Transliteration for Low-Resource Code-Switching Texts: Building an Automatic Cyrillic-to-Latin Converter for Tatar

Chihiro Taguchi[*], **Yusuke Sakai**[*], and **Taro Watanabe**
{taguchi.chihiro.td0, sakai.yusuke.sr9, taro}@is.naist.jp
Nara Institute of Science and Technology

Abstract

We introduce a Cyrillic-to-Latin transliterator for the Tatar language based on subword-level language identification. The transliteration is a challenging task due to the following two reasons. First, because modern Tatar texts often contain intra-word code-switching to Russian, a different transliteration set of rules needs to be applied to each morpheme depending on the language, which necessitates morpheme-level language identification. Second, the fact that Tatar is a low-resource language, with most of the texts in Cyrillic, makes it difficult to prepare a sufficient dataset. Given this situation, we proposed a transliteration method based on subword-level language identification. We trained a language classifier with monolingual Tatar and Russian texts, and applied different transliteration rules in accord with the identified language. The results demonstrate that our proposed method outscores other Tatar transliteration tools, and imply that it correctly transcribes Russian loanwords to some extent.

1 Introduction

Modern Tatar has two orthographies: Cyrillic and Latin. The two alphabets are mostly mutually compatible when an input string consists of only Tatar-origin words. Effectively, however, modern Tatar has a massive amount of Russian loanwords, and, in colloquial texts, even a whole phrase may be switched to Russian. This linguistic phenomenon is known as code-switching or code-mixing.

A difficulty of transliteration from Cyrillic to Latin lies in the following two facts. First, a different set of transliteration rules has to be applied to Tatar and Russian words. This requires a language detection for each token, or worse, for each morpheme, which would additionally require morphological analysis. It is expected that a full implementation of such a system will produce heavy processes. Second, because modern Tatar frequently mixes Russian words, it is not easy to obtain a pure Tatar dataset for developing a language detector.

Existing methods are based on either Tatar monolingual rules or a huge bundle of ad-hoc rules aimed to cover Russian-origin words (Bradley, 2014; Korbanov, n.d.). The experimental results in Section 6 demonstrate that the former monolingual rule-based transliterators show low accuracy because Russian words are not supported. The latter extensively rule-based transliterator has better accuracy, but still misses a certain amount of words. This implies that a strictly rule-based method requires an ever-lasting process of adding rules ad hoc for exceptional words to further improve the accuracy. This is obviously unrealistic and inefficient.

In this study, in contrast, we pursue a simple yet high-accuracy automatic transliteration system from Cyrillic Tatar to Latin Tatar. We prepare two sets of simple rule-based monolingual transliteration rules for Tatar and Russian each. In addition, we train a classifier that automatically determines Tatar or Russian for each subword. Each token is then transliterated to Latin Tatar by applying the rules of the detected language to its subwords. The results demonstrate that our proposed method achieves higher accuracy than the previous tools. Also, our proposed method demonstrates higher accuracy than the transliterator with only Tatar-based rules, indicating that the method can correctly predict and transcribe Russian words to some extent.

2 Tatar: Linguistic Background

Tatar (ISO 639-1 Code: tt) is a Kipchak language of the Turkic language family mainly spoken in the Republic of Tatarstan, Russia. The number of the speakers is estimated to be around five million (Eberhard et al., 2021). The canonical word order is SOV, but allows for free word order with scrambling. Morphological inflections and declensions are derived by means of suffixation. The suffixation obeys the vowel harmony rule where backness

[*] Equal contribution

(or frontness) of vowels is kept consistent.

Through the long history of language contact with Russian, modern Tatar contains a large amount of Russian loanwords. Most of the speakers are bilingual along with Russian, and younger generations living in urbanized cities tend to have better competence in Russian than in Tatar. This bilingualism leads to frequent code-switching (CS) in text and speech, particularly of colloquial genre (Izmailova et al., 2018).

2.1 Tatar Orthographies

The mainstream orthography for modern Tatar is the Cyrillic alphabet, which comprises the Russian standard alphabet and six extended Cyrillic letters. Cyrillic Tatar is mostly used by Tatars living in Russian-speaking countries, while the Latin alphabet is used by Tatars living in other areas such as Turkey and Finland.

Until 1928, Tatar was exclusively written in the Arabic script. Along with the latinization project prevailing in the Soviet Union in 1920–30s, a Latin alphabet system *yañalif* was introduced to Tatar. In 1939, for political reasons, *yañalif* was superseded by the Cyrillic alphabet which has been officially used in Tatarstan as of now. After the demise of the Soviet Union, with the resurgence of the movement for restoring the Latin orthography, a new Latin-based orthography system was adopted by a republic law in 1999 (National Council of the Republic of Tatarstan, 1999). However, the law soon lost its validity in 2002 when a new paragraph stipulating that all the ethnic languages in Russia must be written in Cyrillic was added to the federal law (Yeltsin, 2020). The current Latin alphabet (2013Latin henceforth) based on the Common Turkic Alphabet was officially adopted by a republic law in 2013, and is commonly used in Tatar diaspora communities (National Council of the Republic of Tatarstan, 2013).

We define that the term “Latin alphabet” used in this paper refers to 2013Latin. A detailed rules for the conversion to 2013Latin is given in Timerkhanov and Safiullina (2019).

2.2 Code-switching in Tatar

An example of the former is displayed in (1) with transliteration in the Latin alphabet and the translation. Underlined класс (*klass* “class, grade”) is a Russian word naturalized in Tatar, though it is pronounced with Russian phonology and therefore requires a different transliteration. In addition, a locative suffix -та (*-ta* “at, in”) is attached in the example, as Russian loanwords may take Tatar suffixes, causing CS within a token (intra-word CS).

(1) Безнең класста кызлар сигезенчедән эчә башлаган иде.
translit.: Bezneñ klassta qızlar sigezençedän eçä başlağan ide.
“In our class, girls used to start drinking by the eighth grade.”

3 Related Work

Code-Switching. Even though CS has attracted researchers in NLP, the lack of resource has been a major difficulty, because CS is an exclusively colloquial linguistic phenomenon and CS texts are seldom recorded. Jose et al. (2020) enumerates a list of available CS datasets at the time of the publication. In terms of both the availability of datasets and the popularity of research, CS language pairs in trend are Hindi–English (Srivastava et al., 2020, Singh and Lefever, 2020), Spanish–English (Alvarez-Mellado, 2020, Claeser et al., 2018), Arabic varieties and Modern Standard Arabic (Hamed et al., 2019).

As for the studies of intra-word CS in other languages, Mager et al. (2019) for German–Turkish and Spanish–Wixarika, Nguyen and Cornips (2016) for Dutch–Limburgish, and Yirmibeşoğlu and Eryiğit (2018) for Turkish–English have a similar approach to ours. The differences from ours are that Mager et al. (2019) employs segRNN (Lu et al., 2016) for segmentation and language identification, and that Nguyen and Cornips (2016) uses Morfessor (Creutz and Lagus, 2006) for morphological segmentation. However, our task that combines language detection of intra-word CS and transliteration has never been undertaken in any of these studies.

Tatar Transliteration. At the time of this writing, the following tools are available for Tatar Cyrillic-Latin conversion. The Tatar Transcription Tool (TTT henceforth) (Bradley, 2014) is a transliterator published online by Universität Wien as a part of the Mari Web Project. speak.tatar[1] is an anonymously developed transliteration service. FinTat[2] is a transliteration tool developed as a part of the Corpus of Written Tatar (Saykhunov et al., 2019). Aylandirow is a strictly rule-based transliteration tool

[1] https://speak.tatar/en/language/converter/tat/cyrillic/latin
[2] http://www.corpus.tatar/fintat

available online that extensively covers Russian-origin words as well as Tatar-origin (Korbanov, n.d.). The transliteration system employed in FinTat is based on the Latin alphabet used by Tatars in Finland, whose orthography is somewhat different from 2013Latin.

4 Method

We transliterate Cyrillic Tatar to Latin Tatar word by word, as each word does not affect other words in Tatar transliteration[3].

Taking into account the fact that Tatar has intra-word CS, we created a classifier that detects a language (Tatar or Russian) for each subword. To implement the language classifier, we prepared two monolingual corpora of Tatar and Russian. Given the lack of pure Tatar texts without CS in modern texts[4], we employed Tatar translation of Qur'an[5] (19,691 words with duplication) translated in 1912 that contains no Russian loanwords in order to avoid noise to train the classifier. Its Russian counterpart[6] (21,256 words with duplication) was translated by the Ministry of Awqaf, Egypt.

The training process is as follows. First, the words collected from the dataset were automatically divided into subwords by the Byte Pair Encoding algorithm (Sennrich et al., 2016). Banerjee and Bhattacharyya (2018) reports that, unlike Morfessor, BPE can flexibly solve the OOV problem because some subwords are character-level segments. In our case, due to the meagerness of the monolingual training data, we employed BPE to avoid the OOV problem. Then, assuming that longer subwords are less ambiguous with respect to labels to be assigned, we took the longest match to make it easier to distinguish between Russian and Tatar; for this reason, the subword merge operation was repeated until no further merge was possible. The obtained subwords are then represented in subword embeddings using fastText[7] (Bojanowski et al., 2017). The classification model is the supervised classifier provided by fastText, with the following hyperparameters that are known to perform high accuracy[8]: the number of dimensions is 16, the minimum and maximum character n-gram sizes are 2 and 4. Hierarchical softmax is used as the loss function. Together with this representation, the subword embeddings are annotated with a language label (Joulin et al., 2017).

It is worth noting that, unlike Mager et al. (2019), we do not employ deep learning approach. While the task in Mager et al. (2019) is multi-labeling, our language identification task is a binary classification with a low-resource training dataset; for this reason, deep learning is too superfluous and heavy for achieving our task.

To evaluate the performance of our model, we apply BPE also to the test data in the same manner, and predict a language label for each subword. Adjacent subwords are, when possible, combined to form a longer subword with a single language label for the sake of better accuracy; that is, for example, when two consecutive subwords are both labeled as `tt`, then they are combined into one subword labeled as `tt`. Finally, each subword is converted to the Latin alphabet with the transliteration rules of the predicted language, and combined into a word as an output.

5 Experimental Setup

For the evaluation of the performance, we prepared 700 sentences (shuffled; 8,466 words with duplication, 5,261 without duplication) from the Corpus of Written Tatar (Saykhunov et al., 2019) and their Latin counterpart as the gold data that was manually transcribed by us and verified by a native speaker. Also, we annotated the Cyrillic text data so that CS Russian morphemes are tagged. According to this, the text data contains 1,009 words (with duplication) with a Russian morpheme, and 598 words (with duplication) with intra-word CS.

For evaluation metrics, we calculated BLEU and longest common sequence (LCS) F-measure for each letter in a word as well as word accuracy (ACC) and character error rate (CER). The calculation of LCS F-measure and ACC is based on Chen et al. (2018).

To compare the performances of Tatar monolingual transliteration and of Tatar–Russian bilingual transliteration (i.e., our proposed method; "tt-ru hybrid" henceforth), we also evaluated the data with solely Tatar monolingual transliteration rules

[3] The code is available here: https://github.com/naist-nlp/tatar_transliteration. A demonstration page is published on the website: https://yusuke1997.com/tatar.

[4] In the evaluation dataset we prepared for this study, 1,009 words out of 8,466 contained at least one Russian morpheme.

[5] https://cdn.jsdelivr.net/gh/fawazahmed0/quran-api@1/editions/tat-yakubibnnugman.json

[6] https://cdn.jsdelivr.net/gh/fawazahmed0/quran-api@1/editions/rus-ministryofawqaf.json

[7] https://github.com/facebookresearch/fastText

[8] The training setting was inspired by the blog post by fastText: https://fasttext.cc/blog/2017/10/02/blog-post.html

	BLEU	LCS F-score	CER	ACC	# correct sentence	# error word
speak.tatar	0.869	0.953	0.049	0.952	67	1,747
TTT	0.879	0.956	0.054	0.946	121	1,505
Aylandirow	0.971	**0.994**	0.009	0.991	362	526
tt-based	0.968	0.989	0.011	0.989	365	552
tt-ru hybrid	**0.981**	**0.994**	**0.007**	**0.993**	**437**	**332**

Table 1: The experimental results with 700 sentences (5,261 words without duplication). CER and ACC are complementary in probability (i.e., $ACC = 1 - CER$). Note that the transliteration result is uniquely determined as the proposed method is rule-based.

("tt-based" henceforth). For the comparison with the existing transliteration tools, we computed the scores of speak.tatar, TTT, and Aylandirow[9].

6 Results

As shown in Table 1, the experimental results demonstrate that our tt-ru hybrid marked the best score in all the metrics. CER is the lowest, meaning that the total number of mistakes at a character level is the fewest. The difference is evident between monolingual transliterators (in particular, speak.tatar and TTT) that do not support Russian loanwords and bilingual transliterators (Aylandirow and tt-ru hybrid). Between the two groups, there is more or less a difference by 0.1 points in their BLEU scores. Furthermore, our monolingual tt-based marked higher scores than the other two monolingual transliterators. A possible explanation to this gap will be given in the next section.

Compared to Aylandirow, an extensive rule-based transliterator, CER (i.e., ACC also) in tt-ru hybrid was slightly better by 0.002 points. In LCS F-Score, in contrast, Aylandirow has the same score as our tt-ru hybrid. This fact means that Aylandirow returned transliterations somewhat closer to the gold data than our tt-ru hybrid.

Note that the perfection is not necessarily the ultimate goal of this transliteration. As described in detail in Section 7, there may be several ways of spelling in actual language use, particularly of the spelling of proper nouns. This variance in spelling foreign words is common among languages.

7 Analysis

The results illustrated that the bilingual transliterators generally have higher accuracy than the monolingual transliterators. However, it needs to be examined that tt-based also gained higher scores than the other monolingual ones, even though it does not support Russian loanwords. This is due to the fact that their transliteration rule sometimes does not follow the rules of 2013Latin. For example, тәнкыйть "criticism" is transliterated as *tänqit'* (with a *hamza* at the end) by the TTT, where it should be transliterated as *tänqit* according to 2013Latin. In fact, due to the de-facto absence of any institution that regulates the Latin orthography, different varieties in spelling styles can be observed on the Internet. For this reason, the seemingly high scores in tt-based are merely a product of the orthographical consistency.

Keeping this in mind, the improvement in the scores between tt-based and tt-ru hybrid indicates that the bilingual transliteration method designed to be adaptive to Russian loanwords successfully predicted the language and transcribed them.

Table 3 illustrates the number of error words categorized with respect to the transliteration rules (tt-based and tt-ru hybrid). We can see from the table that, among the error words observed in tt-based ($V(T)$), 336 words were correctly transliterated by tt-ru hybrid.

Examples of successful transliteration in our tt-ru hybrid are закон and совет shown in the upper example of Table 2. The tt-based transliteration converted them as *zaqon* and *sowet*, while tt-ru hybrid correctly returned *zakon* and *sovet*. This shows that tt-ru hybrid successfully identified the language, since they are Russian loanwords.

However, tt-ru hybrid wrongly transliterated 116 words that were correct in tt-based; for example, in the lower example of Table 2, the underlined transliteration *soklanıp* in tt-ru hybrid is a transliteration error, where the correct word form is *soqlanıp*. Because the language classifier identified

[9] Although we mentioned FinTat in Section 2, we excluded it from the evaluation because it is based on a spelling system slightly different from 2013Latin.

Source	РФ Закон чыгаручылар советы президиумы утырышында катнашты.
tt-based	RF Zaqon çığaruçılar sowetı prezidiumı utırışında qatnaştı.
tt-ru hybrid	RF Zakon çığaruçılar sovetı prezidiumı utırışında qatnaştı.
Source	Һәр юлчы туктап, аның хозурлыгына сокланып китә.
tt-based	Här yulçı tuqtap, anıñ xozurlığına soqlanıp kitä.
tt-ru hybrid	Här yulçı tuqtap, anıñ xozurlığına soklanıp kitä.

Table 2: Examples of successful and unsuccessful transliterations in tt-ru hybrid based on subword tokenization. The underlined words are error words (Russian loanwords). The upper sentence is an example where tt-ru hybrid can correctly transliterate the underlined words while tt-based cannot. The lower example is, on the other hand, tt-ru hybrid mistakenly transliterates the underlined word that is correctly transliterated by tt-based.

set (total 5,261 words)	# words
$V(T)$	552
$V(H)$	332
$V(T) \cap V(H)$	216
$V(T) \setminus V(H)$	336
$V(H) \setminus V(T)$	116

Table 3: Comparison of the number of error words (without duplication) between the monolingual tt-based transliteration and our proposed tt-ru hybrid transliteration. $V(T)$ is the set of error words observed in tt-based, $V(H)$ in tt-ru hybrid, $V(T) \cap V(H)$ in both tt-based and tt-ru hybrid, $V(T) \setminus V(H)$ is the set of error words observed only in tt-based, and $V(H) \setminus V(T)$ only in tt-ru hybrid.

	ru words		CS words	
	#	accuracy	#	accuracy
speak.tatar	**805**	**0.798**	455	0.752
TTT	258	0.256	175	0.283
Aylandirow	738	0.731	461	0.762
tt-based	471	0.467	294	0.486
tt-ru hybrid	788	0.781	**464**	**0.767**

Table 4: A comparison of performance in Russian CS words. The left column (ru words) demonstrates the number of correctly transcribed words that contain Russian and its accuracy that is given by dividing by the total Russian words (1,009 with duplication). The right column (CS words) contains the number of correct transcriptions out of 605 words (with duplication) and its accuracy with respect to intra-word CS words.

the first subword as Russian, the Russian transliteration rule was applied, whereas in fact it is not a Russian loanword.

As for the high LCS F-score in Aylandirow (0.994), it implies that Aylandirow is good at correctly transcribing frequent words including Russian loanwords. Because Aylandirow is strictly rule-based without automatic language detection, it can easily suffer from the rule coverage problem; for example, the word такси (*taksi*, a Russian loanword) was mistakenly transliterated as *taqsi*.

As the comparison of performance with respect to Russian CS words in Table 4 shows, tt-ru hybrid demonstrated higher accuracy in transliterating words with Russian morphemes[10]. In particular, the tt-ru hybrid's performance adaptive to Russian morphemes is clearly visible in the accuracies of transcribing words containing Russian, where tt-ru hybrid scored 78.1% and tt-based 46.7%.

Considering that Russian CS in Tatar may arbitrarily occur, our proposed method with automatic language detection is expected to show a stable performance to any Russian words; in contrast, rule-based systems such as Aylandirow are less flexible to unknown Russian words, which, in effect, exist infinitely in natural languages as hapax legomena.

8 Conclusion

In this paper, we proposed a new transliteration system that converts Cyrillic Tatar to Latin Tatar. Taking into account the facts that different transliteration rules are applied to Russian loanwords and that intra-word CS is frequently observed, our proposed method involved language identification for each subword. Even though Tatar resources available at hand for training were limited, the results were significantly better than existing transliteration tools. The simple architecture of language detection employed in this approach is language-agnostic does not need detailed analyses such as syntactic parsing and POS tagging, our method is applicable to other low-resource languages that have intra-word CS.

[10] In this respect, speak.tatar scores the best for Russian words. This is merely because its transliteration rules are designed to work well for Russian words, and, in contrast, its performance to Tatar words is poor as shown in Table 1.

References

Elena Alvarez-Mellado. 2020. An annotated corpus of emerging anglicisms in Spanish newspaper headlines. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 1–8, Marseille, France. European Language Resources Association.

Tamali Banerjee and Pushpak Bhattacharyya. 2018. Meaningless yet meaningful: Morphology grounded subword-level NMT. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 55–60, New Orleans. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jeremy Bradley. 2014. Tatar transcription tool. Available at: https://www.univie.ac.at/maridict/site-2014 [retrieved 28 March 2021].

Nancy Chen, Rafael E. Banchs, Min Zhang, Xiangyu Duan, and Haizhou Li. 2018. Report of NEWS 2018 named entity transliteration shared task. In *Proceedings of the Seventh Named Entities Workshop*, pages 55–73, Melbourne, Australia. Association for Computational Linguistics.

Daniel Claeser, Samantha Kent, and Dennis Felske. 2018. Multilingual named entity recognition on Spanish-English code-switched tweets using support vector machines. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 132–137, Melbourne, Australia. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2006. Morfessor in the morpho challenge. In *PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2021. *Ethnologue*, 24th edition. SIL International, Dallas, Texas. Available at: http://www.ethnologue.com [retrieved 12 March 2021].

Injy Hamed, Moritz Zhu, Mohamed Elmahdy, Slim Abdennadher, and Ngoc Thang Vu. 2019. Code-switching language modeling with bilingual word embeddings: A case study for egyptian arabic-english.

Guzel A. Izmailova, Irina V. Korovina, and Elzara V. Gafiyatova. 2018. A study on tatar–russian code switching (based on instagram posts). *The Journal of Social Sciences Research*, Special Issue. 1:187–191.

N. Jose, B. R. Chakravarthi, S. Suryawanshi, E. Sherly, and J. P. McCrae. 2020. A survey of current datasets for code-switching research. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 136–141.

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.

Dinar Korbanov. n.d. Aylandirow. Available at: http://aylandirow.tmf.org.ru [retrieved 29 March 2021].

Liang Lu, Lingpeng Kong, Chris Dyer, Noah A. Smith, and Steve Renals. 2016. Segmental recurrent neural networks for end-to-end speech recognition. In *Proceedings of Interspeech 2016*, Interspeech, pages 385–389. International Speech Communication Association.

Manuel Mager, Özlem Çetinoglu, and Katharina Kann. 2019. Subword-level language identification for intra-word code-switching. *CoRR*, abs/1904.01989.

National Council of the Republic of Tatarstan. 1999. О восстановлении татарского алфавита на основе латинской графики [on the restoration of the tatar alphabet based on the latin script]. Available at: http://docs.cntd.ru/document/917005056 [retrieved 28 March 2021].

National Council of the Republic of Tatarstan. 2013. Об использовании татарского языка как государственного языка Республики Татарстан [on the use of the tatar language as the national language of the republic of tatarstan]. Available at: http://docs.cntd.ru/document/463300868 [retrieved April 26, 2021].

Dong Nguyen and Leonie Cornips. 2016. Automatic detection of intra-word code-switching. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 82–86, Berlin, Germany. Association for Computational Linguistics.

Mansur R. Saykhunov, R. R. Khusainov, T. I. Ibragimov, J. Luutonen, I. F. Salimzyanov, G. Y. Shaydullina, and A. M. Khusainova. 2019. Corpus of written tatar. Available at: http://www.corpus.tatar [retrieved 28 March 2021].

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.

Pranaydeep Singh and Els Lefever. 2020. Sentiment analysis for Hinglish code-mixed tweets by means of cross-lingual word embeddings. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 45–51, Marseille, France. European Language Resources Association.

Abhishek Srivastava, Kalika Bali, and Monojit Choudhury. 2020. Understanding script-mixing: A case study of Hindi-English bilingual Twitter users. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 36–44, Marseille, France. European Language Resources Association.

Aynur Akhatovich Timerkhanov and Gulshat Rafailevna Safiullina. 2019. *Tatarça-inglizçä, inglizçä-tatarça süzlek: Tatar-English, English-Tatar dictionary*. G. Ibragimov Institute of Language, Literature and Art, Kazan.

Boris Yeltsin. 2020. О языках народов Российской Федерации [on the languages of nations in the russian federation]. First published in 1991, last amended in 2020. Available at: http://pravo.gov.ru.

Zeynep Yirmibeşoğlu and Gülşen Eryiğit. 2018. Detecting code-switching between Turkish-English language pair. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 110–115, Brussels, Belgium. Association for Computational Linguistics.

A Appendix

Cyrillic	Latin (Tatar)	Latin (Russian)
а	a	a
б	b	b
в	w	v
г	g, ğ	g
д	d	d
е	e, ye, yı	e
ё	NA	yo
ж	j	j
з	z	z
и	i	i
й	y	y
к	k, q	k
л	l	l
м	m	m
н	n	n
о	o	o
п	p	p
с	s	s
т	t	t
у	u, uw, w	u
ф	f	f
х	x	x
ц	NA	ts
ч	ç	ç
ш	ş	ş
щ	NA	şç
ъ	—	—
ы	ı	ı
ь	—	—
э	e, ’	e
ю	yu, yü, yuw, yüw	yu
я	ya, yä	ya
ә	ä	NA
ө	ö	NA
ү	ü, üw, w	NA
җ	c	NA
ң	ñ	NA
һ	h	NA

Table 5: Tatar’s Cyrillic–Latin correspondence for Tatar- and Russian-origin words. NA (not applicable) means that the letter does not appear in the language. An em-dash means that the letter is ignored in Latin transcription.

original	РФ Президенты Владимир Путин Россия мөселманнарын изге Рамазан
gold (Latin)	RF Prezidentı Vladimir Putin Rossiyä möselmannarın izge Ramazan
speak.tatar	RF Prezidentı Vladimir Putin Rossiyä möselmannarın izge Ramazan
TTT	RF Prezidentı Wlädimir Pütin Rössiyä möselmannarın izge Ramazan
Aylandirow	RF Prezidentı Vladimir Putin Rossiä möselmännarın izge Ramazan
tt-based	RF Prezidentı Wladimir Putin Rossiyä möselmannarın izge Ramazan
tt-ru hybrid	RF Prezidentı Vladimir Putin Rossiyä möselmannarın izge Ramazan

Table 6: An example of comparison of transliterations. The sequence on the top is the original corpus sentence in Cyrillic, below which is the Latin counterpart manually transcribed. The three sentences in the middle row are transliterations by the previous tools. The first sentence in the bottom row is the monolingual tt-based transliteration, and the second is transcribed by our proposed method.