	13	555
ko	-29	12	43	-3	34	-22	28	19	25	67	0	34	-1	0	-3	76	4	7	17	-11	-1	-0	295
my	-28	4	24	-16	-2	-22	-4	-1	14	37	27	-7	2	1	0	2	1	-4	7	-9	2	5	34
ne	-22	2	18	-23	14	-16	7	7	9	27	17	-3	1	1	-3	0	0	-13	6	-12	4	-0	19
pms	-28	-0	43	-4	33	-25	29	19	1	-7	1	27	0	0	-3	63	0	3	16	-15	9	1	165
ru	-2	0	-1	-31	36	-2	-5	1	7	61	6	37	6	14	42	75	4	0	-3	3	11	11	271
sv	-16	10	41	-28	33	-6	-4	1	25	30	35	8	0	6	-3	36	6	-7	0	-8	3	6	168
ta	-3	12	6	-5	-2	-4	28	19	26	-7	40	-18	4	11	37	3	2	4	-3	0	11	10	171
te	-29	9	41	-5	36	-5	27	17	25	51	36	-7	5	12	33	56	2	2	14	1	0	9	331
zh	-1	15	50	-0	41	1	32	22	37	67	47	-18	8	15	-3	84	8	-17	17	4	12	0	421
Recp.	-381	110	566	-296	388	-225	201	227	338	679	482	233	67	154	231	816	50	-100	201	-125	110	130	

Figure 2: Bilingual finetune scores between language pairs in our *balanced* corpus. Coordinate (i, j) represents $\mathcal{F}(l_i \rightarrow l_j)$, i.e., the performance in MRR[%] (which correlates with perplexity) of an LM pretrained on a bilingual corpus over languages (l_i, l_j) and tested intrinsically on l_j. The last column (marked *Don.*) sums over each line, i.e., index i in the column represents how much language i *donated* to all other languages. Similarly, the j'th index in the last row (marked *Recp.*) sums over column j and represents how much language l_j improved in all configurations.

we use the graph to formulate a set of downstream cross-lingual hypotheses regarding how different languages will affect zero-shot performance, and validate these hypotheses on two downstream tasks.

4.1 Experimental Setup

For all evaluations discussed below, we train a BERT model (Devlin et al., 2019) with 4 layers and 4 attention heads, an MLM task head, and an embedding dimension of 512.[7] We train a single wordpiece tokenizer (Wu et al., 2016) on our entire dataset.[8] We train the models with a batch size of 8 samples, with sentences truncated to 128 tokens.

Each language model was trained up to 4 epochs. This was determined by examining the training loss on 6 diverse languages in our set and observing that they converge around 4 epochs. A subset of 6 languages was trained on 4 additional seeds to verify the stability of the results, as seen in Table 5 and Table 6 in the Appendix. Masks were applied with default settings, generating 15% mask tokens and 10% random tokens for each input sequence (Devlin et al., 2019). We used a single GPU core (nvidia tesla M60, gtx 980, and RTX 2080Ti). Training time varied between 80 - 120 minutes.

4.2 Building a Pretraining Language Graph

Intuitively, we measure MLM performance when pretraining on a pair of languages (l_1, l_2) as a proxy to the extent of how l_1 and l_2 contribute to one another in zero-shot cross-lingual transfer.

This methodology relies on two assumptions. First, we assume that the cross-lingual zero-shot performance as defined in Equation 2 is *monotonic*, i.e., that adding pretraining languages will improve the average downstream performance. This is defined formally as:

$$P' \subseteq P \quad \Rightarrow \quad \mathcal{Z}_{P'}(D) \leq \mathcal{Z}_P(D) \qquad (3)$$

[7] We use the implementation provided by Hugging Face: https://huggingface.co/bert-base-uncased.

[8] To allow future exploration, we also tokenize over 22 additional languages (listed in the Appendix) which are sampled in the same manner but are not included in this study.

Following this assumption will allow us to extend our bilingual observations to a pretraining language set P of arbitrary size.

Second, we assume that MLM performance correlates with downstream task performance, which is often the assumption made when training PLMs to minimize perplexity (Peters et al., 2018; Devlin et al., 2019).

Bilingual MLM finetune score. Formally, for every language pair $s, t \in P$, we compute the following finetune score, $\mathcal{F}$:

$$\mathcal{F}(s \to t) := \frac{\varepsilon(M^{s,t}, t) - \varepsilon(M^t, t)}{\varepsilon(M^t, t)} \tag{4}$$

Where $M^{s,t}$ is a model pretrained on s, t, and ε is an intrinsic evaluation metric for MLM.[9] I.e., $\mathcal{F}(s, t)$ estimates how much the target language t "gains" in the MLM task from additional pretraining on the source language s compared to monolingual pretraining on t.

Figure 2 depicts a weighted adjacency matrix where coordinate (i, j) corresponds to $\mathcal{F}(l_i \to l_j)$. As shown in Figure 1, the same information can be conveyed in a complete directed weighted graph, where each node represents a language, and edges (l_1, l_2) are weighted by $\mathcal{F}(l_1 \to l_2)$.

Language-Level donation and recipience. Next, for each language $l \in P$ we compute a *Donation* score, $\mathcal{D}$, as an aggregate over all of its finetune scores as a source language (i.e., how much it contributed to other languages), and similarly a *recipience* score, $\mathcal{R}$, by aggregating over all its finetune scores as a target language, to measure how much l is contributed to by other languages. Formally:

$$\mathcal{D}(l) := \sum_{\substack{t \in P \\ t \neq l}} \mathcal{F}(l \to t) \tag{5}$$

$$\mathcal{R}(l) := \sum_{\substack{s \in P \\ s \neq l}} \mathcal{F}(s \to l) \tag{6}$$

We depict both donation and recipience scores as aggregate row and column vectors in Figure 2.

Thus, based on the two assumptions above, our hypothesis is that the downstream cross-lingual transfer will be proportional to the sum of recipience scores for all pretraining languages. Formally:

[9] We specifically use mean reciprocal rank (MRR), which correlates with perplexity.

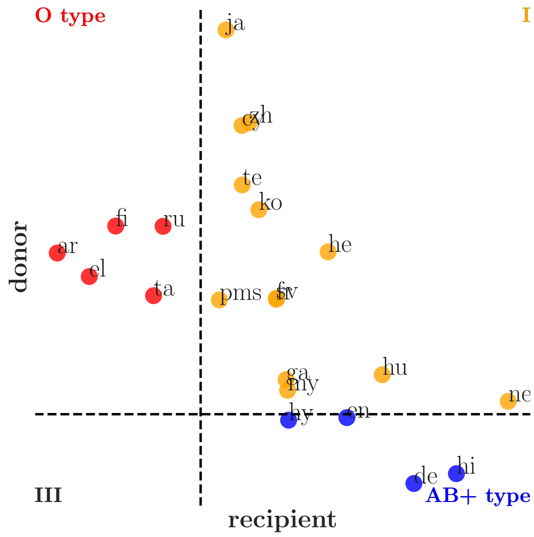


Figure 3: Our languages on a "donor" versus "recipient" axes. A positive coordinate on the "donor" score (X axis) represents a language that on average improved other languages' performance in bilingual pretraining, while a negative score indicates a language which hurts other languages on average. Inversely, a positive score on the Y axis represents languages whose performance was improved by bilingual pretraining, while negative scores represent languages whose performance was hurt by it. The II quadrant represents O type languages (donating but not receiving), languages on the IV's quadrant are AB+ type languages (receiving but not donating)

$$\mathcal{Z}_P(D) \propto \sum_{l \in D} \mathcal{R}(l) \tag{7}$$

Moreover, higher donation scores for languages in the pretrain set will result in higher scores in the downstream task. Formally:

$$\sum_{l \in P} \mathcal{D}(l) \le \sum_{l \in P'} \mathcal{D}(l) \Rightarrow \mathcal{Z}_P(D) \le \mathcal{Z}_{P'}(D) \tag{8}$$

5 Pretraining Graph Analysis

We present several key observations based on the bilingual pretraining graph described in the previous section and summarized by the adjacency matrix in Figure 2, as well as an interactive exploration interface. In the following sections, we use these observations in our downstream evaluations.

Some language combinations are detrimental. Negative finetune scores are present in some of the target languages, e.g., between Korean (ko) and Arabic (ar), which means that initializing a

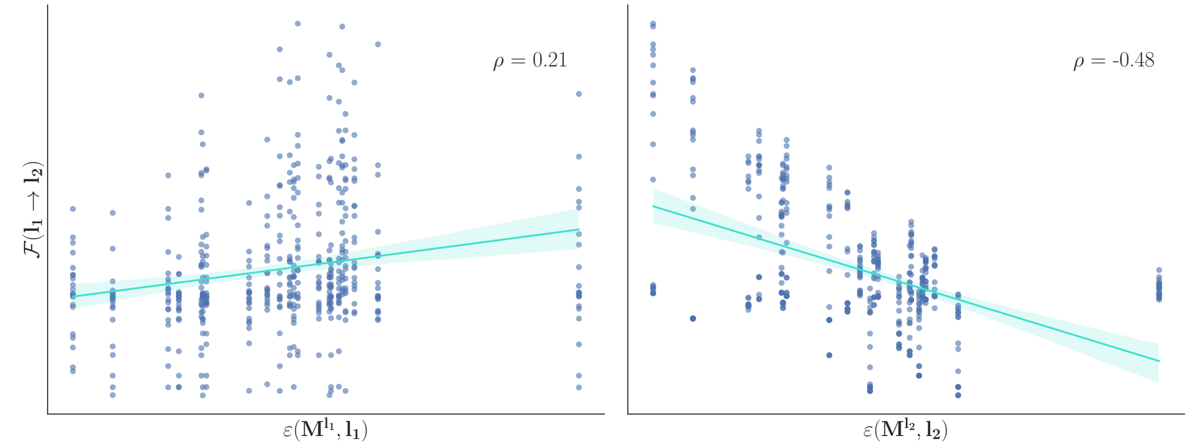


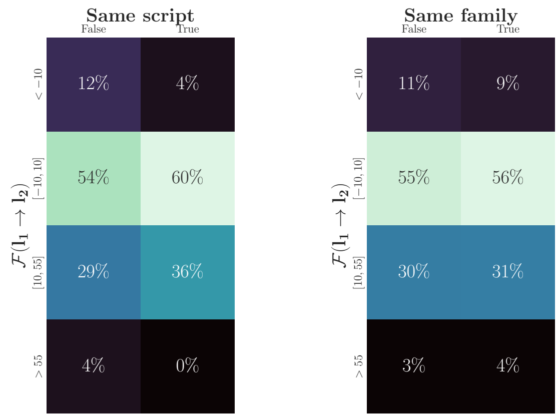
Figure 4: Scatter-plot. Y-axis represents cross-lingual transfer $\mathcal{F}(l_1 \to l_2)$ for a each possible pair of languages, while the x-axis represents the monolingual MRR score for a source language (left) and the target language (right).

language model for Arabic with weights learned for Korean hurts MLM performance on Arabic, compared to an Arabic monolingual baseline. I.e., in these language configurations, initializing the model with another language model's weights leads to worse performance than random initialization.

Bilingual MLM relations are *not* symmetric. In fact, we observe a moderate *negative* correlation between $\mathcal{F}(l_1 \to l_2)$ and $\mathcal{F}(l_2 \to l_1)$, as shown in Figure 3. For example, for German and Finnish we get $0.51 = \mathcal{F}(fi \to de) > \mathcal{F}(de \to fi) = -0.24$. I.e., Finnish initialization improves German MLM, while the inverse is detrimental for Finnish.

Monolingual performance correlates with donation score. Perhaps expectedly, relatively worse-performing models benefit most from the bilingual transfer, while better-performing monolingual models tend to be better donors, although to a lesser extent (Figure 4).[10]

Different script leads to larger variance in bilingual finetuning. However, language family does not affect it. We find that fine-tuning between languages with different scripts is a high-risk, high-reward scenario. The highest transfer scores occur in this setting, but the proportion of negative scores is also higher. A shared script is a safe setting with a high proportion of neutral or positive donations (Figure 5a). In contrast with recent findings (Pires et al., 2019), we did not observe a statistically significant influence for the language family (Figure 5b).



(a) Sharing Script

(b) Sharing Family

Figure 5: We divide language pairs into four bins by bilingual finetune score ($\mathcal{F}(l_1 \to l_2)$).[11] The figures present the percentage of pairs assigned to each bin for samples of language pairs: (a) written in the same or different script; (b) belonging to the same or different language family. Sharing the language family has no significant effect on the transfer score ($p > 0.05$), while the effect of sharing scripts is significant ($p < 0.05$) (p-values based on Pearson's χ^2 test).

Finetuning as transfusion: mapping the linguistic blood-bank. The non-symmetric nature of the scores gives rise to a coarse-grained ontology loosely reminiscent of human blood types, depicted in Figure 3. Languages which on average donate but do not receive ($\mathcal{D}(l) > 0$ and $\mathcal{R}(l) < 0$) are

[10] Correlations are statistically significant ($p < 0.05$ based on Student's T-test).

[11] We motivate our choice of bins in Appendix.

denoted *O type languages*, while the inverse (receiving but not donating) are denoted as *AB+ type*.

5.1 Interactive Exploration

To allow further exploration of our bilingual pretraining graph, we develop a publicly available web-based interactive exploration interface.[12] We enable exploration of interactions between different linguistic features, based on *The World Atlas of Language Structures* (WALS) (Dryer and Haspelmath, 2013), allowing users to filter and focus on specific traits and analyze how they affect bilingual pretraining.

6 Downstream Zero-Shot Performance

In this section, we validate our method for estimating the effect of pretraining language combinations on downstream performance. Towards that end, in Section 6.1, we construct several pretraining configurations, based on pretraining observations. Then, in Section 6.2 we describe the multilingual datasets we use for two downstream tasks. Finally, our results are presented in Section 6.3, showing the influence of pretraining configuration on downstream performance.

6.1 Choosing Pretraining Sets

We use the donation scores to identify pretraining languages projected to lead to better downstream zero-shot performance, and the recipience score to find downstream languages which will perform better languages as source (finetune) languages. Our setup is summarized in Table 2.

Donating languages. We define three sets of languages for pretraining, using the donation score while keeping the sets linguistically diverse: (1) *Most Donating:* Japanese, Telugu, Finnish, and Russian; (2) *Least Donating:* Nepali, Burmese, Armenian, and English. We also include Englishs as it is a popular source language; and (3) *Random:*A randomly selected set of 4 languages: Hebrew, Irish, French and Swedish.

Recipient languages. To validate that lower recipience scores indeed indicate that languages are less likely to improve via cross lingual transfer, we added 6 languages to all configurations described above: 3 *Most Recipient* languages (R_h): Hindi, German, and Hungarian, and 3 *Least Recipient* languages (R_l): Arabic, Greek, and Tamil. Finally, we add a fourth control configuration which was pretrained only on $C := R_h \cup R_l$.

Hypotheses. We hypothesize that the more donating pretraining sets will improve cross-lingual transfer in downstream tasks, and that more recipient languages will have better cross-lingual performance compared to least recipient languages. These can be formally articulated using Equations 9 and 10:

$$\forall P : \mathcal{Z}_P(R_h) > \mathcal{Z}_P(R_l) \quad (9)$$

$$\mathcal{Z}_{MostDon.}(C) > \mathcal{Z}_{Random}(C) > \mathcal{Z}_{LeastDon.}(C) \quad (10)$$

6.2 Tasks

We evaluated all pretraining configurations detailed in Table 2 on two of XTREME's tasks: part of speech tagging (POS) and named entity recognition (NER). Both of which commonly appear in NLP pipelines such as CoreNLP (Manning et al., 2014) and spaCy (Honnibal and Montani, 2017). We aim to balance the data in both tasks across different finetune languages, so as not to skew results towards higher-resource languages.

For part-of-speech tagging, XTREME borrows from universal dependencies (Nivre et al., 2020). Since XTREME is imbalanced across languages, we truncated the data to 1000 sentences to fit the lower-resource languages, e.g., XTREME annotates POS in 909 sentences in Hungarian. For NER, we applied a similar procedure, where XTREME's data was taken from the Wikiann (panx) dataset (Rahimi et al., 2019) which we truncated to 5000 sentences (the data size available for Hindi NER in XTREME).

Experimental setup. We use the code and default hyperparameter default values provided by XTREME to train the downstream tasks (Hu et al., 2020), adapted for multilingual training.

6.3 Results

Several key observations can be made based on the results for both POS tagging and NER across all training configurations, which are presented in Tables 3 and 4. For each configuration P in *Most Donating, Least Donating, Random, Control* we

[12] github.com/SLAB-NLP/linguistic-blood-bank#interactive-exploration

	Base Pretrain Set (Donors)		Shared Pretrain Set Most Recipient (R_h)		Least Recipient (R_l)	Total Data	Summary
Most Donating	{ja, te, fi, ru}	+	{hi, de, hu}	+	{ar, el, ta}	10^8 characters	Most donating pretraining set.
Least Donating	{ne, my, hy, en}	+	{hi, de, hu}	+	{ar, el, ta}	10^8 characters	Least donating pretrain set.
Random	{he, ga, fr, sv}	+	{hi, de, hu}	+	{ar, el, ta}	10^8 characters	Random donating pretrain set.
Control	{}	+	{hi, de, hu}	+	{ar, el, ta}	10^8 characters	No additional donating languages.

Table 2: Four pretraining language configurations. Each consists of *donating* languages (first column) and *recipient* languages (second column). The control group has the same amount of data, equally distributed among its languages.

	NER [$\%F_1$]		**POS** [$\%F_1$]	
	Avg. Monolingual	Avg. Zeroshot	Avg. Monolingual	Avg. Zeroshot
Most Donating	**49.3**±.4	**15.6**±.4	**61.4**±.1	**28.1**±.3
Random	**49.2**±.3	**15.6**±.1	61.3±.1	26.9±.3
Least Donating	48.8±.2	14.8±.3	60.9±.2	26.9±.6
Control	49.0±.2	**15.6**±.2	**61.9**±.1	27.4±.3

Table 3: Donation results for named entity recognition (NER) and part of speech tagging (POS) as mean and standard deviation over five random seeds. For each pretraining language group (*Most Donating*, *Random*, *Least Donating*, and *Control*), we report corresponding average monolingual and zero shot performance. *Most Donating* consistently outperforms *Least Donating* in both tasks, and in both monolingual and zeroshot performance. *Most Donating* is on par with *Control* in monolingual performance in NER, despite having less in-domain data.

	NER [$\%F_1$]		**POS** [$\%F_1$]	
	Avg. Monolingual	Avg. Zeroshot	Avg. Monolingual	Avg. Zeroshot
Most Recipient (R_h)	**50.3**±.6	**18.4**±.6	**64.1**±.3	**28.7**±.7
Least Recipient (R_l)	47.9±.4	12.4±.4	58.6±.4	26.0±.7

Table 4: Recipience results for named entity recognition (NER) and part of speech tagging (POS) as mean and standard deviation over five random seeds. We report results across different training configurations for two groups of downstream recipient languages. In accordance with our pretraining results, the *Most Recipient* set does better than the *Least recipient* set across both tasks in zero-shot and monolingual performance.

calculated zero-shot transfer scores on C, using $\mathcal{Z}_P(C)$ defined by Equation 2. Monolingual results under each pretrain set P were calculated by the average $F1$ performance of each language in C:

$$\frac{1}{|C|} \cdot \sum_{l \in C} \varepsilon(M^{P,l}, l) \qquad (11)$$

Where $\varepsilon(M^{P,l}, l)$ denotes the $F1$ score of a model pretrained on P, finetuned on l and evaluated on l.

Pretraining configuration affects downstream cross-lingual transfer. In both tasks, we observe a variance in results when changing the pretraining configuration, despite all of them having similar amounts of data. This may imply that previous work has omitted an important interfering factor.

Recipience score correlates with downstream cross-lingual performance. We evaluated zero-shot transfer for each language set $R \in \{R_l, R_h\}$ as the average zero-shot transfer scores over all pretraining configurations. Table 4 reveals that the *Most Recipient* set outperforms the *Least Recipient* set in both tasks ($+5.5\%$ in NER, $+2.7\%$ in POS tagging).

Multilingual pretraining can improve *mono-lingual* performance. As seen in Table 3, the *Most Donating* pretraining configuration achieved a monolingual score which is slightly higher than the control group, while the *Least Donating* configuration underperforms all other sets. This suggests that multilingual pretraining datasets can benefit monolingual downstream results compared to more data in a single language.

English might not be an optimal pretraining language. Corresponding with our previous results, if donation score is indicative of a language's contribution in pretraining, English's relative low donation score might indicate that it is not the best language to pretrain upon. English was also part of the *Least Donating* pretraining configuration which scored lower than *Most Donating* as seen in Table 3. Further research can ascertain this finding.

7 Limitations and Future Work

As with other works on cross-lingual transfer, our results are influenced by many hyperparameters. Below we explicitly define our design choices and how they can be explored in future work.

First, data scarcity in low-resource languages restricted us to small data amounts. Although our experiments showed a non-trivial signal for pretraining and downstream tasks, future work may apply our framework to larger data sizes.

Second, for efficiency's sake, we trained relatively small models to enable us to train a large number of language configurations, while ensuring convergence in 6 languages. Furthermore, we did not do any hyper-parameter tuning and used only values reported in previous work, and use only the BERT architecture. Future work may revisit any of these design choices to shed more light on their effect.

Third, similarly to other works, our data was scraped from Wikipedia, and we did not account for language contamination across supposedly monolingual corpora (e.g., due to code switching). Such contamination may confound with cross-lingual transfer, as was recently shown by Blevins and Zettlemoyer (2022).

Finally, our downstream analysis focused on POS tagging and NER since they were available for many languages and are common in many NLP pipelines. Further experimentation can test if our results hold for more NLP tasks.

8 Related Work

To the best of our knowledge, this is the first work to control for the amount of data allocated for each language during pretraining and finetuning while evaluating on many languages.

Perhaps most related to our work, Turc et al. (2021) challenge the primacy of English as a source language for cross-lingual transfer in various downstream tasks. Their work shows that German and Russian are often more effective sources. In all of their experiments, they use mBERT's imbalanced pretraining corpus. Blevins and Zettlemoyer (2022) complement this hypothesis by showing that English pretraining data actually contains a significant amount of non-English text, which correlates with the model's transfer capabilities.

Wu and Dredze (2020) evaluate how mBERT performs on a wide set of languages, focusing on the quality of representation for low-resource languages in various downstream tasks by defining a scale from low to high resource. They show that mBERT underperforms non BERT monolingual baselines for low resource languages while performing well for high resource ones.

While Pires et al. (2019); Limisiewicz and Mareček (2021) show that typology plays a significant role for mBERT's multilingual performance, this is not replicated in our balanced evaluation, and has lesser impact in Wu et al. (2022) as well.

Finally, Conneau et al. (2020a) introduce the transfer-interference trade-off where low resource languages benefit from multilingual training, up to a point where the overall performance on monolingual and cross-lingual benchmarks degrades.

9 Conclusions

We explored the effect of pretraining language selection on downstream zero-shot transfer.

We first choose a diverse pretraining set of 22 languages, and curate a pretraining corpus which is balanced across these languages.

Second, we devise an estimation technique, quadratic in the number of languages, projecting which pretraining languages will serve better in cross-lingual transfer and which specific downstream languages will do best in that setting.

Finally, we test our hypothesis on two downstream multilignual tasks, and show that the choice of pretraining languages indeed leads to varying downstream cross-lingual results, and that our method is a good estimation for downstream performance. Taken together, our results suggest that pretraining language selection should be a factor in estimating cross-lingual transfer, and that current practices which focus on high-resource languages may be sub-optimal.

Acknowledgements

We would like to thank Roy Schwartz for his helpful comments and suggestions and the anonymous reviewers for their valuable feedback. This work was supported in part by a research gift from the Allen Institute for AI. Tomasz Limisiewicz's visit to the Hebrew University has been supported by grant 338521 of the Charles University Grant Agency and the Mobility Fund of Charles University.

References

Giusepppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination explains the cross-lingual capabilities of english pretrained models. *ArXiv preprint*, abs/2204.08110.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Koren Lazar, Benny Saret, Asaf Yehudai, Wayne Horowitz, Nathan Wasserman, and Gabriel Stanovsky. 2021. Filling the gaps in Ancient Akkadian texts: A masked language modelling approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4682–4691, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tomasz Limisiewicz and David Mareček. 2021. Examining cross-lingual contextual embeddings with orthogonal structural probes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4589–4598, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Charles A Perfetti and Ying Liu. 2005. Orthography to phonology and meaning: Comparisons across and within writing systems. *Reading and Writing*, 18(3):193–210.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. 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.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *ArXiv preprint*, abs/2106.16171.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv preprint*, abs/1609.08144.

Zhengxuan Wu, Isabel Papadimitriou, and Alex Tamkin. 2022. Oolong: Investigating what makes crosslingual transfer hard with controlled studies. *ArXiv preprint*, abs/2202.12312.

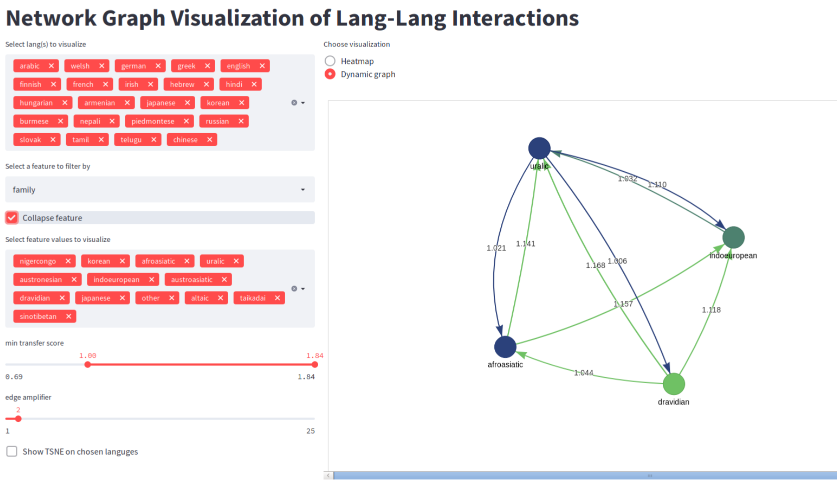


Figure 6: Our visualization tool, based on Streamlit (https://streamlit.io)

A Appendix

Full list of tokenized languages The full list of Wikipedia language codes for languages used in our tokenizer training is:

- pms, ga, ne, cy, fi, hy, my, hi, te, ta, ko, el, hu, he, zh, ar, sv, ja, fr, de, ru, en - languages that are also evaluated and trained. Elaborated in Table 1.
- af, am, ca, cs, da, es, id, is, it, mg, nl, pl, sk, sw, th, tr, ur, vi, yi - Additional languages corresponding to Afrikaans, Amharic, Catalan, Czech, Danish, Spanish, Indonesian, Icelandic, Italian, Malagasy, Dutch, Polish, Slovak, Swahili, Thai, Turkish, Urdu, Vietnamese, and Yiddish.

Transfer Distribution In the histogram of cross-lingual transfers (Figure 7), we observe that the distribution has multiple local maximums (modes). We distinguish four main level of cross-lingual transfer described in Section 4.2 ($\mathcal{F}(l_i \to l_j)$):

- negative transfer $\mathcal{F}(l_i \to l_j) < -10$
- neutral transfer $-10 \leq \mathcal{F}(l_i \to l_j) < 10$
- positive transfer $10 \leq \mathcal{F}(l_i \to l_j) < 55$
- very positive transfer $55 \leq \mathcal{F}(l_i \to l_j)$

The choice of division borders was done in order to separate distinct modes of the distribution and to obtain interpretable bins (e.g. neutral transfer centered around zero).

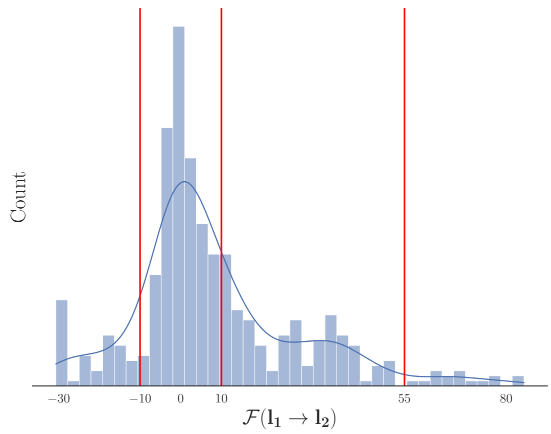


Figure 7: Histogram of cross-lingual transfers $\mathcal{F}(l_i \to l_j)$. Horizontal lines (at -10, 10, and 55) are the borders between four transfer levels.

src/trgt	**de**	**en**	**he**	**ne**	**hi**	**ja**
de	0.2801	0.3177	0.2881	0.2231	0.2685	0.3954
en	0.3401	0.2508	0.2761	0.2238	0.2615	0.3927
he	0.3527	0.3295	0.2612	0.2536	0.2912	0.4041
ne	0.3255	0.2861	0.2716	0.1510	0.2531	0.3887
hi	0.3221	0.2981	0.2873	0.2415	0.2083	0.4045
ja	0.373	0.3536	0.3194	0.2825	0.3232	0.3869

Table 5: Averaged MRR scores for five seeds. Bilingual training was done with five seeds over a group of six diverse languages to verify the results are stable. The table shows mean results. The column indicates the source languages, the row indicates the target languages.

src/trgt	**de**	**en**	**he**	**ne**	**hi**	**ja**
de	0.028	0.0031	0.0118	0.0013	0.0103	0.0068
en	0.0062	0.0229	0.0061	0.0054	0.0071	0.0086
he	0.0037	0.0036	0.0113	0.0019	0.0051	0.0023
ne	0.0287	0.0049	0.0047	0.0041	0.0094	0.0078
hi	0.0033	0.0035	0.0113	0.0276	0.0531	0.0196
ja	0.0057	0.0059	0.0062	0.0066	0.0061	0.0029

Table 6: standard deviations for MRR scores over five seeds. Bilingual training was done with five seeds over a group of six diverse languages to verify the results are stable. The table shows the standard deviation of the results. The column indicates the source languages, the row indicates the target languages.

my	**ne**	**de**	**hi**	**en**	**hu**	**hy**	**ar**	**he**	**ru**	**zh**	**ta**	**ko**	**ga**	**ja**	**fi**	**cy**	**te**	**el**	**fr**	**pms**
23.3	24.2	24.7	25.0	25.9	27.2	28.3	32.1	33.4	34.3	36.3	36.4	36.5	36.7	37.6	38.0	39.0	40.0	41.0	41.4	58.9

Table 7: Monolingual results (MRR scores) for all 22 languages in our study, ordered from low to high. Colors coding follows Figure 3 where *O type languages* are marked in red and *AB+ type languages* languages are marked in blue. Monolingual performance explains some of the pretraining contribution, namely recipient languages appear near the low end of the spectrum while donors appear towards the end.