Super donors and super recipients: Studying cross-lingual transfer between high-resource and low-resource languages

Vitaly Protasov[1] **Elisei Stakovskii**[3] **Ekaterina Voloshina**[3*]
Tatiana Shavrina[3*] **Alexander Panchenko**[1,2]
[1]AIRI [2]Skoltech [3]Independent researcher
{protasov, panchenko}@airi.net {eistakovskii, voloshina.e.yu, rybolos}@gmail.com

Abstract

Despite the increasing popularity of multilingualism within the NLP community, numerous languages continue to be underrepresented due to the lack of available resources. Our work addresses this gap by introducing experiments on cross-lingual transfer between 158 high-resource (HR) and 31 low-resource (LR) languages. We mainly focus on extremely LR languages, some of which are first presented in research works. Across 158 $*$ 31 HR–LR language pairs, we investigate how continued pretraining on different HR languages affects the mT5 model's performance in representing LR languages in the LM setup. Our findings surprisingly reveal that the optimal language pairs with improved performance do not necessarily align with direct linguistic motivations, with subtoken overlap playing a more crucial role. Our investigation indicates that specific languages tend to be almost universally beneficial for pretraining (*super donors*), while others benefit from pretraining with almost any language (*super recipients*). This pattern recurs in various setups and is unrelated to the linguistic similarity of HR-LR pairs. Furthermore, we perform evaluation on two downstream tasks, part-of-speech (POS) tagging and machine translation (MT), showing how HR pretraining affects LR language performance.

1 Introduction

According to the Endangered Languages Project (Belew, 2019), more than 3000 languages are at risk of extinction. In recent years, the NLP community has undoubtedly broadened its efforts and presented very ambitious projects (NLLB Team et al., 2022; Bapna et al., 2022) to incorporate more and more languages into practical use. However, even well-known multilingual transformer models (e.g., mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and mT5 (Xue et al., 2021)) and cross-lingual benchmarks (XGLUE (Liang et al., 2020), XTREME (Ruder et al., 2021)) cover only about 100 of the most presented languages.

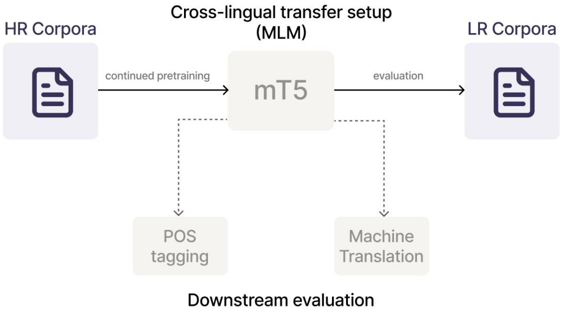

Figure 1: The workflow of cross-lingual transfer between HR and LR languages with further downstream evaluation on POS-tagging and MT tasks.

In this work, we aim to tackle this gap in underrepresentation or even the absence of experiments for LR languages, considering constraints in both labeled and unlabeled text data, as well as linguistic knowledge and experimental base. We present the study of cross-lingual transfer in the case of extremely LR languages, examining how continued pretraining on HR languages impacts model performance on LR languages in the Masked Language Modeling (MLM) setup with additional measurements of its effects on downstream performance (see Figure 1 for more details). We aim to explore whether it is possible to conduct model pretraining on HR languages and observe improvements compared to zero-shot performance when evaluating on unseen LR languages. Additionally, we measure model performance on downstream tasks: POS tagging and MT.

In more detail, we first collect a dataset with raw text data (see Appendix A.3 for the list of lan-

[*]Research was done while at AIRI.

Proceedings of the The Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024), pages 94–108
August 15, 2024 ©2024 Association for Computational Linguistics

guages and Section 3 for criteria of LR and HR languages). Next, we exclude some languages due to data quality issues. Thereby, we experiment with 158 HR and 31 LR languages, resulting in 4898 HR-LR language pairs for further investigation of cross-lingual transfer. Next we assess the zero-shot performance of the mT5 model on the raw data from LR languages. Afterward, we perform continued pretraining of the model with the MLM objective on data from each HR language and evaluate performance of fine-tuned models on all LR languages. Finally, we analyze the factors, such as data and linguistic features, that lead to successful cross-lingual transfer between HR and LR languages. We also measure the downstream performance of successful HR-LR language pairs in POS tagging and MT tasks for LR languages with available annotated data. We do not use data from LR languages during training and use it for evaluation only. We use the term **donor** to denote the languages that serve as sources for knowledge transfer through continued pretraining. On the other hand, we use the term **recipient** to indicate the languages used to evaluate transfer learning efficiency.

The main **contributions** of this work can be summarized as follows: (i) We collect and present the dataset with 189 languages. (ii) We conduct cross-lingual transfer experiments between 158 HR and 31 LR languages. (iii) We interpret cross-lingual transfer results across data and linguistic features. (iv) We investigate how cross-lingual transfer impacts the performance of downstream tasks, focusing mainly on the POS tagging and MT tasks. The code is available[1].

2 Related Work

The cross-lingual transfer involves leveraging existing resources available for HR languages to improve methods for LR languages. This approach can be particularly beneficial for LR languages that lack extensive linguistic resources and data for NLP applications. Wu and Dredze (2020) state that the mBERT's performance for LR languages is not on par with that for HR languages, and there is an unequal representation of languages within models. Libovický et al. (2019) show that mBERT context embeddings capture similarities between languages, but achieving a proper cross-lingual representation requires the availability of parallel corpora, which is lacking for most LR languages. Malkin et al. (2022) show that the selection of pretraining languages significantly impacts the performance, indicating that there are more effective donors than English. Additionally, Turc et al. (2021) show that Russian and German can serve as better donors for reliable transfer. Fujinuma et al. (2022) experiment with different number of languages during pretraining and find it promising in terms of impact on performance on unseen languages. There are also different suggested strategies for choosing the proper donor language. Kocmi and Bojar (2018) propose using vocabulary overlap to find a better HR donor. Lauscher et al. (2020) demonstrate that typological motivation in language selection positively impacts the transfer learning scores, as well as the size of the source language. Muller et al. (2021) show that the type of language script used plays an essential role, and transliteration helps to improve the quality of transfer learning. Eronen et al. (2023) also show that fine-tuning on linguistically similar languages (defined using WALS (Dryer and Haspelmath, 2013) improves the performance on several downstream tasks. Muller et al. (2022) investigate cross-lingual transfer using diverse data, revealing that morphology and language modeling performance are strong predictors of its success. Dolicki and Spanakis (2021); Lin et al. (2019) establish that no individual WALS feature stands out as the most crucial across various tasks.

Transfer learning has long emerged as a pivotal technique in machine translation, particularly for LR languages. Zoph et al. (2016) introduce an approach that uses HR language data to enhance neural machine translation (NMT) for LR language pairs, achieving notable improvements in their translation quality. Further exploration by Aji et al. (2020) reveal that word embeddings are a critical component of transfer learning, and their proper alignment is essential for optimal results. These findings highlight the critical role of transfer learning in addressing challenges associated with the limited availability of linguistic data in NMT.

The studies mentioned above focus on well-resourced languages with labeled data, which has resulted in neglecting LR languages that already lack available data. This study aims to address this gap by investigating cross-lingual transfer for several understudied LR languages.

[1] Code: https://github.com/Vitaly-Protasov/LR_Transfer

3 HR-LR Multilingual Corpus

We assemble from various existing sources a text corpus. Appendix A.3 lists all used languages.

3.1 Text sources

To assemble a corpus for the need of cross-lingual experiments, we use a wide range of linguistic resources in addition to commonly used corpora. We deliberately do not include projects, such as Oscar (Ortiz Suárez et al., 2019) and Cleaned Colossal Common Crawl (Raffel et al., 2020) because they are already partially represented in the training set of large language models such as XLM-R and mT5. The general corpus includes text materials from the following projects: (i) Wikipedia in every language available (CC BY-SA); (ii) Universal Dependencies project[2] (de Marneffe et al., 2021) (original texts without annotation, the license for every treebank is different, mainly GNU GPL 3.0/LGPLLR/CC BY-based); (iii) The Hamburg Center for Language Corpora (HZSK-PUB)[3] (primary linguistic research textual data, not restricted by copyright or personal data protection); (iv) The Endangered Languages Archive[4] (text content only, no multimedia, non-commercial private research or educational activity); (v) Corpora with annotated languages of CIS countries[5] (Krylova et al., 2015).

3.2 Text processing

We aggregate languages according to their official names and codes presented in a large database, The World Atlas of Language Structures (WALS[6]) (Dryer and Haspelmath, 2013). To ensure high data quality for language processing, we exclude languages with a high presence of HTML tags in the collected data, accounting for 15% of the gathered data. We assume that a large amount of code would significantly affect results, as HTML tags are easier to predict than words in natural languages.

We collect both HR and LR languages. We define a LR language based on a specific range of tokens: 10k tokens as a lower bound and 350k tokens as an upper bound (Yang et al., 2019). Thus, we categorize languages exceeding the upper bound as HR ones. Refer to Appendix A.2 for the list of all languages we collected.

4 Cross-lingual Transfer Methodology

The main goal of our experiments is to figure out whether the training on HR language donors improves the modeling of LR recipient languages, try to interpret it and to observe possible performance of transfer learning in downstream tasks.

4.1 Base model

In our experiments, we utilize the widely used pretrained multilingual language model mT5[7](Xue et al., 2021). It is an encoder-decoder model trained on 101 languages from the mC4 dataset. It was originally pretrained in the transfer learning procedure and has shown itself well in transferring knowledge. We think the encoder-decoder architecture is more flexible and has more possible applications for future works than only encoder or decoder-based models. Due to its multitask finetuning, we decided not to use its another version, mT0 (Muennighoff et al., 2023). Considering our lack of labeled data, exploring its multitask zero-shot performance is unnecessary here.

4.2 MLM pretrainig on donor languages

Following the original article of the mT5, we use the Masked Language Modeling (MLM) objective for the continued pretraining on HR donor languages. More specifically, in the case of mT5 model, this is a denoising task for prediction masked spans (sequential set) of tokens. Similarly to the original paper, we also utilize *early stopping*. We limit the data to perform the training for all languages under the same conditions and minimize the training time: 500k sampled sentences are chosen for continued pretraining on each HR language. We conduct this procedure 5 times to consider the variance of results based on different subsets of training data. We then average the results from the top-performing checkpoints within each training iteration. Textual data sourced from HR languages is employed during both the training and validation steps, while LR language data is utilized during the testing step only to measure the performance on unseen LR languages.

4.3 Evaluation on low-resource languages

We use the perplexity metric (Brown et al., 1992) to evaluate the MLM step. Perplexity has its limitations when evaluating language modeling performance, which may become evident in downstream

[2] https://universaldependencies.org

[3] https://corpora.uni-hamburg.de/hzsk

[4] https://www.elararchive.org

[5] http://web-corpora.net

[6] https://wals.info/feature

[7] https://huggingface.co/google/mt5-base

tasks. Since we deeply explore the results of cross-lingual transfer, we also plan to conduct downstream evaluation afterward to see whether the continued pretraining on HR languages impacts the downstream performance on LR languages. Here, we first measure the zero-shot model's performance in modeling all 31 LR languages. Secondly, we use the best model's checkpoints from each run after continued pretraining on different HR language and evaluate them across all LR languages.

4.4 Analysis of transfer learning results

We are also interested in exploring potential factors that affect cross-lingual transfer results, determining whether they lead to success or failure in various language pairs. Following Lin et al. (2019), we utilize linguistic and data-level features to interpret cross-lingual transfer results.

4.4.1 Data-level analysis

Regarding the data level, we calculate the subtoken overlap between languages. We measure the overlap between unique subtokens in HR-LR pairs:

$$o_{12} = \frac{S_1 \cup S_2}{S_2}, \tag{1}$$

where S_1 is the set of unique subtokens of donor (HR) languages, and S_2 is the set of unique target (LR) languages subtokens. In our experiments, we use the mT5 tokenizer. Here, we consider how the subtoken overlap between HR and LR languages relates to the model's performance in LR languages after continued pretraining in donor languages.

4.4.2 Linguistic-level analysis

We investigate language similarity by leveraging their typological characteristics. According to previous works, we consider WALS features. We deliberately avoid relying on GramBank (Skirgård et al., 2023) and `lang2vec` (Littell et al., 2017). GramBank, despite its extensive data coverage, offers features that are quite specific and narrow in scope, resulting in a heterogeneous and uninformative set for our purposes. On the other hand, `lang2vec` represents each feature from the WALS as one-hot encoding vectors, which increases the number of features. This might affect the outcomes of statistical tests due to the interdependence of many of these features. To interpret the results, we employ the Logistic regression model (Hosmer and Lemeshow, 2000) to obtain coefficients associated with input features. These coefficients serve as indicators of the strength in the relationship between input features and target variables learned during model training. This analysis enables us to assess the importance of various language features in achieving successful transfer learning results.

To mitigate biases in absolute perplexity values, we use binary targets to indicate if perplexity decreases (1) or remains unchanged (0) after continued pretraining. Each data point in our training dataset is represented as a binary vector, where every element signifies whether both languages share the same value for a specific linguistic feature (1) or not (0). Finally, we consider the regression coefficients to identify which typological characteristics should be shared between donor and target languages for successful cross-lingual transfer.

4.5 Downstream evaluation

4.5.1 POS tagging task

We consider the POS tagging task since it is one of the few tasks with annotated data available for the LR languages. Specifically, we utilize datasets from UD treebanks. However, only 6 out of the 31 LR languages have data available: *Bambara, Bhojpuri, Cantonese, Coptic, Guarani, Komi-Zyryan.*

To evaluate the cross-lingual transfer performance of different HR-LR pairs in POS tagging, we train logistic regression on top of mT5 embeddings. Our training and validation data come from a donor HR language's training and validation sets extracted from the UD corpus. Afterward, we assess the model's performance on target LR languages' test sets taken from their UD corpora.

4.5.2 Machine translation task

Additionally, we aim to experiment with another downstream task – Machine Translation (MT). Here, we use data from the NLLB project (NLLB Team et al., 2022; Bapna et al., 2022) as the only open-source dataset with MT data for extremely LR languages that we consider. This dataset contains parallel corpora for numerous language pairs, but in the case of LR languages, we see that only a few of them contain parallel HR-LR corpora, and only for two HR languages, such as English and Afrikaans, there are HR-LR datasets: 8 pairs for English and 7 pairs for Afrikaans.

While evaluating the MT setup, we aim to explore how cross-lingual transfer impacts the model's performance on different HR-LR pairs. To do this, we follow a series of steps. First, we conduct separate experiments for each HR-LR pair

LR language	LR perplexity (zero-shot)	HR language (best)	LR perplexity after training
Akan	33.07	Afrikaans	**30.04**
Atikamekw	61.72	Afrikaans	**49.77**
Bambara	51.67	Lithuanian	**38.39**
Bhojpuri	**31.27**	Hindi	113.48
Cantonese	58.27	Slovene	**53.3**
Chichewa	**13.72**	Afrikaans	43.55
Coptic	**4.72**	Afrikaans	10.21
Dagbani	**47.81**	Slovene	57.71
Greenlandic (South)	**35.55**	Afrikaans	39.68
Guaraní	3.99	French	**3.04**
Kashmiri	**26.27**	Lithuanian	34.90
Komi-Zyrian	110.02	Yazva	**66.56**
Koryak	88.66	Slovene	**53.28**
Kurmanji	**32.44**	Afrikaans	66.22
Madurese	33.61	Afrikaans	**31.81**
Nanai	72.91	Slovene	**38.38**
Quiché	165.78	Slovene	**63.78**
Romani (Lovari)	**25.1**	Afrikaans	40.43
Rundi	**21.92**	Afrikaans	33.50
Samoan	**12.52**	Lithuanian	23.88
Sesotho	**12.77**	Afrikaans	26.21
Shor	167.74	Slovene	**98.91**
Sranan	35.44	Afrikaans	**14.09**
Swati	**40.65**	Afrikaans	53.08
Tabassaran	57.19	Slovene	**50.54**
Tat (Muslim)	**70.32**	Afrikaans	82.90
Tofa	62.38	Slovene	**61.98**
Tsakhur	41.74	Slovene	**25.60**
Tsonga	**40.41**	Afrikaans	48.76
Udi	**55.01**	Afrikaans	72.88
Yukaghir (Kolyma)	104.8	Slovene	**68.45**

Table 1: Comparison of zero-shot results of the mT5 model on LR languages with the results after continued pretraining on HR languages. We highlight the best scores among language pairs.

by training the out-of-the-box model in the MT setup and evaluating its performance afterward. This allows us to measure the performance before the cross-lingual transfer. Then, we select the best-performing model checkpoints for each HR donor after continued pretraining and repeat training and evaluation on MT data using these selected checkpoints. Finally, we compare the model's performance on the MT task before and after cross-lingual transfer across various HR-LR pairs.

To ensure consistency in the obtained results, we intend to use identical test sets across all experiments with specific HR-LR pairs. This enables us to compare and analyze the impact of cross-lingual transfer on the MT results more accurately. Also, following the approach we do in Section 4.5.1, we restrict the training data to match the number of tokens available in the least-resourced language pair to ensure a fair comparison.

5 Results and Analysis

5.1 General cross-lingual transfer results

In Table 1, you can see the comparison of the zero-shot results with the best results after continued pretraining on HR donors. Across 16 out of 31 LR languages, continued pretraining resulted in diminished perplexity scores. In this context, «best» refers to the lowest perplexity score attained among all iterations of continued pretraining. We also examine the most effective HR donors. Figure 4 illustrates relative perplexity scores between zero-shot results and results after continued pretraining across the most effective HR-LR pairs (refer to Appendix A.1 and Appendix A.2 for results for all 4898 HR-LR pairs). After pretraining on *Slovene*, the model demonstrates lower perplexity for 14 LR languages. Similar results are observed for *Afrikaans*, also with 14 LR languages, *Lithuanian* with 12, and *French* with 11.

As well as Turc et al. (2021), we observe that *English* may not be the optimal language for cross-lingual transfer. In our experiments, *Afrikaans* and *Slovene* show the best performance in cross-lingual transfer for extremely low-resource languages. Thus, we consider them as «super-donors» in the scope of our experiments. There are also instances where such languages as *Guaraní* and *Coptic* exhibit universal target characteristics after pretraining on various HR languages.

5.2 Correlation with subtoken overlap

To assess how the overlap of subtokens in data affects the performance of different HR-LR language pairs, we calculate the Pearson correlation coefficient between these data features and the results of cross-lingual transfer for these pairs.

We observe a moderate correlation between subtoken overlap and Δperplexity ($r_{stat} = -0.33$, $p_{value} < 0.01$), where Δperplexity is a difference between results before and after continued pretraining. This indicates that as the degree of subtoken overlap grows, we observe a decrease in perplexity on LR languages (see Figure 2 for the distribution of different pairs in such axes).

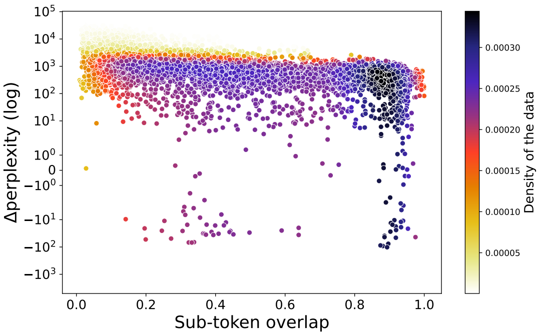

Figure 2: The correlation of subtoken overlap between HR and LR languages and Δperplexity (perplexity values are given in logarithmic scale). Darker colors show a greater density of points, where each point represents a HR-LR pair.

5.3 Interpretation using linguistic features

We also utilize linguistic features to interpret cross-lingual transfer results. To maintain the validity of our findings, we exclude features not annotated in at least half of the considered languages. Thus, we have only 21 out of 194 WALS features for analysis; 12 are specifically related to word order, and the rest to morphology. It is important to note that the absence of annotation in WALS may lead to possible gaps in our analysis. Thus, some crucial factors may be missed. We utilize these features for training the logistic regression model (see Appendix 3 for regression coefficients of 21 linguistic features we relied on).

Surprisingly, the genealogical family feature has a negative coefficient, suggesting that the model performs better when the donor and target languages are unrelated. At the same time, positive coefficients are observed for similar morphological features (e.g., *prefixing* vs. *suffixing*), indicating whether features like *Tense* or *Number* tend to be expressed with prefixes or suffixes. The word order typically does not play an important role, as only the order of *verb* and *object* appears to be significant.

5.4 Downstream evaluation results

5.4.1 POS tagging results

In this downstream evaluation, we investigate whether continued pretraining can help in POS tagging experiments. Here, for each LR language, we compare how well models trained on the best donors from the MLM setup perform against models trained in three randomly chosen languages. We want to determine if training on the best language yields better results for POS tagging than training on random ones. As described in 4.5.1, we train logistic regression using word embeddings to identify part-of-speech. If a word consists of multiple tokens, we use their average embedding. Additionally, we limit the training data to the number of tokens available in the least-resourced donor language for fair comparison. We evaluate performance using both Accuracy and F1-score metrics.

Table 2 shows the results for 6 LR languages with available data. When trained on the best HR donors, such as *Bambara, Bhojpuri, Guarani, Komi-Zyryan*, the model achieves the best performance in at least one metric. However, for *Cantonese* and *Coptic*, the best donors from MLM experiments do not result in the highest performance. In Figure 5, you can see the heatmap of POS tagging results for all considered HR languages in the case of the aforementioned 6 LR languages.

LR	HR	Setup	Accuracy	F1-score
Bambara	Arabic	random	0.266 ± 0.000	0.284 ± 0.0
	Armenian	random	0.344 ± 0.000	0.381 ± 0.000
	Dutch	random	0.159 ± 0.0	0.175 ± 0.0
	Lithuanian	best	**0.378** ± 0.004	**0.368** ± 0.004
Bhojpuri	Arabic	random	0.364 ± 0.0	0.428 ± 0.0
	German	random	0.5 ± 0.0	0.504 ± 0.0
	Persian	random	0.648 ± 0.000	0.627 ± 0.000
	Hindi	best	**0.705** ± 0.000	**0.722** ± 0.000
Cantonese	Italian	random	0.165 ± 0.000	0.175 ± 0.000
	Polish	random	**0.468** ± 0.000	**0.447** ± 0.000
	Russian	random	0.411 ± 0.000	0.388 ± 0.000
	Slovene	best	0.351 ± 0.000	0.373 ± 0.000
Coptic	Danish	random	0.052 ± 0.000	0.013 ± 0.000
	German	random	0.073 ± 0.000	0.092 ± 0.000
	Irish	random	**0.208** ± 0.000	**0.121** ± 0.000
	Afrikaans	best	0.146 ± 0.000	0.1 ± 0.000
Guarani	Faroese	random	0.208 ± 0.000	0.214 ± 0.000
	French	random	0.188 ± 0.000	**0.229** ± 0.000
	Irish	random	0.167 ± 0.000	0.167 ± 0.000
	Slovak	best	**0.25** ± 0.000	0.186 ± 0.000
Komi-Zyryan	Chinese	random	0.369 ± 0.000	0.334 ± 0.000
	Estonian	random	0.519 ± 0.000	0.42 ± 0.000
	Urdu	random	0.431 ± 0.000	0.456 ± 0.000
	Slovene	best	**0.581** ± 0.000	**0.536** ± 0.001

Table 2: The comparison shows evaluation results on POS tagging for 6 LR languages after pretraining on the most effective donors from MLM experiments, compared to 3 randomly selected HR languages. We observe that utilizing the best donors for transfer learning achieves better results in the POS tagging task compared to employing random HR languages.

5.4.2 Machine translation results

In contrast to the POS tagging, our data availability here is significantly limited. As mentioned in Section 4.5.2, we have a substantial amount of annotated data only for Afrikaans and English, total-

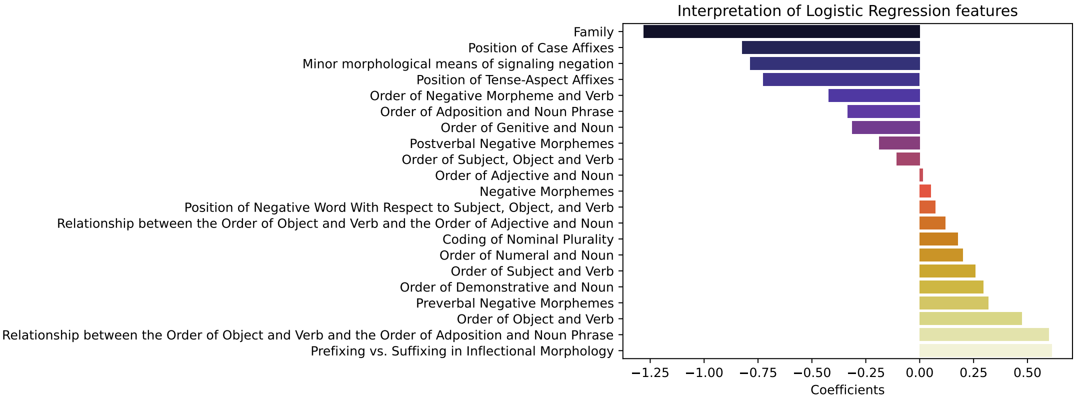

Figure 3: Coefficients of the WALS features obtained from the Logistic Regression model for the interpretation of cross-lingual transfer results.

ing 7 and 8 HR-LR pairs, respectively. Based on the general cross-lingual transfer results (Section 5.1), we observe that *Afrikaans* tends to perform better as a super-donor, while *English* does not show satisfactory results in that regard. Therefore, we decide to proceed with evaluation in Machine Translation setup using data of *Afrikaans* only. As supported by Figure 4, we limit our scope only to the top-performing HR donors.

In Figure 6, you can see the results of the MT setup where we perform fine-tuning on 7 *Afrikaans*-LR pairs. We report the comparisons using the $\Delta chrf$ metric, which indicates the difference in results with and without transfer learning on different HR languages. You can see that cross-lingual transfer significantly boosts MT performance in evaluating *Afrikaans-Sesotho*, *Afrikaans-Swati*, and *Afrikaans-Tsonga*. However, there is no improvement in the evaluation of *Afrikaans-Chichewa*. When looking at absolute values, cross-lingual transfer experiments demonstrated the most significant improvement in performance when applied to *Afrikaans-Akan* and *Afrikaans-Sesotho* pairs, with an increase of more than 0.2 in the *chrf*.

5.5 Super donors and super recipients

Surprisingly, as Figures 4,5,6 (but much more apparent from the heatmaps presented in the Appendix A.2) corresponding to the MLM, POS, and MT tasks respectively, show, a fraction of LR languages tend to be *super recipients*, benefiting unconditionally from all other HR languages. Additionally, some HR languages tend to be *super donors*, benefiting all languages unconditionally, regardless of the donor's or recipient's linguistic characteristics. Furthermore, the sets of such super donors and super recipients do not tend to generalize completely across tasks (MLM, POS, MT).

5.6 Further discussion

Our experiments demonstrate that transferring knowledge from HR languages during continued pretraining can enhance the performance of the mT5 model in the MLM setup across various LR languages. Subsequent downstream evaluation also exhibits such improvement, particularly in the case of the POS tagging and Machine Translation tasks.

The analysis of the MLM experiments reveals that most word order features have no significant impact. HR-LR language pairs from the same families correlate negatively with the results; however, this correlation shifts to positive when they share the same morphological feature as *affix*. Additionally, a higher degree of overlap between subtokens in languages tends to yield better performance in cross-lingual transfer. Meanwhile, other characteristics show no significant correlation with the results from MLM experiments. At the language level, *Afrikaans* and *Slovene* are determined as the best donor languages, and continued pretraining on them tends to better results across most LR languages examined in this study. It aligns with the findings of Turc et al. (2021), who identified a different pair of Germanic and Slavic languages, German and Russian, as the best donors. Most languages that exhibit promising results as donors belong to the Indo-European language family, specifically falling under the classification of Standard Average European (SAE) languages (Haspelmath, 2008). However, languages with the best results are peripheral members of the SAE continuum, i.e., have only some characteristics of SAE languages.

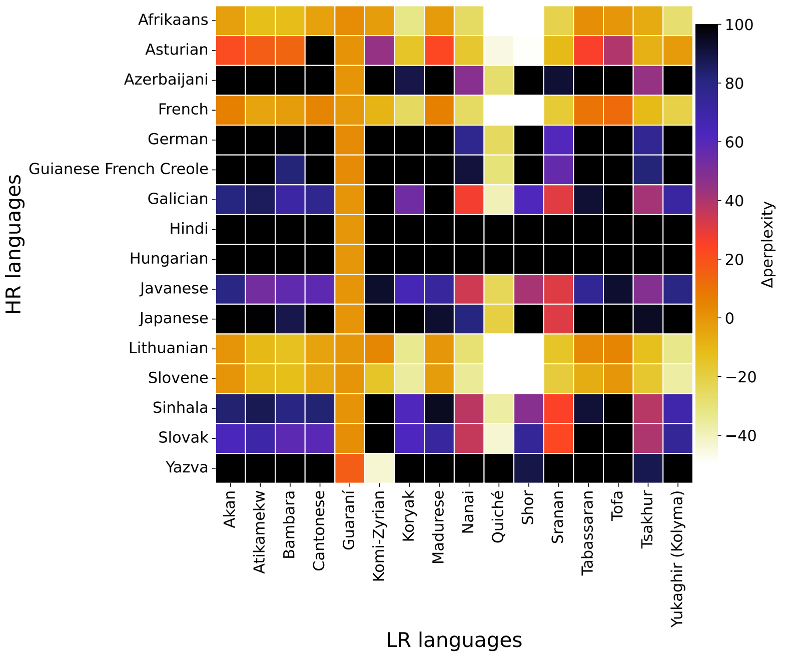

Figure 4: This heatmap illustrates the HR and LR languages where mT5 achieved lower perplexity scores than zero-shot performance. The colors represent the difference (Δperplexity) in perplexity after cross-lingual transfer by the continued pretraining on donors versus the zero-shot setup. Please refer to Appendix A.2 for detailed results of all LR and HR languages considered.

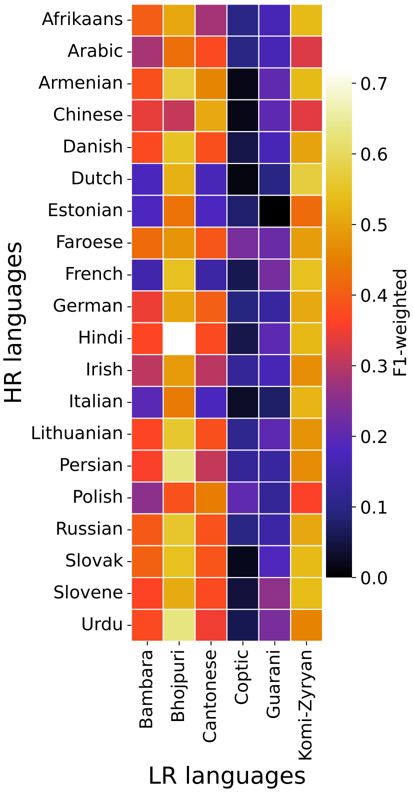

Figure 5: POS tagging results for 6 LR languages with available data in Universal Dependencies after training on HR donors from Table 2.

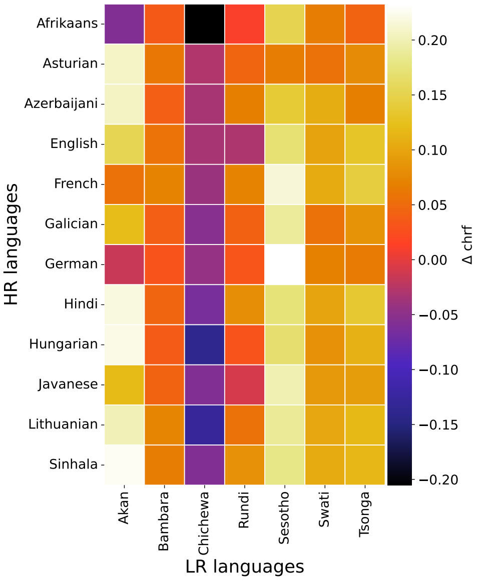

Figure 6: Machine Translation results for 7 *Afrikaans*-LR pairs. Here, we measure the difference between the model's performance with and without continued pretraining on different HR donors and report the results in the $\Delta chrf$ metric. It is important to note that fine-tuning and evaluation for MT were explicitly conducted on these 7 pairs.

In downstream evaluation, the findings from POS tagging demonstrate that employing optimal donor languages during pretraining outperforms pretraining on randomly selected languages in most cases. In Machine Translation, our investigation focuses on a particular pipeline for translation from *Afrikaans* to various LR. Consequently, we can conclude that pretraining on well-performing HR donors from the MLM step contributes to enhanced translation performance, particularly for extremely low-resource languages that we consider.

6 Conclusion

In this work, we present extensive experiments on cross-lingual transfer for HR-LR language pairs, leveraging HR languages for continued pretraining of the mT5. We observe that the performance of cross-lingual transfer significantly correlates with morphological features. Additionally, a higher degree of overlap between subtokens can contribute to better performance. Meanwhile, other characteristics do not significantly correlate with cross-lingual transfer results. We also observe improved downstream evaluation results, showing successful POS tagging and MT tasks performance. Finally, some LR languages tend to be *super recipients*, namely benefiting from all languages, and some HR languages tend to be *super donors* namely benefiting all languages with no apparent linguistic relation between donor and recipient.

References

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7701–7710, Online.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages.

Anna Belew. 2019. The endangered languages project (elp): Collaborative infrastructure and knowledge-sharing to support indigenous and endangered languages. In *Proceedings of the Language Technologies for All (LT4All)*.

Peter F. Brown, Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra, and Jennifer C. Lai. 1992. An estimate of an upper bound for the entropy of english. *Comput. Linguist.*, 18(1):31–40.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. 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.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Comput. Linguistics*, 47(2):255–308.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Blazej Dolicki and Gerasimos Spanakis. 2021. Analysing the impact of linguistic features on cross-lingual transfer. *CoRR*, abs/2105.05975.

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

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Inf. Process. Manag.*, 60(3):103250.

Yoshinari Fujinuma, Jordan L. 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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1500–1512. Association for Computational Linguistics.

Martin Haspelmath. 2008. 107. the european linguistic area: Standard average european. In *2. Halbband Language Typology and Language Universals 2. Teilband*, pages 1492–1510. De Gruyter Mouton.

David W. Hosmer and Stanley Lemeshow. 2000. Introduction to the logistic regression model. In *Applied Logistic Regression*.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Irina Krylova, Boris Orekhov, Ekaterina Stepanova, and Lyudmila Zaydelman. 2015. Languages of russia: Using social networks to collect texts. In *Russian summer school in information retrieval*, pages 179–185. Springer.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Bruce Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *CoRR*, abs/2004.01401.

Jindrich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *CoRR*, abs/1911.03310.

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.

Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. 2022. A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4903–4915. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15991–16111. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Benjamin Muller, Deepanshu Gupta, Jean-Philippe Fauconnier, Siddharth Patwardhan, David Vandyke, and Sachin Agarwal. 2022. Languages you know influence those you learn: Impact of language characteristics on multi-lingual text-to-text transfer. In *Transfer Learning for Natural Language Processing Workshop, 03 December 2022, New Orleans, Louisiana, USA*, volume 203 of *Proceedings of Machine Learning Research*, pages 88–102. PMLR.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Sebastian Ruder, Noah Constant, Jan A. Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10215–10245. Association for Computational Linguistics.

Hedvig Skirgård, Hannah J Haynie, Damián E Blasi, Harald Hammarström, Jeremy Collins, Jay J Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *CoRR*, 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, RepL4NLP@ACL 2020, Online, July 9, 2020*, pages 120–130. Association for Computational Linguistics.

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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Ze Yang, Wei Wu, Jian Yang, Can Xu, and Zhoujun Li. 2019. Low-resource response generation with template prior. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1886–1897, Hong Kong, China. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Appendix

A.1 The most efficient high-resource languages

Table 3 lists the most effective HR languages for cross-lingual transfer with the corresponding LR languages with decreased perplexity.

A.2 Cross-lingual transfer results

Figures 7 and 8 depict the heatmaps of various LR-HR language pairs, with corresponding colors indicating perplexity scores measured in the MLM setup. These scores represent the average values across 5 runs of continued pretraining on each HR language. Here, you can see 158 high-resource and 31 low-resource languages.

A.3 Statistics of the collected corpus

As we discussed in Section 3, we gathered the corpus of textual data for 189 languages. In order to divide language by high and low resource, we first calculated some statistics for each language. In Table 4, you can see official names, number of symbols and tokens for each language. We used the tokenizer from the mT5 model for text tokenization.

HR language	**LR languages with lowered perplexity**
Afrikaans	Akan, Atikamekw, Bambara, Cantonese, Komi-Zyrian Koryak, Madurese, Nanai, Quiché, Shor Sranan, Tofa, Tsakhur, Yukaghir (Kolyma)
Asturian	Koryak, Nanai, Quiché, Shor, Sranan Tsakhur, Yukaghir (Kolyma)
Azerbaijani	Quiché
French	Atikamekw, Bambara, Guaraní, Komi-Zyrian, Koryak Nanai, Quiché, Shor, Sranan, Tsakhur Yukaghir (Kolyma)
German	Quiché
Guianese French Creole	Quiché
Hindi	Guaraní
Hungarian	Guaraní
Japanese	Quiché
Javanese	Quiché
Lithuanian	Atikamekw, Bambara, Cantonese, Guaraní, Koryak Madurese, Nanai, Quiché, Shor, Sranan Tsakhur, Yukaghir (Kolyma)
Sinhala	Quiché
Slovak	Quiché
Slovene	Atikamekw, Bambara, Cantonese, Komi-Zyrian, Koryak Madurese, Nanai, Quiché, Shor, Sranan Tabassaran, Tofa, Tsakhur, Yukaghir (Kolyma)
Yazva	Komi-Zyrian

Table 3: High-resource languages that were used for training the mT5-Base, which achieved a lower perplexity metric than in zero-shot performance on low-resource languages.

Name	N_tokens, kk	N_symbols, kk
Abaza	1.33	2.35
Acehnese	1.47	3.09
Arabic (Egyptian)	157.47	291.76
Afrikaans	46.5	126.33
Akan	0.33	0.62
Albanian	0.002	0.005
Amharic	5.37	6.85
Arabic (Moroccan)	1.1	2.07
Arabic (Modern Standard)	1.83	1.83
Apurinã	0.002	0.0035
Archi	0.0012	0.0019
Arabic (Lebanese)	0.0015	0.0026
Armenian (Eastern)	0.12	0.26
Armenian (Western)	0.09	0.17
Armenian (Iranian)	212.94	569.39
Adyghe (Shapsugh)	3.32	5.58
Altai (Southern)	2.46	4.6
Assamese	11.18	18.87
Asturian	117.6	304.59
Atayal	0.71	1.35
Atikamekw	0.33	0.71
Avar	5.73	10.82
Awadhi	0.57	1.04
Aymara (Central)	0.92	1.81
Azerbaijani	90.5	230.74
Azari (Iranian)	39.73	74.9
Balinese	2.04	4.87
Bambara	0.17	0.31
Beja	0.003	0.003
Bengali	28.14	56.44
Bhojpuri	0.01	0.02
Bikol	3.39	8.63
Belorussian	168.71	467.04
Breton	21.79	43.06
Burmese	26.12	64.79
Bashkir	58.9	122.15
Bislama	0.13	0.28
Basque	12.73	35.51
Bugis	0.98	2.05
Bulgarian	147.44	367.2
Buriat	172.16	344.66
Choctaw	0.001	0.002
Cebuano	1365.37	3319.73
Chamorro	0.06	0.13
Chechen	378.15	477.3
Cherokee	0.17	0.22
Chukchi	4.08	5.12
Chuvash	277.48	442.77
Chichewa	0.28	0.75
Cantonese	0.02	0.02
Coptic	0.1	0.13
Crimean Tatar	1.24	2.6
Cornish	0.9	1.88
Catalan	463.18	1168.29
Chatino (Yaitepec)	1.21	3.05
Cheyenne	0.06	0.1
Czech	336.81	819.13
Dagbani	0.28	0.55
Dogri	0.006	0.009
Dhivehi	4.35	5.77
Dargwa	11.64	23.4
Danish	0.52	1.46
Dutch	598	1675.76
Dutch (Zeeuws)	1.43	3.12
English	7920.93	24002.62
Estonian	97.8	265.66
Even	0	0.01
Ewe	0.085	0.153
Faroese	4.17	9.4
Finnish	243.97	715.36
Frisian (North)	2.74	5.65
French	1541.64	4046.63
Frisian	0.007	0.016
Frisian (Western)	29.12	69.47
Fuzhou	1.95	2.86
Gaelic (Scots)	4.52	9.25
Gagauz	0.43	1.02
Georgian	65.7	149.49
German	6.94	21.3
Guianese French Creole	0.53	1.06

Name	N_tokens, kk	N_symbols, kk	Name	N_tokens, kk	N_symbols, kk	Name	N_tokens, kk	N_symbols, kk
Gilaki	1.33	2.38	Kinyarwanda	0.71	1.65	Samoan	0.31	0.61
Guajajara	0.0016	0.0023	Kurmanji	0.02	0.04	Sango	0.04	0.06
Galician	108.96	282.46	Karakalpak	0.71	1.55	Serbian-Croatian	375.92	828.15
Greek (Modern)	167.59	387.4	Kannada	40.49	94.87	Sindhi	9.28	15
German (Ripuarian)	1.01	2.21	Kongo	0.12	0.26	Seediq	1.14	2.28
Gorontalo	1.3	3.1	Komi-Permyak	1.82	3.19	Sesotho	0.22	0.49
Greenlandic (South)	0.15	0.36	Korean	254.45	318.2	Shan	5.09	7.59
German (Timisoara)	1761.41	5469.18	Kapampangan	1.85	4.41	Shona	1.32	3.11
Guaraní	0.03	0.02	Karachay-Balkar	4.38	9.14	Shor	0.18	0.31
Gujarati	19.35	34.67	Kurdish (Central)	16.77	29.46	Slovene	92.81	239
German (Viennese)	9.44	21.27	Karelian	0.01	0.02	Seminole	0	0
German (Zurich)	0.003	0.006	Koryak	0.25	0.43	Sinhala	18.78	37.48
Hakka	1.67	2.68	Khanty	0.0004	0.0005	Saami (Northern)	1.27	2.62
Hausa	5.58	13.62	Kumyk	1.16	2.45	Solon	1.49	2.93
Hawaiian	0.53	1.02	Komi-Zyrian	0.03	0.05	Somali	3.51	8.41
Haitian Creole	7.72	15.97	Lak	16.46	30.72	Sorbian (Upper)	3.71	7.64
Hebrew (Modern)	308.46	623.18	Lao	1.74	4.17	Spanish	1233.15	3408.23
Hindi	81.33	160.75	Latvian	54.26	135.41	Sranan	0.2	0.43
Hungarian	297.86	779.47	Luganda	1.47	3.38	Sardinian	3.39	7.95
Icelandic	23.58	55.33	Ladin	0.45	0.94	Sorbian (Lower)	0.83	1.69
Igbo	1.16	2.27	Lezgian	10.34	18.87	Santali	6.21	8.12
Ilocano	4.39	10.03	Low German	20.84	51.3	Sotho (Northern)	0.84	1.86
Indonesian	0.28	0.89	Lingala	0.53	1.06	Sundanese	12.54	31.15
Ingush	8.09	14.27	Lithuanian	78.83	199.57	Slovincian	1.2	2.14
Indonesian (Jakarta)	209.58	612.28	Liv	0.004	0.009	Slovak	91.75	224.02
Irish	0.27	0.6	Ladino	1.31	3.25	Swahili	14.32	35.39
Irish (Munster)	15.82	34.2	Luxemburgeois	17.24	41.58	Swedish	0.36	0.99
Italian	1018.44	2776.36	Mari (Hill)	1.71	2.91	Swati	0.14	0.34
Itelmen	0.0005	0.0008	Maithili	2.86	5.27	Swedish (Västerbotten)	882.57	2204.72
Italian (Genoa)	2.57	4.9	Maori	1.2	2.25	Tagalog	21.96	54.45
Italian (Napolitanian)	1.99	3.9	Macedonian	83.34	211.18	Tahitian	0.11	0.19
Italian (Turinese)	12.27	23.48	Madurese	0.2	0.42	Tajik	20.91	43.76
Javanese	17.44	43.98	Meithei	0.47	0.94	Tashlhiyt	0.53	0.89
Jamaican (Creole)	0.42	0.9	Mingrelian	4.43	8.19	Tabassaran	0.06	0.11
Japanese	750.08	1244.1	Marathi	17.82	36.2	Telugu	79.39	178.59
Kabardian	18.8	29.62	Minangkabau	34.7	86.25	Thai	79.96	226.41
Kashmiri	0.13	0.22	Mongol (Khamnigan)	13.48	30.69	Tigrinya	0.13	0.14
Kazakh	72.64	184.75	Malgwa	21.64	48.87	Turkmen	4.05	8.47
Kabyle	1.58	2.83	Maltese	5.37	11.79	Tamil	0.03	0.08
Kabiyé	1.84	2.39	Malay	83.7	241.85	Tibetan (Modern Literary)	34.98	41.26
Galician	108.96	282.46	Mari (Meadow)	189.06	275.65	Tat (Muslim)	0.06	0.11
Greek (Modern)	167.59	387.4	Mordvin (Moksha)	1.45	2.24	Tongan	0.36	0.65
German (Ripuarian)	1.01	2.21	Mandarin	497.71	659.72	Tofa	0.03	0.06
Gorontalo	1.3	3.1	Mansi	2.58	3.99	Tok Pisin	0.16	0.34
Greenlandic (South)	0.15	0.36	Manx	1.57	3.07	Tsakhur	0.09	0.15
German (Timisoara)	1761.41	5469.18	Mordvin (Erzya)	7.81	15.25	Tsonga	0.25	0.58
Guaraní	0.03	0.02	Mon	4.98	7.33	Tamil (Spoken)	65.88	188.99
Gujarati	19.35	34.67	Marshallese	0.002	0.003	Tswana	0.5	1.15
German (Viennese)	9.44	21.27	Mundurukú	0.002	0.002	Tetun	0.42	1.01
German (Zurich)	0.003	0.006	Malayalam	45.49	118.33	Tulu	1.14	2.15
Hakka	1.67	2.68	Mazanderani	2.73	5.08	Tupi	0.006	0.008
Hausa	5.58	13.62	Nanai	0.24	0.41	Turkish	169.37	468.8
Hawaiian	0.53	1.02	Nauruan	0.13	0.26	Tuvan	24.8	51.9
Haitian Creole	7.72	15.97	Navajo	5.67	8.52	Tatar	59.84	127.89
Hebrew (Modern)	308.46	623.18	Ndonga	0.003	0.008	Udi	0.1	0.16
Hindi	81.33	160.75	Nadroga	0.2	0.43	Udmurt	4.76	9.24
Hungarian	297.86	779.47	Nepali	13.5	27.54	Ukrainian	619.13	1523.74
Icelandic	23.58	55.33	Nias	0.36	0.77	Urdu	57.23	110.15
Igbo	1.16	2.27	Norwegian	224.43	608.76	Urubú-Kaapor	0.001	0.001
Ilocano	4.39	10.03	Narom	0.98	1.96	Uyghur	12.66	18.24
Indonesian	0.28	0.89	Neo-Aramaic (Assyrian)	0.0026	0.0021	Uzbek	45.15	104.75
Ingush	8.09	14.27	Nenets (Tundra)	0.47	0.78	Venda	0.11	0.25
Indonesian (Jakarta)	209.58	612.28	Nivkh (South Sakhalin)	0.73	1.14	Veps	2.96	6.83
Irish	0.27	0.6	Newar (Dolakha)	24.93	45.29	Vietnamese	424.35	742.98
Irish (Munster)	15.82	34.2	Oirat	7.97	14.12	Welsh	40.94	82.34
Italian	1018.44	2776.36	Ossetic	2.96	4.91	Wolof	1.3	2.54
Itelmen	0.0005	0.0008	Oriya	16.53	18.51	Warlpiri	0	0
Italian (Genoa)	2.57	4.9	Panjabi	27.68	42.51	Wu	6.67	9.11
Italian (Napolitanian)	1.99	3.9	Papiamentu	11.04	19.5	Waray-Waray	187.45	446.09
Italian (Turinese)	12.27	23.48	Pangasinan	0.63	1.29	Xhosa	0.61	1.56
Javanese	17.44	43.98	Polish	629.24	1616.18	Yi	0.001	0.001
Jamaican (Creole)	0.42	0.9	Portuguese	600.72	1550.9	Yukaghir (Kolyma)	0.026	0.044
Japanese	750.08	1244.1	Provençal	32.42	76.25	Yakut	16.84	33.18
Kabardian	18.8	29.62	Persian	298.12	601.61	Yiddish (Lithuanian)	7.43	14.58
Kashmiri	0.13	0.22	Qafar	0	0.0001	Yoruba	4.69	8.03
Kazakh	72.64	184.75	Quiché	0.02	0.04	Yup'ik (Central)	0.004	0.009
Kabyle	1.58	2.83	Romani (Lovari)	0.2	0.49	Yurt Tatar	2.28	4.04
Kabiyé	1.84	2.39	Rundi	0.14	0.31	Yukaghir (Tundra)	0	0.01
Kirghiz	27.32	65.52	Romanian	195.88	485.41	Yazva	14.23	25.9
Khakas	1.44	2.89	Romansch (Sursilvan)	5.07	12.59	Zazaki	5.46	10.8
Khmer	10.98	28.84	Russian	2372.32	6397.2	Zhuang (Northern)	0.28	0.52
Kikuyu	0.18	0.34	Rutul	0.36	0.67	Zulu	1	2.34

Table 4: All languages presented in the data we collected and used for the experiments.This table includes all collected languages with number of symbols and tokens (tokenized by mT5-Base tokenizer), where "kk" - millions.

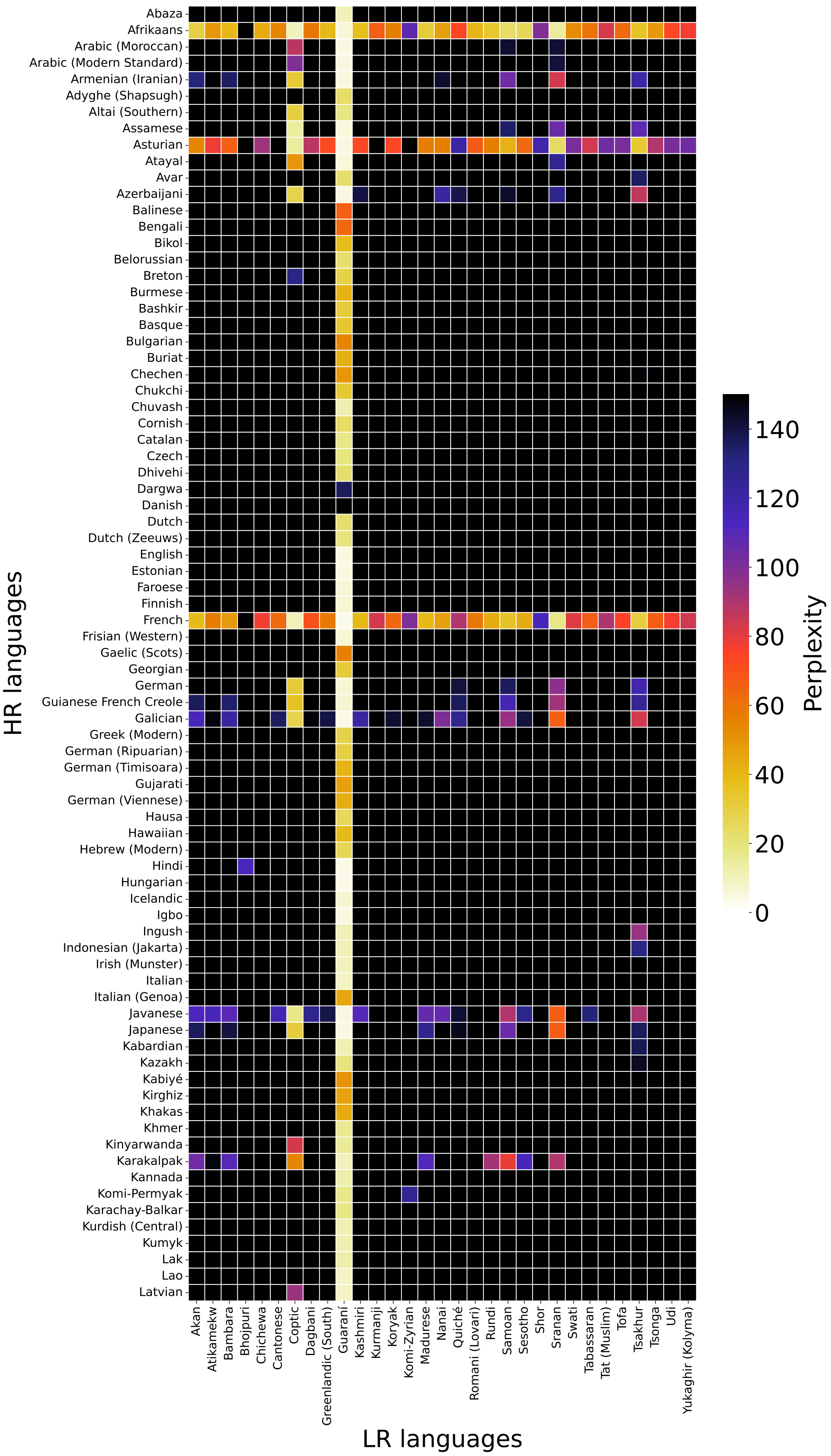

Figure 7: Heatmap for the first 79 high-resource languages with absolute perplexity scores of 31 low-resource languages.

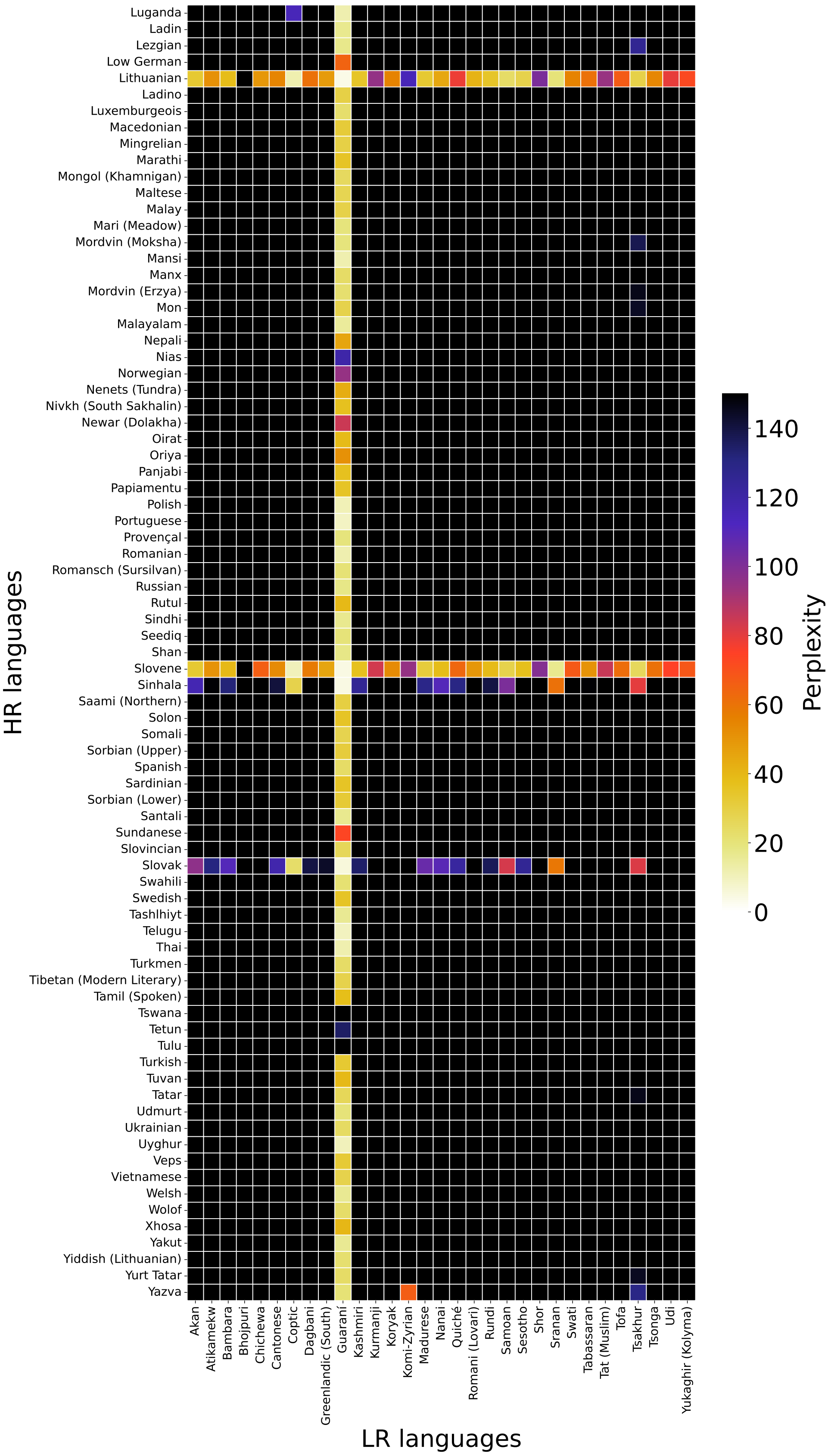

Figure 8: Heatmap for the second 79 high-resource languages with absolute perplexity scores of 31 low-resource languages.