Subword Evenness (SuE) as a Predictor of Cross-lingual Transfer to Low-resource Languages

Olga Pelloni[1] **Anastassia Shaitarova**[2] **Tanja Samardzic**[1]

[1]Text Group, URPP Language and Space,

[2]Department of Computational Linguistics, University of Zurich, Switzerland

olga.pelloni@gmail.com, {anastassia.shaitarova, tanja.samardzic}@uzh.ch

Abstract

Pre-trained multilingual models, such as mBERT, XLM-R and mT5, are used to improve the performance on various tasks in low-resource languages via cross-lingual transfer. In this framework, English is usually seen as the most natural choice for a transfer language (for fine-tuning or continued training of a multilingual pre-trained model), but it has been revealed recently that this is often not the best choice. The success of cross-lingual transfer seems to depend on some properties of languages, which are currently hard to explain. Successful transfer often happens between unrelated languages and it often cannot be explained by data-dependent factors. In this study, we show that languages written in non-Latin and non-alphabetic scripts (mostly Asian languages) are the best choices for improving performance on the task of Masked Language Modelling (MLM) in a diverse set of 30 low-resource languages and that the success of the transfer is well predicted by our novel measure of *Subword Evenness (SuE)*. Transferring language models over the languages that score low on our measure results in the lowest average perplexity over target low-resource languages. Our correlation coefficients obtained with three different pre-trained multilingual models are consistently higher than all the other predictors, including text-based measures (type-token ratio, entropy) and linguistically motivated choice (genealogical and typological proximity).

1 Introduction

Since pre-trained multilingual models became available, the most common approach to NLP tasks on low-resource languages has been cross-lingual transfer learning. After initial tests on cross-lingual ability within the languages covered by the multilingual models (Pires et al., 2019; Wu and Dredze, 2020; Libovický et al., 2020), a more recent line of research shows that new languages (not seen in the training set) are best processed by continued training or fine-tuning (depending on the task) of a pre-trained multilingual model on *one transfer language*, even if the transfer language is different from the target language. As a matter of fact, a single transfer language can help improve the performance on *many target languages* (Lin et al., 2019; Lauscher et al., 2020; Turc et al., 2021). This insight is especially important for low-resource languages, for which only test data is typically available. Multilingual models are usually fine-tuned or (additionally) trained on a high-resource language and the resulting weights are applied as zero- or few-shot transfer to a target low-resource language (Tunstall et al., 2022).

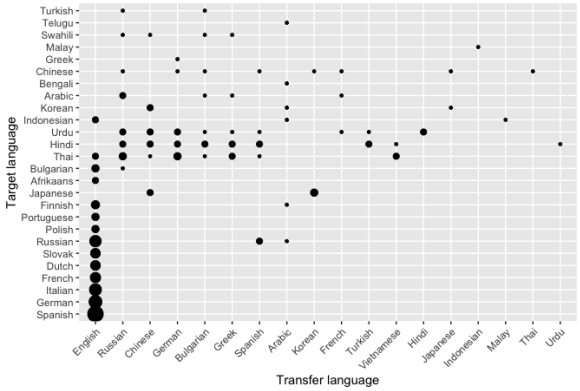


Figure 1: Best cross-lingual transfer results mentioned in previous work. The size of each circle is proportional to the number of mentions of a given transfer-target pair. English is the best transfer language for European target languages. For target languages outside of Europe, the best transfer languages are hard to predict.

It seems that injecting a bias towards some particular linguistic traits, while slightly "forgetting" some others, makes the multilingual models adapt better to a new language. An evident but hard question arises here: Which languages are especially well-suited to serve as transfer languages for many low-resource target languages? This is the first

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7428–7445

research question that we address in our study.

Most research on cross-lingual transfer regards English as the most natural choice for any target language due to the abundance of training data. However, more recent studies reveal that other languages are often better choices. For instance, Russian turns out to be a good transfer language for Thai, Arabic, Swahili, and Chinese (Turc et al., 2021). It is hard to see why this happens as these languages are neither related nor similar by any of the known criteria. Our own review of several studies on cross-lingual transfer, summarized in Figure 1, shows that English is the best transfer language when European languages are the target. For other target languages, the choice of the transfer languages is anything but clear. English is rarely the best choice, while Russian, Chinese, German, Greek, or Arabic seem to help in many cases.

This observation leads us to the second question that we address in our study: Which linguistic traits should be used to improve the performance of pre-trained multilingual models? In other words, what are the traits that make some languages more suitable for transfer?

To answer these questions, we test language models on a highly diverse set of low-resource languages, whose geographical distribution is shown in Figure 2. We follow the current cross-lingual transfer learning framework in order to train the models. We consider three popular multilingual pre-trained models and 19 high-resource languages as potential transfer languages. In these experiments, we look for the best transfer languages, but also, more importantly, for a strong predictor of the transfer results across all target low-resource languages.

As a predictor, we propose a novel text-based measure, *Subword Evenness (SuE)*, a parameter that describes the differences in the length of subword units (one value per language). The motivation for looking into the properties of subword units to assess the suitability of a given language to be a transfer language comes from the fact that language models are trained over the output of subword tokenization. We know that subword splits can depend on the properties of a language. For example, Finnish is known for its regular subword structure (morphology), which is expected to give more even subword tokens, while other languages might have less subword regularity, leading to more uneven splits. Our measure shows the preference

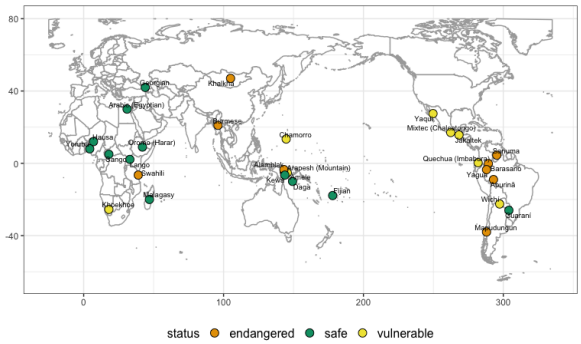


Figure 2: Distribution of 30 test languages in our study: highly diverse set in terms of linguistic families (23), geographic areas (5) and endangerment status (3).

for even splits in a given language. An example of a relatively even split would be *co-work-ing* → set of lengths [2, 4, 3] compared to an uneven split *co-working* → set of lengths [2, 7]. A detailed explanation of the measure is in Section 3.

The main finding of our work is that the property of Subword Evenness matters for successful transfer to diverse non-Indo-European languages. Transfer languages that score low on our measure give the lowest average perplexity over target low-resource languages. Compared to other text-based statistics and alternative linguistically motivated choices, our measure is the best predictor of transfer results.

The code and data used for this work are publicly available in the repository https://github.com/olgapelloni/subword_evenness.

2 Related Work

The main idea behind cross-lingual transfer learning is that both similarities and differences between languages can be exploited for improving model performances. In early work on multilingual models, similarities between languages were exploited to increase training data, for instance in multilingual syntactic parsing (Zeman and Resnik, 2008; Snyder et al., 2009) and pivot-based statistical machine translation (Paul et al., 2013). Cross-linguistic differences, as a form of distant supervision, have been shown to help the disambiguation of lexical meaning (van der Plas and Tiedemann, 2006) and part-of-speech (POS) tags (Snyder et al., 2008). Neural models made this idea even more attractive, leading to many proposals on how to share some parameters across languages, for instance through cascading and multi-task learning (p. 218–227, Goldberg, 2017; Ruder, 2017). With

Transformer-based models (Vaswani et al., 2017), the transfer approach has proven to be widely applicable and successful, as shown by the 2018 model ULMFiT (Howard and Ruder, 2018).

Previous analyses of multilingual Transformer-based models have shown that the representations produced during training are multilingual but the language-specific information is preserved (Pires et al., 2019; Wu and Dredze, 2019; Libovický et al., 2020; Wang et al., 2020b). Forgetting some of the language-specific information has been helpful for the task of question-answering (Yang et al., 2021). Wu and Dredze (2019) show a positive correlation between the number of shared tokens and transferability between languages, but this finding has not been confirmed in later studies. K et al. (2020) do not find a strong influence of lexical overlap on successful language transfer. Pires et al. (2019) add that the shared tokens help to create cross-lingual representations, but language transfer success depends more on the structural similarity between languages.

Studies on low-resource languages demonstrate how a lack of resources impacts performance (Wu and Dredze, 2020; Wang et al., 2020a; Goyal et al., 2021). Ruder et al. (2021) show that the performance of XLM-R is lower on low-resource languages than on high-resource ones. Moreover, performance is lower on languages with non-Latin scripts, such as Hebrew, Japanese, Thai or Chinese.

The current processing workflows often include continued training of a multilingual model on a single high-resource transfer language and then applying the resulting weights in a few- or zero-shot manner to many low-resource languages. Focusing on the question of which transfer languages give good results over multiple target languages, we count the mentions of the best transfer pairs in several previous studies (Ruder et al., 2021; Turc et al., 2021; Vázquez et al., 2021; Hu et al., 2020; Lauscher et al., 2020; Lin et al., 2019; Paul et al., 2013). The counts are plotted in Figure 1, showing an interesting asymmetry between European and other languages. When European languages are the target of transfer, English seems to be the best transfer language. This is in line with the findings on the XTREME benchmark for evaluating cross-lingual transfer (Hu et al., 2020), which led the authors to conclude that English is the most common and the most universal choice for a transfer language. Paul et al. (2013) found that English as a pivot language in statistical machine translation works well in approximately half of the observed language pairs (22 Indo-European and Asian languages). Our review shows that English is not the best choice when non-European languages are the target.

One approach to choosing a transfer language is to rely on structural similarity, which is measured using grammar features from the URIEL database (Littell et al., 2017). Lauscher et al. (2020) find that transfer is better in language pairs that are closer in the URIEL vector space regarding POS tagging and syntactic dependency parsing. Other factors, such as data size, are better predictors on the tasks of question answering and inference. de Vries et al. (2022) find that surface string similarity is the best predictor for POS tagging. Shaitarova and Rinaldi (2021) develop their own linguistic typology based on negation constructions, which helps to choose a better transfer language for negation scope resolution. Lin et al. (2019) suggest aggregating various types of linguistic features, including geographic location. Aggregated measures select better transfer candidates than single features.

We depart from previous research in terms of both data and methods. We work with a much bigger and more diverse sample of languages than any previous studies. The main methodological novelty of our work is the proposed text-based parameter, which captures an interesting subword feature of good transfer languages.

3 Subword Evenness (SuE) as a Language Parameter

Defining formal properties of languages relevant to NLP is still a challenging task. Mielke et al. (2019), for instance, notice that language model sentence surprisal scores (aggregated at the level of a language) differ across languages, but do not manage to identify any properties of a language that would predict the surprisal. When it comes to transfer learning, previous research shows little agreement on which languages should be chosen for cross-lingual transfer and why.

The full workflow for calculating the SuE score (one value per language) is shown in Figure 3. We first apply a subword tokenization algorithm to split words in an unsupervised fashion. We then convert each segmented word $W = w_1, w_2, \ldots w_n$ in the data into a sequence of integers $L = l_1, l_2, \ldots l_n$, where n is the number of subword segments in

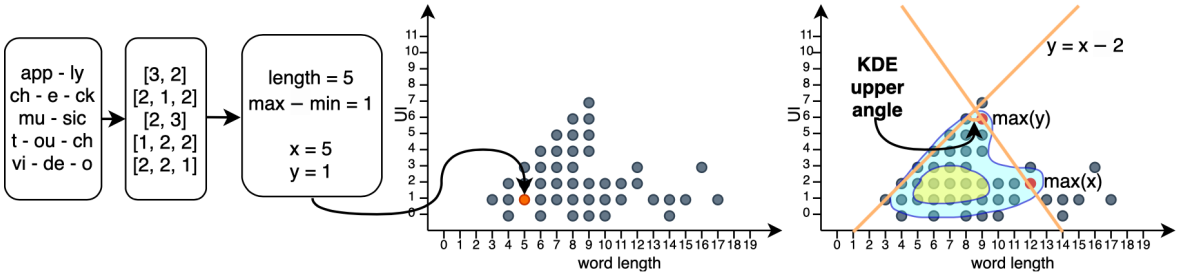


Figure 3: Extracting our measure of Subword Evenness (SuE).

each word and each integer l_i represents the length of the subword segment w_i measured in Unicode characters. We then map each sequence of integers to a single integer $L \rightarrow UI$ (**u**nevenness **i**ndex), where $UI = max(L) - min(L)$ is the difference between the maximum length (longest segment) and the minimum length (shortest segment) in the sequence. The values of UI are plotted against the values of word length measured in Unicode characters (the x-axis in Figure 3), resulting in a distribution, which then undergoes a Kernel Density Estimation (KDE)[1] analysis for identifying the shape of the density region.

We aim to describe the whole distribution with a single parameter by approximating the shape of the density area with two lines and measuring the angle between them. The first line (intercepting the x-axis on the left side in Figure 3) is a natural bound of the distribution $f(x) = x - 2$ determined by the fact that the difference between any two subword segments cannot be greater than the length of the whole word. The second line is a linear function fitted to two points at the edge of the density area: $max(x)$ and $max(y)$, $g(x) = kx + b$. Finally, the measure of *Subword Evenness* (Equation 1) is the upper angle in the triangle formed by these two lines and the x-axis.

$$SuE = 180° - |\arctan 1| - |\arctan k| \quad (1)$$

For Equation 1, we first calculate the inclination of each intersecting line using the formula $m = tan(\theta)$, where m is the slope of the line and θ is the inclination, i.e. the angle between the x-axis and the line (Larson and Hostetler, 2007, p. 430). The slope of the first line $f(x)$ equals 1, the slope of the second line $g(x)$ equals k. In order to get the value of the inclination θ from $tan(\theta)$, we apply the inverse trigonometric function $\arctan(\tan(\theta)) = \theta$ (Larson and Hostetler, 2007, p. 109, 193). We use absolute values in order to work with the angles measured in degrees. Once we know the inclination of both lines (their angles to the x-axis), we can find the upper angle inside the triangle. Since all angles in a triangle sum up to 180°, we subtract both inclinations from 180° to find the upper angle[2]. The upper angle shows the preference for *Subword Evenness*, thus called SuE. The higher the value of this angle, the higher the preference towards even subword splits in longer words.

There are many methods that can be used to perform unsupervised subword tokenization. The most widely used methods are Byte-Pair Encoding (BPE, Gage, 1994) implemented by Sennrich et al. (2016) and also in the SentencePiece library (Kudo and Richardson, 2018), WordPiece (Wu et al., 2016) and the SentencePiece Unigram model. The output of all these algorithms is highly dependent on the hyperparameters that determine the size of the resulting subword vocabulary. However, there are currently no general criteria to determine the value of these hyperparameters (Mielke et al., 2021).

We follow the method of Gutierrez-Vasques et al. (2021), which is motivated by information theoretic properties of segmented text in many languages: the BPE algorithm is run until the text redundancy value reaches its minimum. This criterion is independent of any particular NLP task and is cross-lingually aligned, which is important for our study. Cross-lingual alignment allows us to normalize, to a certain degree, the difference in writing systems: minimum redundancy is reached faster in alphabetic scripts and slower in syllabic and logographic scripts. For reference, we extract the SuE measure using several other segmentation algorithms and hyperparameters. These results can be found in Table 9 in the Appendix.

[1] KDE is a common method in spatial data science, described in more detail by Ripley (2002).

[2] The result of the arctan function is measured in radians, which we convert into degrees.

4 Data and Experiments

We obtain our training and test data from the Text Data Diversity Sample (TeDDi, Moran et al., 2022). The TeDDi corpus contains texts representing the languages included in the 100-language sample, published by the World Atlas of Language Structures (WALS, Haspelmath et al., 2005). This sample is compiled by experts in linguistic typology with the goal of representing overall linguistic diversity: language families, geographical areas and typological features. Following this sample, the TeDDi corpus maximizes linguistic diversity and also covers different textual genres for languages with more available resources (Gutenberg Project[3], the Parallel Bible Corpus (Mayer and Cysouw, 2014), the OpenSubtitles corpus (Lison and Tiedemann, 2016) and Universal Declaration of Human Rights[4]

For the set of transfer languages (training set), we select 19 languages which contain at least one million tokens: Basque (eus), English (eng), Finnish (fin), French (fra), German (deu), Greek (ell), Hebrew (heb), Hindi (hin), Indonesian (ind), Japanese (jpn), Korean (kor), Mandarin (cmn), Persian (pes), Russian (rus), Spanish (spa), Tagalog (tgl), Thai (tha), Turkish (tur), Vietnamese (vie). We balance the genres when possible and cap the length of the sampled text to 1M tokens per language. We use the scikit-learn library (Pedregosa et al., 2011) to shuffle each dataset and split it into train (80%) and validation (20%) sets.

For the set of target languages (test set), we define the threshold of at least 100K tokens available per language. There are 30 such languages: Alamblak, Amele, Apurina, Arabic (Egyptian), Arapesh (Mountain), Barasano, Burmese, Chamorro, Daga, Fijian, Georgian, Guarani, Hausa, Jakaltek, Kewa, Khalkha, Khoekhoe, Lango, Malagasy, Mapudungun, Mixtec (Chalcatongo), Oromo (Harar), Quechua (Imbabura), Sango, Sanuma, Swahili, Wichi, Yagua, Yaqui, Yoruba. We fix the size of the test sets to be 100K tokens per language. The chosen languages belong to 23 different language families and are spoken in 5 geographical areas (Figure 2).

4.1 Pre-trained Models

We work with three popular multilingual Transformer models in our experiments: mBERT (Devlin et al., 2018), XLM-RoBERTa (Conneau et al., 2019) and mT5 (Xue et al., 2021; Raffel et al., 2019). The pre-training objectives for all three models include the Masked Language Modeling (MLM) objective which is also the objective of our trained models. These models were pre-trained on vast amounts of multilingual data such as CommonCrawl for XLM-R and mT5 and Wikipedia for mBERT. We employ the base variants of the models (Table 1).

Model	Param	Segment	Lang	Data
mBERT	110M	WordPiece	104	Wikipedia
XLM-R	270M	SentPiece	100	CCNet
mT5	580M	SentPiece	101	mC4

Table 1: Models' specifications. All models are base models of the respective model families.

mBERT PPL: 42.90		**XLM-R** PPL: 98.08		**mT5** PPL: 3.02	
Lang	**Gain**	**Lang**	**Gain**	**Lang**	**Gain**
jpn	-14.13	kor	-57.90	rus	-0.13
ell	-12.14	cmn	-55.86	ell	-0.10
heb	-11.82	heb	-55.73	tha	-0.08
tha	-11.79	jpn	-55.58	pes	-0.08
rus	-11.75	ell	-54.28	jpn	-0.07
eng	-10.20	eng	-53.18	eng	-0.06

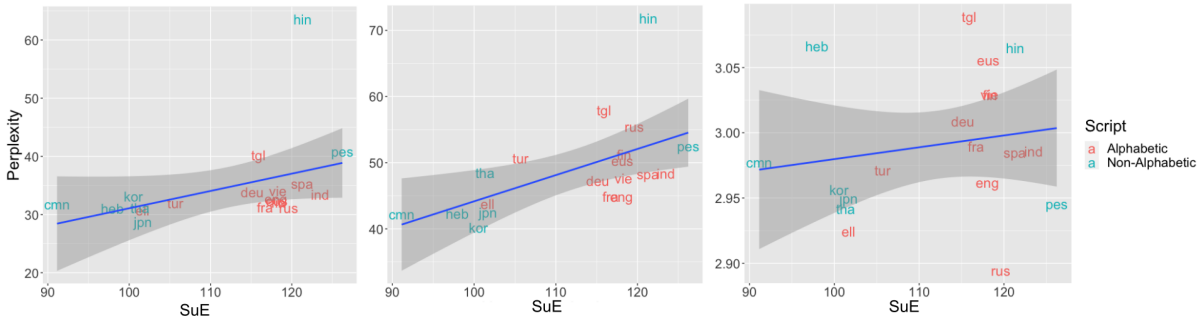
Table 2: Top 5 transfer languages and English (for reference) with their respective gains (reduction in average target language model perplexity) compared to the average baseline perplexity (PPL). The bigger the reduction, the better. The colours highlight re-occurring languages across the models.

Note that we use the pre-trained multilingual models only as the starting point for our own monolingual training. This is currently the best method for obtaining high performance on many test languages.

4.2 Continued Training and Testing

We continue training the three pre-trained models on our set of transfer languages. We keep the same hyperparameters for mBERT and XLM-R and train them for 5 epochs with a learning rate 3e-5 and maximum sequence length 256. In order to maintain the batch size of 128 with infrastructural constraints, we set the batch size to 4 and use gradient accumulation after 32 steps. For mT5, we use an available T5 architecture from Huggingface (Wolf et al., 2019) which allows continued

[3] https://www.gutenberg.org/
[4] http://unicode.org/udhr/.



(a) mBERT (b) XLM-R (c) mT5

Figure 4: Correlation (Pearson) of our measure SuE and the average perplexity across all test languages in (a) mBERT (0.4), (b) XLM-R (0.56) and (c) mT5 (0.18). Transfer languages closer to the bottom stand for the best transfer results.

Lang	SuE	Lang	SuE	Lang	SuE
cmn	**91.14**	deu	115.14	fin	118.35
heb	**97.94**	tgl	115.89	rus	119.61
kor	**100.54**	fra	116.70	spa	121.27
tha	**101.33**	eng	118.04	hin	121.33
ell	**101.66**	vie	118.29	ind	123.49
jpn	101.69	eus	118.15	pes	126.21
tur	105.68				

Table 3: Transfer languages with their respective SuE values in ascending order. Top 5 (lowest SuE) are highlighted.

training of the language model in an unsupervised fashion. We load the mT5-base checkpoint and continue training following the early stopping approach as in Turc et al. (2021), where we evaluate each model every 200 steps and stop training at 2000 steps. Only the best identified checkpoint is kept as a final model. As a result, mT5 is trained for approximately the same number of epochs (3–5) as mBERT and XLM-R, making the results comparable.

All models are re-trained 5 times with 5 different seeds. Thus, we obtain 285 fine-tuned models (5 models $\times$ 19 training languages $\times$ 3 architectures) plus 3 pre-trained models without continued training as baselines. We test the baseline and additional trained models on all 30 test languages in a zero-shot fashion. We evaluate all the models measuring the perplexity, which is an exponentiated average of negative log likelihood of a sequence. Lower perplexity means better performance with the chosen transfer language. We report the average results across 5 seeds for each language pair.

Only 4 out of 30 target languages[5] in our datasets are included in the pre-training of all three pre-trained models: Burmese, Georgian, Malagasy, and Swahili, making our experiments almost completely zero-shot.

5 Results and Discussion

The first general outcome of our experiments is the ranking of transfer languages according to the average perplexity score on the target languages shown in Table 2. There is considerable overlap between the three sets of top 5 languages (one set per pre-trained model).[6] In addition, we note that English is never in the top 5, which confirms observations from previous work summarized in Figure 1. An interesting observation regarding these results concerns the scripts: all of the best transfer languages are written in a non-Latin script, in contrast to the general preference for Latin scripts (e.g. over 70% of languages in the TeDDi sample are written in a Latin script).

The second general outcome is the ranking of transfer languages according to the SuE measure shown in Table 3. There is considerable overlap between the top languages in the two rankings (comparing Table 3 with Table 2). To quantify this dependence, we perform correlation tests.

[5]Two language varieties are not specified exactly in the models' documentation, such as Arabic Egyptian and Khalkha. If we map them to Arabic and Mongolian in the models' documentation, then we get 6 languages covered.

[6]The full table of the scores for all the transfer languages can be found in Appendix B.

5.1 Transfer language SuE correlates with average target language perplexity

Figure 4 shows the main finding of our study: a relatively high correlation score between the Subword Evenness (SuE) measure and the transfer-learning performance per transfer language, measured as language model perplexity averaged over all 30 target languages. This correlation means that transferring from languages with uneven subword splits (low SuE) leads to better performance (lower perplexity) on a diverse set of target languages. In other words, adding information from languages whose words are split unevenly to a pre-trained multilingual model is more helpful for the performance on many target languages than adding information from more regular languages whose subwords are evenly split.

While we find this correlation with all three pre-trained models, we notice that the coefficients differ. For instance, the Pearson correlation is 0.56 in the case of XLM-R, 0.40 for mBERT and only 0.18 for mT5. These differences are well aligned with the baseline performance and transfer gains. For example, the baseline perplexity of mT5 is much lower than the other two models, the gains are consequently smaller and the correlation is weaker. However, with only three observations, we cannot know whether this alignment is due to chance.

Another pattern that appears in the plots is the grouping of languages according to the script, with non-alphabetic scripts being more associated with low SuE values. This might point to other potential explanations for the observed correlation such as, for instance, the fact that words tend to be shorter in non-alphabetic scripts leading to smaller angles. To exclude such factors, we perform tests with several other measures and compare them with SuE.

	Model	P	S	All P	All S
Subword Evenness	mBERT	**0.40** $p = .09$	**0.62** $p = .005$	**0.5** $p = .03$	**0.77** $p = .0002$
	XLM-R	**0.56** $p = .01$	**0.69** $p = .002$		
	mT5	**0.18** $p = .47$	0.11		
Avg word length	mBERT	-0.01	0.06	0.13	0.39
	XLM-R	0.27	0.40		
	mT5	**0.18** $p = .47$	**0.19** $p = .43$		
Type-token ratio	mBERT	-0.21	-0.23	-0.17	0.04
	XLM-R	-0.12	0.11		
	mT5	-0.11	-0.16		
Unigram entropy	mBERT	-0.15	-0.14	-0.19	-0.08
	XLM-R	-0.21	-0.13		
	mT5	-0.07	-0.01		

Table 4: SuE vs. other text-based measures: Pearson (P) and Spearman (S) correlation coefficients. P-values are given for the highest correlation coefficients.

Target	Transfer	Avg PPL Change	Lang Family (WALS)
Arabic	Hebrew	+0.04	Afro-Asiatic
Burmese	Mandarin	+0.11	Sino-Tibetan
Chamorro	Indonesian	+7.15	Austronesian
	Tagalog	+16.45	
Fijian	Indonesian	+3.14	Austronesian
	Tagalog	+5.72	
Hausa	Hebrew	+1.2	Afro-Asiatic
Khalkha	Turkish	+9.23	Altaic
Malagasy	Indonesian	+0.17	Austronesian
	Tagalog	+0.63	
Oromo	Hebrew	+0.39	Afro-Asiatic

Table 5: Average change in perplexity (across 3 models), when using a genealogically close language instead of a language with low SuE (best PPL option among top 5 is chosen). Higher numbers mean worse performance.

5.2 SuE vs. other text-based measures

Table 4 shows the correlation coefficients obtained with SuE compared to other, simpler text-based measures. We include the mean word length in this analysis to see whether the known impact of the script on the word length may explain the observed patterns in cross-lingual transfer. As for type-token ratio (TTR) and unigram entropy, we include them as indicators of language complexity, which is also studied in previous work.

While the mean word length indeed shows a moderate correlation with the transfer performance (likely due to the differences in scripts), SuE comes out as a much better predictor. TTR and unigram entropy are only weakly correlated, which can be seen as an indirect replication of the negative results of Mielke et al. (2019). It is reasonable to expect that more complex languages (high TTR and entropy indicate higher complexity) could be more helpful for non-Indo-European languages (often complex themselves), but we do not observe this. Subword Evenness (SuE) explains the transfer success much better, probably by capturing a text property that is more relevant to the model's performance.

Target	Transfer	Avg PPL Change	Feature (Lang2vec)
Arabic	Hebrew	+0.04	Syntactic, Phonological
	Persian	+0.3	Inventorial
Burmese	Korean	+0.12	Syntactic
	Thai	+0.23	Phonological
	English	-0.07	Inventorial
Chamorro	Tagalog	+16.45	Syntactic, Phonological
	Indonesian	+7.15	Inventorial
Fijian	Tagalog	+5.72	Syntactic, Phonological
	Indonesian	+3.14	Inventorial
Hausa	Thai	+1.22	Syntactic
	Vietnamese	+1.95	Phonological
	Indonesian	+2.89	Inventorial
Khalkha	Japanese	-1.35	Syntactic
	English	+7.12	Phonological
	Finnish	+13.43	Inventorial
Malagasy	Tagalog	+0.63	Syntactic, Inventorial
	Indonesian	+0.17	Phonological
Oromo	Basque	+2.04	Syntactic
	Vietnamese	+2.85	Phonological
	Indonesian	+2.33	Inventorial

Table 6: Average increase in perplexity (across 3 models), when using a typologically close language instead of a language with low SuE (best PPL option among top 5 is chosen). Higher positive numbers mean worse performance.

5.3 SuE vs. genealogical and typological proximity

Other than text-based measures, typical predictors of transfer learning studied in previous work are measures of language similarity extracted from linguistic databases. To see how SuE compares with such measures, we analyse changes in complexity when SuE is replaced by linguistic measures. We do not perform correlation tests in these cases because linguistic measures are not available for all the languages in our experiments.

Table 5 shows the comparison between SuE and genealogical proximity. For this comparison, we look for pairs of languages where both the transfer and the target language belong to the same family according to the genealogical hierarchy presented in WALS (Haspelmath et al., 2005). Since our corpus maximizes linguistic diversity, most of the languages in our study do not have such close relatives. Nevertheless, we identify a sub-sample of 8 target languages whose relatives are among transfer languages. For each of these languages, we check whether the language model's perplexity is reduced more if the related language is used for transfer compared to one of the SuE top 5 languages (with lowest SuE), which are not related to the target language. The positive scores in Table 5 mean that at least one of the SuE top 5 languages is always a better transfer language than the genealogically closest language in our sample for all the target languages which are tested.

Table 6 shows the comparison between SuE and typological proximity for the target languages for which we could extract reliable feature vectors from the URIEL database (Littell et al., 2017). To assess the typological proximity, we use syntactic, phonological and inventorial features. Again, we obtain mostly positive scores, meaning that at least one of the SuE top 5 languages is almost always a better option than the typologically closest language in our sample. There are two cases where the language chosen according to the typological proximity performed better: English as a transfer for Burmese and Japanese as a transfer for Khalkha (Mongolian). The case of Khalkha is quite special, since this language appears to be hard to model: its perplexity is among the highest values across the three models. On the other hand, Japanese is actually among the good predictors according to SuE, just not in the strict top 5 (Table 3). In case of transfer between English and Burmese, we can see that the change in perplexity is rather low (-0.07), and generally English does not appear to be a good transfer language for any other languages in our experiments.

Our results show that proximity to the transfer language in terms of genealogical or typological properties is often not the best criterion for choosing a good transfer language when working with low-resource languages. At least one transfer language that we identify as most suitable (top 5 lowest SuE values) gets consistently better (lower) perplexity scores than genealogically or typologically close languages. Sometimes the best transfer language overlaps between all three choice methods, but our measure SuE is still more consistent, and does not depend on the test low-resource language and does not need any linguistic annotation (purely data-driven).

6 Stability of SuE across different corpora

Additionally, we check how corpus size and change of data can influence our measure of SuE. A systematic study on the corpus dependence would exceed

lang	SuE (TeDDi 1mio)	SuE (TeDDi full)	diff
eus	118.15	117.05	-1.11
eng	118.04	114.54	-3.51
fin	118.35	115.95	-2.40
fra	116.70	119.61	2.91
deu	115.14	114.25	-0.89
ell	101.66	108.10	6.44
heb	97.94	115.90	17.96
hin	121.33	112.52	-8.81
ind	123.49	116.15	-7.34
jpn	101.69	77.72	-23.97
kor	100.54	77.39	-23.15
cmn	91.14	71.71	-19.42
pes	126.21	109.55	-16.67
rus	119.61	111.63	-7.98
spa	121.27	124.51	3.25
tgl	115.89	109.00	-6.90
tha	101.33	104.25	2.93
tur	105.68	112.24	6.56
vie	118.29	99.36	-18.93

Table 7: Difference in SuE values when measured on 1 mio TeDDi sample and on the full TeDDi corpus.

lang	SuE (TeDDi 1mio)	SuE (TeDDi full)	SuE (Aalto)
eng	118.04	114.54	126.14
fin	118.35	115.95	121.19
tur	105.68	112.24	118.07

Table 8: Difference in SuE values when measured on 1 mio TeDDi sample, on the full TeDDi corpus and on the Aalto corpus.

the scope of this paper, so we provide only several comparisons here.

We measured SuE on the full TeDDi corpus and compared the results with the SuE values obtained on a balanced sample from the TeDDi corpus with 1 million tokens (Table 7). Then, we used the Aalto MorphoChallenge corpus (Kurimo et al., 2010)[7], which provides large amounts of data in English, Finnish and Turkish coming mostly from Europarl (Koehn, 2005) in order to check SuE on data from a different source. We compare the obtained values to the two TeDDi versions in Table 8.

Table 7 demonstrates that the differences in languages with Latin scripts are smaller than in languages written in non-Latin scripts. Nevertheless, average difference across all languages is about 5 degrees, which is rather low. The correlation between the values of the TeDDi sample with 1 million tokens and the full TeDDi is high, we get a 0.7 Pearson coefficient correlation (p-value = .002).

Table 8 shows that fluctuations range between 2 and 12 degrees with no apparent correlation to the data size.

While these experiments show that the rankings of SuE values remain generally stable after changing data source or size, we find that the best predictions are found when using the balanced 1 million tokens sample from the TeDDi corpus.

7 Conclusion

In this study, we tackled the question of what languages are preferable for cross-lingual transfer when modelling diverse low-resource languages, and what motivates this selection. Our experiments on the task of masked language modelling (MLM) with three multilingual pre-trained Transformer-based models show that there is a small set of *generally* good transfer languages: Japanese, Greek, Hebrew, Thai, and Russian. What is common to these languages is the fact that all of them are written in a non-Latin script. However, the script alone is not the best general predictor of transfer performance. We show that the best predictor of the language model perplexity on a wide range of target low-resource languages is the Subword Evenness (SuE) score of transfer languages, which we have presented in this paper. Most of the languages that repeatedly come out as good transfer languages have low SuE scores.

Our results are largely in line with the observations about good transfer languages made in previous work. In addition to providing new evidence confirming previous findings on a very diverse sample of languages (19 transfer and 30 target languages), we identify the strongest predictor of the observed performances up to now. The proposed SuE measure is a better predictor for transfer languages than other text-based measures (mean word length, type-token ratio, unigram entropy) as well as genealogical and typological proximity. With this finding, we make a further step towards explaining why continued training on transfer languages is helpful for modelling low-resource languages.

Acknowledgements

This research is supported by the Swiss National Science Foundation (SNSF) grant 176305.

[7] http://morpho.aalto.fi/events/morphochallenge2010/

Limitations

Our focus on low-resource languages limits to a certain degree the generalization of our findings. While our data represents a carefully designed language sample, the decisions made by the authors of the sample are arbitrary, which means that our samples are not random and the finding might not generalize to all languages. For example, our sample of target languages does not include any Indo-European languages, such as Germanic or Romance low-resource languages. These languages have been studied before and it has been shown that the best choice for them is transferring from a genealogically related rich-resource language (Aepli and Sennrich, 2021). It might be interesting to see how our proposed measure would compare with other measures in these cases, but this would require a different study design, which we leave for future work.

Another limitation of our work concerns the data size thresholds that we use to divide the languages into low-resource (target) and high-recourse (transfer) languages. In our experiments, the size of 100K tokens is assigned to the low-resource group. We took this decision taking into account two criteria. First, we wanted to have reasonable models for comparing the performance and we judged this size reasonable. Second, we identified this threshold by checking how many test languages we could still draw out of the TeDDi sample, and the number of 30 (1/3 of TeDDi's languages) seems to be a relatively good test size. While this procedure might leave some of the really low-resource languages out of our sample, we still cover a wide variety of languages for which at least some texts (100K tokens is approximately the size of a shorter novel) are available.

Finally, our experiments are limited to one task (masked language modelling) and the results might not generalize to tasks requiring annotated data. We note that, despite this limitation, our results are largely in line with previous work on various tasks.

References

Noëmi Aepli and Rico Sennrich. 2021. Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. *arXiv preprint arXiv:2109.06772*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Yoav Goldberg. 2017. *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online. Association for Computational Linguistics.

Ximena Gutierrez-Vasques, Christian Bentz, Olga Sozinova, and Tanja Samardzic. 2021. From characters to words: the turning point of BPE merges. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3454–3468, Online. Association for Computational Linguistics.

Martin Haspelmath, Matthew S Dryer, David Gil, and Bernard Comrie. 2005. *The world atlas of language structures*. OUP Oxford.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Mikko Kurimo, Sami Virpioja, Ville T Turunen, et al. 2010. Proceedings of the morpho challenge 2010 workshop. In *Morpho Challenge Workshop; 2010; Espoo*. Aalto University School of Science and Technology.

Ron Larson and Robert Hostetler. 2007. *Trigonometry*. Houghton Mifflin Company.

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.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

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.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings from LREC 2016*, pages 923–929. European Language Resources Association.

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.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 3158–3163.

Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *CoRR*, abs/2112.10508.

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

Steven Moran, Christian Bentz, Ximena Gutierrez-Vasques, Olga Sozinova, and Tanja Samardzic. 2022. TeDDi sample: Text data diversity sample for language comparison and multilingual NLP. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1150–1158, Marseille, France. European Language Resources Association.

Michael Paul, Andrew Finch, and Eiichrio Sumita. 2013. How to choose the best pivot language for automatic translation of low-resource languages. *ACM Transactions on Asian Language Information Processing (TALIP)*, 12(4):1–17.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Brian D Ripley. 2002. *Modern applied statistics with S*. Springer.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, et al. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. *arXiv preprint arXiv:2104.07412*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Anastassia Shaitarova and Fabio Rinaldi. 2021. Negation typology and general representation models for cross-lingual zero-shot negation scope resolution in Russian, French, and Spanish. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 15–23, Online. Association for Computational Linguistics.

Benjamin Snyder, Tahira Naseem, and Regina Barzilay. 2009. Unsupervised multilingual grammar induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 73–81, Suntec, Singapore. Association for Computational Linguistics.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2008. Unsupervised multilingual learning for POS tagging. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1041–1050, Honolulu, Hawaii. Association for Computational Linguistics.

Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. *Natural Language Processing with Transformers. Building Language Applications with HuggingFace*. O'Reilly Media.

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 arXiv:2106.16171*.

Lonneke van der Plas and Jörg Tiedemann. 2006. Finding synonyms using automatic word alignment and measures of distributional similarity. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 866–873, Sydney, Australia. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 255–264, Online. Association for Computational Linguistics.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020a. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. 2020b. On negative interference in multilingual models: Findings and a meta-learning treatment. *arXiv preprint arXiv:2010.03017*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

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 arXiv:1609.08144*.

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

Ziyi Yang, Yinfei Yang, Daniel Cer, and Eric Darve. 2021. A simple and effective method to eliminate the self language bias in multilingual representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5825–5832.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

A Results with alternative subword tokenization algorithms

Measure	Model	P	S	Avg, P	Avg, S
SuE (BPE Gutierrez-Vasques et al. (2021))	mBERT	0.4	**0.62** $p = .005$	**0.5** $p = .03$	**0.77** $p = .0002$
	XLM-R	**0.56** $p = .01$	**0.69** $p = .002$		
	mT5	0.18	0.11		
SuE (BPE Mielke et al. (2019))	mBERT	0.47	0.28	0.44	0.36
	XLM-R	0.37	0.33		
	mT5	**0.38** $p = .11$	**0.37** $p = .12$		
SuE (Morfessor)	mBERT	**0.49** $p = .03$	0.26	0.49	0.29
	XLM-R	0.44	0.31		
	mT5	0.17	0.24		
SuE (WordPiece)	mBERT	0.32	0.03	0.3	0.07
	XLM-R	0.26	0.11		
	mT5	0.15	0.06		
SuE (SentencePiece Model)	mBERT	0.21	0.24	0.15	0.03
	XLM-R	0.1	0.1		
	mT5	0.11	0.15		

Table 9: Pearson (P) and Spearman (S) correlation coefficients between SuE, measured on different segmentation outputs (same input as described for the training datasets, Section 4), and the average perplexity values (across all test languages). P-values are given for the highest correlation coefficients. The method we use (BPE Gutierrez-Vasques et al. (2021)) gives the highest average correlations (across all 3 models) and the highest results for XLM-R and mBERT (Spearman), that is why it was chosen as a final segmentation method to base our analysis on. If one aims at working with mT5 only, the BPE method of Mielke et al. (2019) (more merges), would provide SuE values that predict the perplexity better. Note, however, that working with the actual segmentation methods used in mBERT (WordPiece), XLM-R and mT5 (SentencePiece) does not provide good predicitons.

B Detailed language model perplexity scores

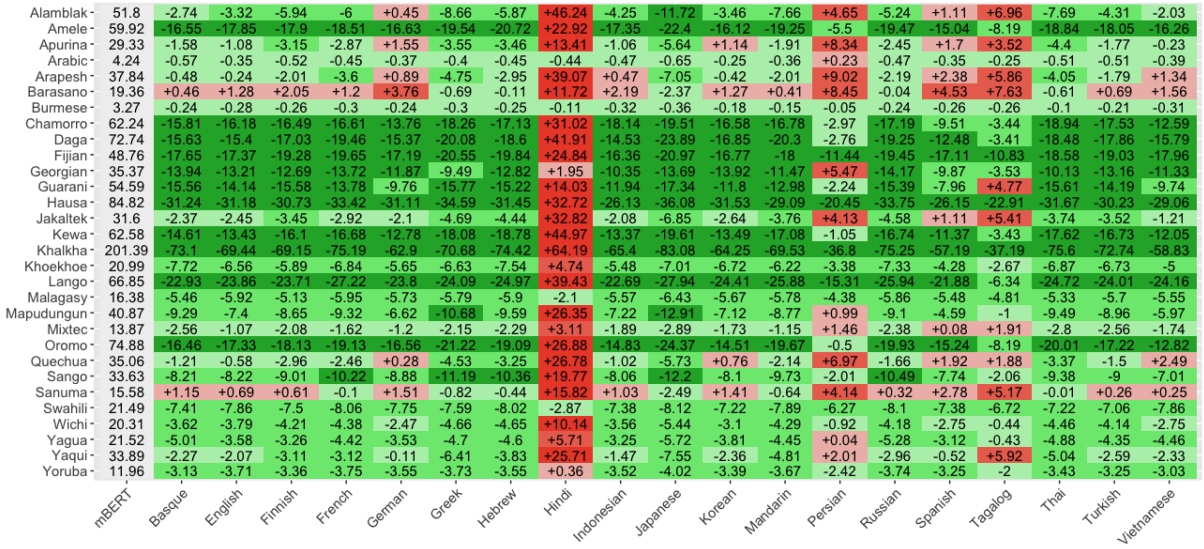


Figure 5: Difference in perplexity compared to pre-trained mBERT. Green means better results (lower perplexity).

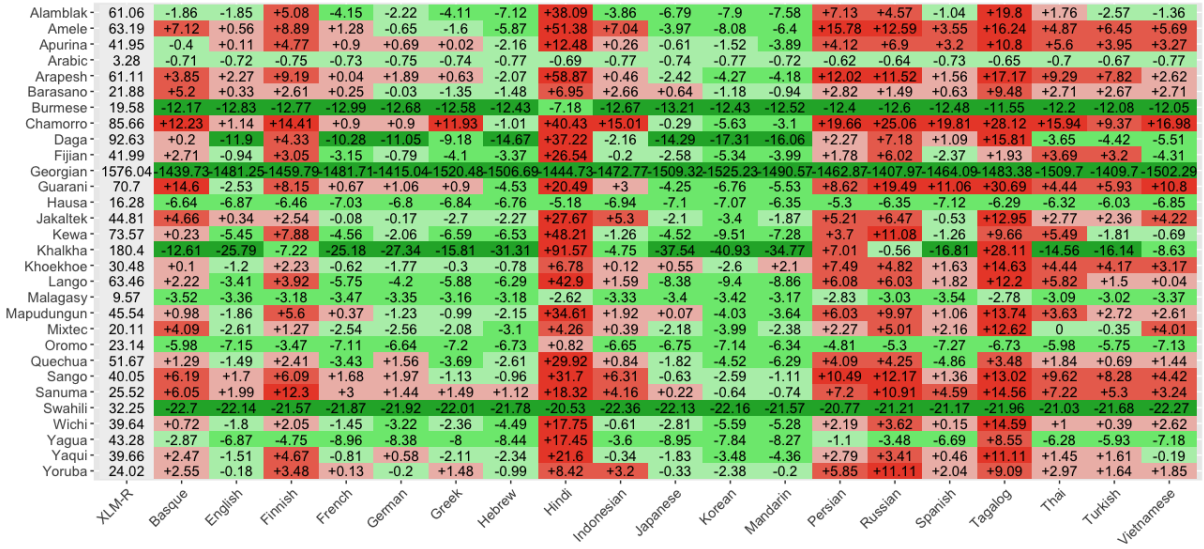


Figure 6: Difference in perplexity compared to pre-trained XLM-R. Green means better results (lower perplexity).

Figure 7: Difference in perplexity compared to pre-trained mT5. Green means better results (lower perplexity).

C Details on the previous work

Transfer language	Target language (+ Source for MT)	Task/Dataset	Comments	References
English	Spanish –> 10 languages of Americas	NMT	Best model in the AmericasNLP shared task	Vazquez et al. 2021
	Spanish, German, Russian (mBERT, XLM-R)	LAReQA	Best performance of the multilingual models (top 3); zero-shot transfer	Ruder et al. 2021
	German, Spanish, Polish (mBERT) German, Polish, Spanish (XLM-R)	Mewsli-X (EL)		
	Dutch, Italian, Portuguese (mBERT) Italian, Afrikaans, Dutch (XLM-R)	UD-POS		
	Spanish, French, German (mT5)	NLI		
	Slovak, Italian, Russian (mBERT) Slovak, Italian, Finnish (XLM-R)	DEP	Smallest performance drop compared to English (top 3); zero-shot transfer	Lauscher et al. 2020
	Slovak, Russian, Italian (mBERT) Slovak, Russian, Finnish (XLM-R)	POS		
	Italian, Finnish, Slovak (mBERT) Finnish, Slovak, Italian (XLM-R)	NER		
	French, Spanish, German (mBERT) Spanish, French, Bulgarian (XLM-R)	NLI		
	German, Spanish, Russian (mBERT) Spanish, Thai, Russian (XLM-R)	XQuAD		
	Dutch, Italian, Portuguese (mBERT) Dutch, Portuguese, Bulgarian (XLM-R)	NER	Best performance of the multilingual models (top 3); zero-shot transfer	Hu et al. 2020
	Dutch, Italian, Spanish (mBERT) Afrikaans, Dutch, Russian (XLM-R)	POS		
	Spanish, French, German (mBERT) Spanish, Bulgarian, German (XLM-R)	NLI		
	Spanish, French, German (mBERT) French, Spanish, German (XLM-R)	PAWS-X		
	Spanish (mBERT, XLM-R)	MLQA	Best performance of the multilingual models (top 1); zero-shot transfer	
	Indonesian (mBERT, XLM-R)	TyDiQA		
	Spanish (mBERT, XLM-R)	XQuAD		
	203 language pairs	Pivot-based MT	Best pivot choice according to the BLEU score	Paul et al. 2013

Chinese	Hindi, Urdu (mBERT, mT5) Swahili, Thai (mT5)	NLI	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021
	Japanese, Korean (mBERT, mT5)	PAWS-X	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021
	Japanese <–> Korean	Pivot-based MT	Gain in BLEU points compared to English pivot (avg 4.7)	Paul et al. 2013
Turkish	Hindi (mBERT, mT5) Urdu (mT5)	NLI	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021
	Azerbaijani –> English	NMT	Best transfer language for NMT into English among 54 transfer options; concatenated transfer	Lin et al. 2019
Hindi	Urdu (mBERT, mT5)	NLI	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021
	Telugu –> English	EL	Best transfer language for EL to an English knowledge database among 53 transfer options; zero-shot transfer	Lin et al. 2019
Japanese	Chinese, Korean (mT5)	PAWS-X	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021
	Korean –> 19 languages 6 languages –> Korean Chinese –> Hindi	Pivot-based MT	Gain in BLEU points compared to English pivot (avg 2.5)	Paul et al. 2013
Korean	Japanese (mBERT, mT5), Chinese (mT5)	PAWS-X	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021
	Japanese –> 9 languages 10 languages –> Japanese	Pivot-based MT	Gain in BLEU points compared to English pivot (avg 1.9)	Paul et al. 2013
Spanish	Hindi (mBERT, mT5) Urdu, Thai, Chinese (mBERT)	NLI	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021
	English –> 12 languages French, Dutch –> English French, Italian –> Chinese	Pivot-based MT	Gain in BLEU points compared to English pivot (avg 0.8)	Paul et al. 2013
	Russian (mBERT, XLM-R)	Automated negation detection	Higher accuracy compared to French, English (XLM-R)	Shaitarova et al. 2021
Russian	Hindi, Urdu, Thai (mBERT, mT5) Swahili, Arabic, Chinese (mT5) Bulgarian (mBERT)	NLI	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021
	Arabic, Thai, Turkish (mBERT)	XQuAD	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021

German	Hindi, Urdu, Thai (mBERT, mT5) Chinese (mBERT)	NLI	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021
	Greek, Thai (mBERT)	XQuAD	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021
Greek	Hindi, Thai (mBERT, mT5) Swahili, Arabic, Urdu (mT5)	NLI	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021
Bulgarian	Hindi (mBERT, mT5) Swahili, Turkish, Arabic, Urdu, Thai, Chinese (mT5)	NLI	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021
Vietnamese	Thai (mBERT, mT5) Hindi (mBERT)	NLI	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021
French	Arabic, Urdu, Chinese (mBERT)	NLI	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021
Thai	Chinese (mT5)	NLI	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021
Urdu	Hindi (mT5)	NLI	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021
Arabic	Russian, Bengali, Finnish, Indonesian, Korean, Telugu (mT5)	XQuAD	Gain compared to transfer from English (>= 3.0 points)	Turc et al. 2021
Hungarian	Bengali –> English	NMT	Best transfer language for NMT into English among 54 transfer options; concatenated transfer	Lin et al. 2019
Indonesian	Malay –> 20 languages 15 languages –> Malay	Pivot-based MT	Gain in BLEU points compared to English pivot (avg 2.4)	Paul et al. 2013
Portuguese	English <–> Spanish Italian, Arabic, Tagalog, Vietnamese –> English	Pivot-based MT	Gain in BLEU points compared to English pivot (avg 2.3)	Paul et al. 2013
Malay	Indonesian –> 17 languages 16 languages –> Indonesian Chinese –> Russian Vietnamese –> Chinese	Pivot-based MT	Gain in BLEU points compared to English pivot (avg 1.8)	Paul et al. 2013
Dutch	Danish, German –> English English, Chinese –> German Chinese –> Arabic	Pivot-based MT	Gain in BLEU points compared to English pivot (avg 0.6)	Paul et al. 2013

Table 10: Detailed examples of transfer languages which yielded the best results in the previous works.