Inria-ALMAnaCH at the WMT 2022 shared task: Does Transcription Help Cross-Script Machine Translation?

Jesujoba O Alabi, Lydia Nishimwe, Benjamin Muller, Camille Rey, Benoît Sagot, Rachel Bawden

▶ **To cite this version:**

Jesujoba O Alabi, Lydia Nishimwe, Benjamin Muller, Camille Rey, Benoît Sagot, et al.. Inria-ALMAnaCH at the WMT 2022 shared task: Does Transcription Help Cross-Script Machine Translation?. EMNLP 2022 - Seventh Conference on Machine Translation (WMT22 - Workshop on Statistical Machine Translation), Dec 2022, Abu Dhabi, United Arab Emirates. hal-03836180

HAL Id: hal-03836180
https://inria.hal.science/hal-03836180

Submitted on 1 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Inria-ALMAnaCH at the WMT 2022 shared task: Does Transcription Help Cross-Script Machine Translation?

Jesujoba O. Alabi[1]* **Lydia Nishimwe**[2] **Benjamin Muller**[2,3]
Camille Rey[2] **Benoît Sagot**[2] **Rachel Bawden**[2]
[1]Spoken Language Systems (LSV), Saarland University, Saarland Informatics Campus, Germany
[2]Inria, Paris, France [3]Sorbonne Université, France
jalabi@lsv.uni-saarland.de firstname.lastname@inria.fr

Abstract

This paper describes the Inria ALMAnaCH team submission to the WMT 2022 general translation shared task. Participating in the language directions {cs,ru,uk}→en and cs↔uk, we experiment with the use of a dedicated Latin-script transcription convention aimed at representing all Slavic languages involved in a way that maximises character- and word-level correspondences between them as well as with the English language. Our hypothesis was that bringing the source and target language closer could have a positive impact on machine translation results. We provide multiple comparisons, including bilingual and multilingual baselines, with and without transcription. Initial results indicate that the transcription strategy was not successful, resulting in lower results than baselines. We nevertheless submitted our multilingual, transcribed models as our primary systems, and in this paper provide some indications as to why we got these negative results.

1 Introduction

This paper describes the Inria ALMAnaCH team submission to the WMT 2022 general translation shared task. We chose to explore the language directions {cs,ru,uk}↔en and cs↔uk in order to concentrate on the Slavic language family. Due to some experimental problems that impacted the into-Slavic directions most heavily, we only submitted {cs,ru,uk}→en and cs↔uk language directions, but we present all results we obtained here.

A major area of interest in machine translation (MT) research is transfer between languages, particularly related ones and for lesser resourced languages (Zoph et al., 2016; Kocmi and Bojar, 2018). One way of encouraging transfer is to train multilingual models, whereby several language directions are trained simultaneously, often sharing some (Firat et al., 2016) or all model parameters (Ha et al., 2016; Johnson et al., 2017; Aharoni et al., 2019), with the hope that similarities between the languages can boost performance, particularly for the lower-resourced languages.

To encourage lexical sharing and therefore the transfer capacity of such models, joint subword segmentation models (Sennrich et al., 2016b) and MT vocabularies are often used (Sennrich et al., 2016a), and techniques such as phonetisation and transliteration/transcription can be applied to texts in a bid to overcome differences in writing systems and spelling (Nguyen and Chiang, 2017; Chakravarthi et al., 2019; Goyal et al., 2020; Muller et al., 2021).

In our submission to the WMT 2022 general translation shared task, we experimented with multilingual models and the use of customised transcription into a common writing system designed to maximise lexical sharing, similar to the one used in (Muller et al., 2021). We choose to work with the language directions involving Slavic languages, that is {cs,ru,uk}↔en and cs↔uk. We find that our transcription method unfortunately leads to degraded results, likely a consequence of errors being injected and notably the necessity to apply a learned detranscription model as a post-processing step for into-Slavic language directions. Our multilingual models achieved largely inferior results to our bilingual baseline models for the same number of parameters, showing that multilingual transfer cannot compensated for sharing the vocabulary over a larger number of languages. Transcribing the languages in the multilingual setup results narrows the gap slightly, but the results remain lower than the bilingual baselines. We nevertheless decided to submit our multilingual models with common-Slavic transcription rather than our superior baseline results in the full knowledge that these results would not achieve the best results in the shared task.[1]

*Contributions made whilst at Inria.

[1]We believe it was more interesting to submit these results to test our hypothesis rather than to submit more standard

cs	original	*Sníh pokryl stromy vedle zámku.*
cs	transcribed	*Snig pokril stromi vedle zamku.*
uk	original	Сніг вкрив дерева біля замку.
uk	transliterated	*Snih vkryv dereva bilja zamku.*
uk	transcribed	*Sneg vkriv dereva bela zamku.*
ru	original	Снег покрыл деревья возле замка.
ru	transliterated	*Sneg pokryl derev'ja vozle zamka.*
ru	transcribed	*Sneg pokril dereva vozle zamka.*
en	original	The snow has covered the trees next to the castle.

Table 1: Constructed example illustrating the difference between standard transliteration and our linguistically motivated transcription.

2 Related Work

There has been a considerable body of work in MT dedicated to multilingual models, whereby several language directions are trained simultaneously, with different degrees of parameter sharing, ranging from separate encoders and decoders (Firat et al., 2016) to the sharing of a single encoder and a single decoder for all languages with a single shared vocabulary (Ha et al., 2016; Johnson et al., 2017; Aharoni et al., 2019). As well as being practical by providing a single MT model that can be used for multiple directions, the models have the advantage of aiding the representations of lower-resourced languages, particularly if related, higher-resourced languages are also included in training (Kudugunta et al., 2019; Aharoni et al., 2019; Tchistiakova et al., 2021).

In addition to approaches such as joint subword segmentation models (Sennrich et al., 2016b) and the use of a joint vocabulary for all languages (Johnson et al., 2017), strategies to encourage more lexical sharing have also been explored in order to overcome surface differences introduced by orthographic conventions, notably phonetisation (Liu et al., 2019; Rosales Núñez et al., 2019; Sun et al., 2022) and transliteration (Nakov and Tiedemann, 2012; Nguyen and Chiang, 2017; Goyal et al., 2020). These approaches can be particularly useful for borrowings and for proper nouns, which can be made to be identical (or near-identical) across languages once transliteration has been applied.

Transliteration is the mapping of one writing system to another, and therefore is relevant when languages are written in different scripts (e.g. Latin, Cyrillic, Devanagari, etc.). In particular for related languages, it can be interesting to apply transliteration in order to exploit the fact that many words can be made to be similar on the surface once transliteration has been applied. Much of the work that has explored transliteration for MT has focused on Indian languages, for which the mapping between scripts is relatively straightforward (Bawden et al., 2019; Goyal et al., 2020; Kunchukuttan and Bhattacharyya, 2021; Sun et al., 2022), but there has also been research on other language families (Maimaiti et al., 2019; Sun et al., 2022), including Slavic languages (Maimaiti et al., 2019). In our systems, we follow a similar approach to test whether a form of transliteration that maximises lexical overlap between Slavic languages could help translation in a multilingual setup, even in the relatively high-resource scenario provided by the shared task.

3 Multilingual Slavic models with transcription

Building on the previous work on multilingual MT and on transliteration to encourage lexical sharing, we propose multilingual models with a custom linguistically motivated transcription scheme for translation between English and the Slavic languages Czech (cs), Ukrainian (uk) and Russian (ru).

Multilingual Slavic translation models We train multilingual Slavic translation models with a single encoder-decoder architecture as in (Johnson et al., 2017) over the following language directions: {cs,uk,ru} from and into English and cs to and from uk. Given that a single shared encoder and a single shared decoder is used, the same vocabulary is used across all languages, and we also share embeddings across the encoder and decoder. To further encourage sharing, we train a joint subword segmentation model. To test the performance of this multilingual model, we compare against bilingual baselines trained uniquely on parallel data for the

baseline systems. Due to human error, these submitted models perform less well than the results presented in this paper, as described in Section 5.3.

specific language pair, which also share encoder and decoding embeddings.

Linguistically motivated transcription We experiment with the use of a customised common Slavic writing system designed with the aim of maximising lexical overlap between the Slavic languages we study. The underlying idea is that MT models, both bilingual and multilingual, should benefit from an increase in the similarity between languages including in training. Since Slavic languages share a common ancestor, Proto-Slavic, they display similarities in terms of phonetics, grammar and vocabulary. Lexical overlap, though, can be further improved in at least two ways:

- Whereas Czech uses the Latin script with a number of diacritics, Russian and Ukrainian use the Cyrillic script. Using a common script would inevitably increase the lexical overlap and make it more explicit. For instance, using a standard Latin transliteration scheme for Russian,[2] the Russian word рука 'hand' can be rendered as *ruka*, which is identical to Czech *ruka* 'hand'.

- Each Slavic language has undergone a number of changes from Proto-Slavic, including regular sound changes. Examples such as Ru. рука∼*ruka* vs. Cz. *ruka*, where transliteration alone is enough to create a perfect lexical overlap, are therefore rare. However, there are a large number of cognates (words in related languages that share a common ancestor), which, independently of the script, are still similar and only differ in partly systematic ways. For instance, Ru. корень 'root', Uk. корінь 'id.' and Cz. *kořen* 'id.' are cognates. Using standard transliteration schemes, the Russian and Ukrainian words can be rendered as *koren'* and *korin'*, respectively. This is closer to Cz. *kořen* but is not identical. More importantly, it fails to identify the fact that Uk. і often corresponds to Cz. *e* and that Cz. *ř* often corresponds to Ru. and Uk. р.

To further increase lexical overlap and with the aim of encouraging more transfer between the languages than what is permitted by standard transliteration schemes, we developed transformation rules for all three Slavic languages based on systematic patterns, based on observations from cognate lists in the three languages and knowledge about their morphology, in order to lower the differences introduced between them by sound changes and morphological particularities, similarly to (Muller et al., 2021). For Russian and Ukrainian, this involves a script change, but Czech is also modified. We call this transformation *linguistically motivated transcription*.[3] Going back to the example above, the output of our transcription scripts for Ru. корень, Uk. корінь and Cz. *kořen* is the same, namely *koren*. Table 1 illustrates our linguistically motivated transcription strategies on a constructed multilingual example.

Transcription and detranscription Our common Slavic transcription is applied during preprocessing to the training data. For into-English language directions, no further processing is required following translation, because we only transcribe the Slavic languages and not English. However for from-English directions and for cs↔uk, the output of the MT model will require detranscription in order to transform the outputs into the correct form for that language. We therefore also train small transcription models, which are essentially individual translation models trained to translate from the transcribed text to the original writing system. This step can be trained on large quantities of monolingual data rather than being limited to parallel data, which is important if error propagation is to be kept to a minimum.

4 Data

We developed systems for four of the several language combinations taken into account for the general translation task. They are {cs,ru,uk}↔en and cs↔uk. We took part in the challenge under its constrained track, using only a portion of the data made available for the task. The following sections describe the data we used and how we processed and filtered it. We present the data sizes and their corresponding sources in Table 8 in Appendix A.

[2]Here and elsewhere in this paper, we use the so-called "scientific transliteration" when we transliterate Russian and Ukrainian. It is a standard, slightly language-dependent, transliteration scheme.

[3]Transliteration is generally defined as a bijective script change, a constraint that is too strict to allow for a significant increase in lexical overlap. Relaxing the bijectivity constraint, on the other hand, means that some information is lost. It is no longer a transliteration *stricto sensu*. Contrarily to (Muller et al., 2021), we therefore use the term *transcription* rather than transliteration to denote a transformation process that performs non-bijective changes.

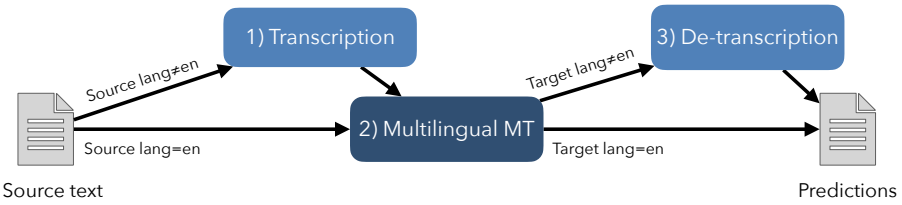

Figure 1: Illustration of our multilingual MT approach using common Slavic transcription.

4.1 Parallel Data

We used all of the parallel data provided for the language pairs we selected, with the exception of the back-translated news data, CzEng2.0 and two more datasets released at a later stage of the challenge, ELRC-EU acts, and Yakut parallel data, for the training of our NMT systems. We excluded the back-translated news data[4] and CzEng 2.0,[5] which are both back-translated data sources, after inspecting their respective content and discovering a large proportion of poorly translated sentences. To assess their quality and gauge the amount of noise present, the other parallel data were carefully examined. This was important especially for the web-mined data such as CCAligned, Wikimatrix, and CommonCrawl, which all contained a variety of quality issues identified in (Kreutzer et al., 2022).

Parallel Data Filtering: Each parallel corpus was subjected to a generic filtering pipeline involving the removal of blank lines and sentences without corresponding translations. We carried out language identification on the web-mined parallel corpora using FastText (Joulin et al., 2016a, 2017), thus removing sentence pairs where either the source or target is not in the intended language. Finally, the parallel corpora for each language pair were combined, and duplicate translation pairs were removed. Table 2 shows the original number of parallel sentences for the different language pairs and their corresponding sizes after filtering.

Language pair	Original	Filtered
cs–en	56,289,558	54,495,258
cs–uk	3,163,969	2,490,622
en–ru	31,052,852	25,584,007
en–uk	23,355,100	22,322,394

Table 2: Number of parallel sentences.

[4] http://data.statmt.org/wmt20/translation-task/back-translation/

[5] https://ufal.mff.cuni.cz/czeng/czeng20

4.2 Monolingual Data

We used monolingual data to train the detranscription models. As with the parallel data, we removed empty lines, duplicated lines and also sentences that were not from the target language by doing language identification with FastText (Joulin et al., 2017, 2016b). This process was necessary since most of these sentences were web-mined text. The statistics of the monolingual data for each language are shown in Table 9 in Appendix A, along with their sizes before and after pre-processing.

For the transcription experiments, we randomly selected 20M sentences from the pre-processed monolingual texts for each of the Slavic languages.

4.3 Validation and Test Data

For each language pair, we chose 2000 and 3000 sentence pairs from the pre-processed parallel texts as our internal validation and test sets respectively, and the remaining sentences were used for training. In order to compare the various systems we developed, we also used the development set provided for the shared task (the FLORES development set and the WMT2018 test set depending on the language pair). This was also done for the systems with transcription. En↔uk and cs↔en models were only evaluated on the in-house test and the FLORES development sets because they were not in covered by the WMT2018 test sets. We also provide automatic scores on the WMT2022 test sets.

4.4 Subword Tokenisation

We tokenised all data using a joint SentencePiece (Kudo and Richardson, 2018) unigram model with a character coverage of 1.0 and a maximum sentence length of $4,096$ tokens. Specifically, for the bilingual systems, we uniformly sampled $5M$ monolingual sentences from the parallel training data of each language pair to have $10M$ sentences in total over which we trained a SentencePiece tokeniser. Similarly, for multilingual systems, we

sampled a total of $10M$ monolingual sentences evenly from all monolingual data available for each language that the tokeniser was trained on.

5 Experiments and training

We submitted three categories of NMT systems: (i) the baseline bilingual translation models for each of the four language pairs in their original scripts, (ii) a multilingual model with common-Slavic transcription for {cs,uk,ru}→en, and (iii) a bilingual model with common-Slavic transcription for cs↔uk. Below, we provide details of these submitted systems, as well as the additional systems developed before and after the task's deadline.

5.1 NMT architecture and training

All models used the transformer-base architecture (Vaswani et al., 2017) within the Fairseq[6] toolkit (Ott et al., 2019). We use the `multilingual_translation` architecture for all models, except for those trained on a single language pair. We used batch sizes of $10,240$ tokens, a maximum sentence length of $1,024$, and a dropout of 0.3. For optimisation, we used Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.998$, a learning rate of $1 * 5e^{-}5$ and a warm-up of $4,000$ updates. The optimiser uses a label-smoothed cross-entropy loss function with a label-smoothing value of 0.1. For multilingual models we use temperature sampling with $T = 1.5$. All models were trained until convergence based on the BLEU score on the development set. We use BLEU (Papineni et al., 2002) to evaluate our models and to choose the best checkpoints, calculated using SacreBLEU[7] (Post, 2018).

5.2 Baseline models

We trained a bilingual translation model for each of the four language pairs we covered. We chose a vocabulary size of $64k$ for all systems after experimenting with different sizes ($16k$, $32k$, and $64k$). We then fine-tuned each bilingual model to each of the two directions in the language pair (taking the best checkpoint of the bilingual model), resulting in a baseline model for each of the 8 translation directions.

We also trained a multilingual system for all of the language pairs, i.e. a single model that can translate in every direction, which was then finetuned to each language direction. We chose to use a vocabulary size of $64k$ based on the trends we found from the bidirectional model experiments. We chose not to go bigger in order to keep the model compact and comparable to the bilingual baselines, at least in terms of the number of parameters. Our comparison therefore tests whether for a same number of parameters multilingual models (and transcription) can be beneficial, despite the fact that multilingual vocabs are likely to result in a higher degree of segmentation for the individual languages.

5.3 Common Slavic transcription

To assess the impact of transcription, we trained bilingual and multilingual models on the transcribed versions of the Slavic parallel data. We follow the same setup as for the baseline models (i.e. bilingual/multilingual training and then fine-tuning on the specific language direction), simply substituting the original Slavic text with the transcribed versions.[8] When presenting the results, we refer to the transcribed version of Russian (ru), Czech (cs) and Ukranian (uk) as rl, cl and ul respectively.

Due to human error, our submitted multilingual systems were trained with a vocabulary of 16k rather than 64k, which severely penalised them and resulted in very low official scores. We report results with the intended vocabulary size of 64k in this article.

5.4 Detranscription models

For each Slavic language, we trained a detranscription model on $20M$ parallel sentences (transcribed→original), consisting of monolingual sentences and their automatically transcribed versions. We used a joint SentencePiece model of size $16k$ and used the same architecture as before. These models were applied after translation to make sure that transcribed Slavic outputs were in their original writing system.

6 Results

6.1 Baseline models (without transcription)

We first report results for our baseline models in Table 3 (i.e. without transcription).

We provide results for our in-house test set (from the same distribution as the training data), the FLORES devtest subset and the WMT2018 test

[6] https://github.com/facebookresearch/fairseq

[7] With the following parameters: `case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1`

[8] SentencePiece models were also retrained on the new data, keeping a vocabulary size of 64k.

	en→cs	cs→en	cs→uk	uk→cs	en→ru	ru→en	en→uk	uk→en
				Bilingual				
*In-house Test	43.11	45.38	39.16	40.20	42.66	47.07	33.32	38.16
$\text{FLORES}_{devtest}$	29.05	33.70	19.76	20.86	24.60	28.65	24.11	30.03
WMT 2018	20.81	29.03	–	–	23.77	28.15	–	–
WMT 2022	**33.62**	**39.45**	**27.40**	**25.65**	**23.60**	**34.71**	**20.72**	**34.42**
				Multilingual				
*In-house Test	36.02	39.42	27.08	28.50	35.71	40.74	34.19	38.97
$\text{FLORES}_{devtest}$	21.53	27.22	11.56	13.41	15.22	21.43	17.48	24.76
WMT 2018	15.36	21.18	–	–	15.04	21.19	–	–
WMT 2022	24.95	26.89	18.61	17.43	16.70	26.66	16.66	27.49

Table 3: BLEU score results for bilingual and multilingual baseline models (i.e. without transcription).

	en→cs	cs→en	cs→uk	uk→cs	en→ru	ru→en	en→uk	uk→en
				Bilingual				
In-house Test	41.72	44.94	35.80	38.04	38.67	46.50	28.89	37.69
$\text{FLORES}_{devtest}$	28.83	33.56	19.05	20.09	21.77	27.93	21.94	28.86
WMT 2018	20.66	27.64	–	–	21.64	27.57	–	–
WMT 2022	33.42	37.83	26.43	24.96	21.22	34.43	18.69	32.7
				Multilingual				
In-house test	36.75	40.08	30.64	32.99	34.57	41.19	29.73	38.71
$\text{FLORES}_{devtest}$	22.34	28.15	15.90	16.22	17.96	22.64	20.62	24.78
WMT 2018	15.85	22.39	–	–	17.85	22.24	–	–
WMT 2022	26.08	29.00	23.20	21.15	17.66	27.32	18.36	28.48

Table 4: BLEU score results for bilingual and multilingual models using transcription for all Slavic languages.

set (when available). Although the BLEU scores are not directly comparable across test sets, the baseline results are generally quite high. The highest results are seen for cs↔en and for all sets other than the WMT2022 test set, the lowest are generally seen for cs↔ul, which correspond to the highest and lowest resourced language pairs respectively. Interestingly, the en↔uk test set are comparatively tougher than the other sets we evaluate with.

When we compare bilingual and multilingual results, it is clear that the bilingual models are largely superior for all language directions, with very large differences in BLEU scores across evaluate sets. The only BLEU scores that are higher for the multilingual model is for en↔uk, for which the in-house test set gives slightly higher results. However, this does not hold for the other test sets, indicating overfitting of the models. These results are not so surprising given the relatively small vocabulary size of 64k for the four languages included in training. This is to compare with the bilingual models' 64k vocabulary sizes spread over two languages only. The obligation to share a same vocabulary size amongst more languages (and more scripts) is certainly not compensated by any gain that could possibly be had through multilingual transfer.

6.2 Results with transcription

In Table 4 we provide the results of bilingual and multilingual models with transcription (and detranscription where necessary).

Although transcription should not help the bilingual models that translate to and from English since there is only one Slavic language involved, we include these results for comparative purposes. Ideally, these results (for {cs,uk,ru}↔en) should be identical to the baseline results, showing that transcription does not introduce noise into the process. In reality, we see a systematic drop in results when transcribing for into-English directions, and a greater drop in BLEU score for into-Slavic directions, most likely due to errors introduced by the detranscription model. Interestingly, some directions suffer much more than others (e.g. en→uk and en→ru have a drop of over 2 BLEU vs. en→cs's drop of 0.20 BLEU on WMT2022). This could well be a reflection of the fact that the transcription scheme was centred around Czech, with fewer modifications being made to this language than to the others.

For the multilingual models, the scores are again much lower than the bilingual models with transliteration for all directions, although some slight im-

provements are seen for into-English directions, although the performance is much closer for en→uk. We do however see an improvements across the board on the results of the baseline multilingual models (i.e. without transcription), suggesting that transcribing helps to marginally make up some of the lost scores. Unfortunately, it is unclear whether this is due to the vocabulary now being spread over fewer different scripts or whether transcription does help provide better transfer in some other ways.

7 Discussion

Given these disappointing results, it is important to make a first step to understanding why transcription does not help. We therefore look at some additional results concerning the noise that the transcription step might be introducing: (i) the translation results for the detranscription step itself and (ii) comparative results for cs↔uk when transcribing the source, the target or both.

Detranscription quality We show the results for the detranscription step itself in Table 5, where we apply our detranscription models to the texts to which our transcription rules have been applied. The BLEU scores are very high, but not exactly perfect, suggesting that errors are being introduced in this step. The results are highest for Czech, therefore confirming our earlier hypothesis that this step is degrading less for this language given that fewer changes are made.

We also provide results (Table 6) of the raw output of the from-English bilingual models with transcription (i.e. before applying detranscription). We compare these to the results of the bilingual baselines (trained to produce the correct script) but with automatic transcription applied to the outputs in order to provide a point of comparison in terms of the BLEU score. The results are lower for the bilingual models with transcription for Russian and Ukrainian, suggesting that the outputs of the MT models are also far from perfect, and that transcription may be introducing ambiguities and making it harder for the models to learn. However, as can be seen in previous results, the same cannot be said for Czech, where the results are actually slightly higher for the bilingual model with transcription compared to the bilingual baseline with transcription applied.

Comparative results for cs↔uk with different combinations of transcription Table 7 shows the results for cs↔uk when transcribing the source, target or both. The best results when using the original scripts, as can be seen in previous results. However, the results suggest that in some scenarios, it could be better to transcribe just the source rather than to transcribe both source and target. The advantage of this for uk→cs is that Ukrainian is being made to look more like Czech, but without there being extra errors added by the detranscription step. (Muller et al., 2021) showed that transcribing could be useful for lower-resource languages, so a possibility here is that the languages are sufficiently high-resource for transcription not to help so much and for the errors introduced in detranscription to outweigh any potential benefits.

	cl→cs	**rl→ru**	**ul→uk**
FLORES$_{devtest}$	97.49	94.74	96.29
WMT 2022 (src)	96.47	95.34	94.70
WMT 2022 (ref)	97.33	96.24	97.12

Table 5: BLEU score results for detranscription.

	en→cl	**en→rl**	**en→ul**
Bilingual with transcription			
FLORES$_{devtest}$	29.53	22.90	22.62
WMT 2022	34.06	22.56	19.27
Transcribing bilingual baseline's output			
FLORES$_{devtest}$	29.37	25.35	24.60
WMT 2022	33.87	23.75	20.94

Table 6: Comparison of bilingual models with transcription and of baseline bilingual models with transcription applied to the outputs. Results (BLEU scores) are provided on transliterated references.

	none	source	target	both
cs→uk				
FLORES$_{devtest}$	19.76	19.64	19.32	19.05
WMT 2022	27.40	27.01	26.82	26.43
uk→cs				
FLORES$_{devtest}$	20.86	20.41	18.50	20.09
WMT 2022	25.65	25.15	25.41	24.96

Table 7: BLEU score results for bilingual cs↔uk models when transliterating neither source nor target (none), just the source, just the target or both. Results are shown after detranscription.

8 Conclusion

Setting aside the fact that multilingual models provide very inferior results to specific bilingual models for the same number of parameters, our results

suggest that the answer to the question "Does transcription help cross-script machine translation?" is no. This is at least for the languages on which we experimented and given the amount of training data we had at our disposition. Our bilingual model results show that transcription harms performance, whether it is done on the source side, the target side or both sides. There are several possible explanations for this: (i) the relatively high-resource scenario we are working in, where baselines can already achieve good results and where little gain can be achieved through this type of transfer, (ii) the possibility that transcription introduced ambiguities that could harm translation, and (iii) the detranscription step itself also introducing errors.

Acknowledgements

This work was partly funded by Rachel Bawden's and Benoît Sagot's chairs in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001 and also by the Emergence project, DadaNMT, funded by Sorbonne Université.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Rachel Bawden, Nikolay Bogoychev, Ulrich Germann, Roman Grundkiewicz, Faheem Kirefu, Antonio Valerio Miceli Barone, and Alexandra Birch. 2019. The University of Edinburgh's submissions to the WMT19 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 103–115, Florence, Italy. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Bernardo Stearns, Arun Jayapal, Sridevy S, Mihael Arcan, Manel Zarrouk, and John P McCrae. 2019. Multilingual multimodal machine translation for Dravidian languages utilizing phonetic transcription. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 56–63, Dublin, Ireland. European Association for Machine Translation.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Vikrant Goyal, Sourav Kumar, and Dipti Misra Sharma. 2020. Efficient neural machine translation for low-resource languages via exploiting related languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168, Online. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alex Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomás Mikolov. 2016b. Fasttext.zip: Compressing text classification models. *CoRR*, abs/1612.03651.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol

Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. 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 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2021. *Machine Translation and Transliteration Involving Related and Low-resource Languages*. Chapman and Hall/CRC.

Hairong Liu, Mingbo Ma, Liang Huang, Hao Xiong, and Zhongjun He. 2019. Robust neural machine translation with joint textual and phonetic embedding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3044–3049, Florence, Italy. Association for Computational Linguistics.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2019. Multi-Round transfer learning for Low-Resource NMT using multiple High-Resource languages. *ACM Transactions on Asian Low-Resource Language Information Processing*, 18(4):1–26.

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.

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 301–305, Jeju Island, Korea. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

José Carlos Rosales Núñez, Djamé Seddah, and Guillaume Wisniewski. 2019. Phonetic normalization for machine translation of user generated content. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 407–416, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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.

Simeng Sun, Angela Fan, James Cross, Vishrav Chaudhary, Chau Tran, Philipp Koehn, and Francisco

Guzmán. 2022. Alternative input signals ease transfer in multilingual machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5291–5305, Dublin, Ireland. Association for Computational Linguistics.

Svetlana Tchistiakova, Jesujoba Alabi, Koel Dutta Chowdhury, Sourav Dutta, and Dana Ruiter. 2021. EdinSaar@WMT21: North-germanic low-resource multilingual NMT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 368–375, Online. 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*, volume 30. Curran Associates, Inc.

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 Data sources

Tables 8 and 9 give the amount of data for each data source in the parallel and monolingual data respectively.

Source	en-ru	en-cs	en-uk	cs-uk
AirBaltic	1092			
ECB		3100		
CZECHTOURISM	7328			
RAPID		263287		
EMA		495234		
EESC		1329010		
UNCorpus4	23239280			
NEWS Commentary	333899	253639		
WorldBank	25849		1628	
Paracrawl	5377911	50632492	13354365	
WikiTitles	1189107	410978		
WikiMatrix		2094650		
EUROPARL		645330		
Commoncrawl	878386	161838		
Opus				
Bible			15901	7953
Open Subtitles			877780	730804
EUBooks			1793	1506
TED2020			208141	115351
Wikimedia			348143	1959
MultiCCAligned			8547349	2306396
Dev set				
FLORES (dev)	997		997	
FLORES (devtest)	1012		1012	
NEWSTEST2018	991			

Table 8: Parallel data sources.

Source	Cs	En	Ru	Uk
News crawl	12203274	39361312	15441304	411439
Europarl v10	669676			
News Commentary	282139	660667	404978	
Common Crawl	333498145	-	1168529851	
UberText Corpus				-
fiction				1811548
news				31021650
ubercorpus				48620146
wikidump				15786948
Leipzig Corpora	-	-	-	-
ukr_mixed_2012				1000000
ukr_news_2020				1000000
ukr_newscrawl_2018				1000000
ukrua_web_2019				1000000
ukr_wikipedia_2021				1000000
Legal Ukrainian				7568246
Common Crawl (filt.)	275825036	-	1150607428	-
Total (concat.)	293980125	-	1170453710	110219977
Total (dedup.)	290477308	-	1155825622	53177077

Table 9: Monolingual data sources.