ParsText: A Digraphic Corpus for Tajik-Farsi Transliteration

Rayyan Merchant*, **Kevin Tang***
*University of Florida
Department of Linguistics, College of Liberal Arts and Sciences
rayyan.merchant@gmail.com

Heinrich Heine University Düsseldorf
Department of English Language and Linguistics, Faculty of Arts and Humanities
kevin.tang@hhu.de

Abstract

Despite speaking dialects of the same language, Persian speakers from Tajikistan cannot read Persian texts from Iran and Afghanistan. This is due to the fact that Tajik Persian is written in the Tajik-Cyrillic script, while Iranian and Afghan Persian are written in the Perso-Arabic script. As the formal registers of these dialects all maintain high levels of mutual intelligibility with each other, machine transliteration has been proposed as a more practical and appropriate solution than machine translation. Unfortunately, Persian texts written in both scripts are much more common in print in Tajikistan than online. This paper introduces a novel corpus meant to remedy that gap: ParsText. ParsText contains 2,813 Persian sentences written in both Tajik-Cyrillic and Perso-Arabic manually collected from blog pages and news articles online. This paper presents the need for such a corpus, previous and related work, data collection and alignment procedures, corpus statistics, and discusses directions for future work.

Keywords: parallel text, Persian, Tajik, Farsi, orthography, transliteration, Cyrillic, Perso-Arabic

1. Introduction

ParsText is a new digraphic Persian corpus created for the express purpose of transliteration between two Persian dialects and their scripts: Tajik-Cyrillic in Tajikistan and Perso-Arabic in Iran and Afghanistan. The corpus consists of 2,813 sentences, with average Tajik-Cyrillic and Perso-Arabic sentence lengths of 15.00 and 15.57 words, respectively.

To the best of our knowledge, only two previous efforts have investigated machine transliteration between these two scripts thus far, and both of them lacked parallel corpora with which to directly evaluate their models (Davis, 2012; Megerdoomian and Parvaz, 2008). While digraphic texts are available within Tajikistan in print form, similar texts rarely make appearances online, even on the website of Tajikistan's embassy in Iran.[1] ParsText fills this gap as a corpus made up of blog posts and news articles written by native Persian speakers in both scripts. The goal of ParsText is to enable future efforts to train or evaluate their transliteration systems. In an independent study (under review), we use ParsText to train Tajik-Farsi transliteration models. This data will be made available on OSF[2] and Github[3].

In Section 2, the importance of Tajik-Farsi transliteration and why ParsText, a digraphic parallel-text corpus at the sentence level, is preferable to lists with word pairs in isolation is discussed. Section 3 introduces previous and related work. Section 4 describes how the corpus was developed. Section 5 provides corpus statistics and observations. Finally, Section 6 concludes the paper.

2. Background

2.1. Motivation

Tajik Persian (henceforth, Tajik) is the formal register of Modern Persian spoken in Tajikistan. While spoken Tajik has evolved separately for centuries, the formal register retains extremely high levels of mutual intelligibility with the formal Persian of Iran and Afghanistan (both henceforth referred to as Farsi) (Perry, 2005). Unlike these two countries which use the traditional Perso-Arabic script, Tajikistan uses the relatively new Tajik-Cyrillic script due to its Soviet heritage. Proposals have been made to shift Tajik back to the Perso-Arabic script, but any significant shift will likely not occur soon as Tajikistan's former Minister of Culture stated in 2008 that "...some 90-95% of Tajikistan's population is not familiar with Arabic script..." (Ghufronov, 2008). As a result, the vast majority of the 10 million Persian speakers in Tajikistan cannot read written Persian media produced by the 100 million Persian speakers in Iran and Afghanistan. This restriction extends to the Internet, where Farsi dominates. For example, as of September 2023, the Tajik Wikipedia had 269,857 articles and 10.5 million

[1] https://mfa.tj/tg/tehran

[2] https://doi.org/10.17605/OSF.IO/37GZX

[3] https://github.com/merchantrayyan/ParsText

Second Workshop on Computation and Written Language (CAWL 2024) @LREC-COLING-2024, pages 1–7
May 21, 2024.

words across all content pages compared to the Farsi Wikipedia's 5.5 million articles and 194 million words (Wikimedia Foundation, 2023b). These two scripts are highly incongruous (Perry, 2005). The Perso-Arabic script, as an impure abjad, often omits vowels, and those that are written are ambiguous. Meanwhile, the Tajik-Cyrillic script, as an alphabet, writes out all vowels, making it a better phonetic representation of the language than the Perso-Arabic script. Table 1 illustrates how the same sentence is represented in both scripts, with a Latin transliteration and an English translation provided for clarity.

Script	**Sentence**
Farsi (Perso-Arabic)	'زبان فارسی'
Tajik (Tajik-Cyrillic)	'забони форсӣ'
Latin Translit.	'zaboni forsī'
English Translation	'The Persian language'

Table 1: Example sentence written in Farsi and Tajik with Latin transliteration and English translation

2.2. Challenges with Typical Tajik-Farsi Parallel Corpora

Although Tajik and Farsi descend from a common root, they have nonetheless diverged in several aspects, including grammar, lexicon, and pronunciation (Perry, 2005). As a result, Tajik and Farsi versions of the same text made in isolation from each other, such as the United Nations Declaration of Human Rights, are often quite divergent (Gacek, 2015). As a result, they do not align on a word-to-word basis and cannot be used for the task of transliteration, the conversion of text in one script to another. Additionally, several discrepancies between Tajik and Farsi mean that any transliteration system must take into account features at the interword level.

The Persian 'Ezafe' is one example of such an interword feature, as a grammatical feature that links a modifier to a preceding head noun (or preceding modifier) (Perry, 2005). In accordance with its phonetic nature, the Tajik standard typically writes the Ezafe as an -и attached to the previous word. In contrast, the Perso-Arabic script often omits the Ezafe, so the reader must infer its location from the surrounding context. Typically, the Ezafe is only written when added to the plural marker ها ('ho'). Otherwise, usually if needed to disambiguate a phrase, it is written as a diacritic at the end of the head noun.

As the Ezafe has the potential to drastically change a sentence's phrasal boundaries, and thus its meaning, detection of the Ezafe is an important step in a Natural Language Processing pipeline for Persian (Asghari et al., 2014; Doostmohammadi et al., 2020).

On a basic level, several affixes are always attached to the stem word in Tajik, but written either separately or conjoined with a Zero Width Non-joiner character (ZWNJ) in Farsi (Megerdoomian and Parvaz, 2008). These discrepancies are what make the transliteration task challenging. Further challenges are described in Appendix A.

3. Related Work

Previous investigations of Tajik-Farsi transliteration systems have made use of non-digraphic datasets, resulting in indirect methods of system evaluation. Megerdoomian and Parvaz (2008) created a Tajik-only dataset from the news site Radio Ozodi, judging performance of Tajik to Farsi transliteration through detection of correctly-spelled Farsi words. Davis (2012) utilized a Tajik-Farsi word list of 3,503 pairs as training data for a statistical transliteration system. To evaluate said system, two unrelated Tajik and Farsi datasets were used. None of these datasets or transliteration systems have been made public.

To the authors' knowledge, only one other dataset has been made publicly available for the same purpose as ParsText: the training data released by Github user *stibiumghost*[4] for a Tajik-to-Farsi transliteration project[5] based on work by Talafha et al. (2021). These data were uploaded to Github on December 2022, well after the ParsText corpus was created in February 2022.

This dataset consists of a 43,535 word dictionary and collection of poetry and news with 404,755 Farsi tokens and 392,562 Tajik tokens. We note that in-depth exploration of this corpus and comparison to our own present avenues for further research. We also believe that our corpus, despite its smaller size, would be appropriate for use as an evaluation dataset in combination with the larger *stibiumghost* dataset.

Beyond Persian, there exist several datasets made for machine transliteration of a single language. For Jordanian Arabic, Talafha et al. (2021) created a dataset in Arabic and a non-standard romanization known as Arabizi. Ahmadi et al. (2022) compiled a corpus of Kurdish news articles written in the Sorani (Arabic-based) and Kurmanji (Latin-based) orthographies. More recently, Gow-Smith et al. (2022) reconstructed part of a 16th-century Scottish Gaelic manuscript in modern orthography. These corpora all focus on low-resource lan-

[4] https://github.com/stibiumghost/tajik-to-persian-transliteration/tree/main/training_data

[5] https://github.com/stibiumghost/tajik-to-persian-transliteration

guages and tackle similar challenges in transliteration due to non-phonetic orthographies.

4. The Corpus

4.1. Data Collection

After an extensive online search, two main sources of parallel data presented themselves: blog pages and British Broadcasting Corporation (BBC) News articles. The two blogs we found were written by native Persian speakers who knew both orthographies and dealt with a wide variety of topics ranging from poetry to politics.[6][7] Latin orthographies for Persian, such as Dabire, were not considered as they are not standard in any Persian-speaking country (Maleki, 2008). These blogs and articles, as opposed to individual word lists, provide interword details such as the aforementioned affixes and ezafe which are critical to Tajik-Farsi transliteration. Moreover, they deal with a variety of formal topics, and are therefore written in a formal register of Persian.

To filter out posts that lacked such sentence alignment, as well as those written in only one script or in other languages (usually Russian), we opted to manually collect these data rather than use an automatic website scraping tool.

We were also able to find 23 BBC News articles written in both orthographies[8][9][10] during the time BBC Tajik operated from 1993 to 2015 (BBC, 2015). These articles almost exclusively deal with politics, and exhibited a similar degree of word-to-word alignment. Due to the small number of articles, we decided to collect these manually.

As the first author is a non-native speaker of Persian, he conducted manual inspection of texts for word-to-word alignment during collection, along with spot checking at later points. In this manner, texts that did not meet this standard were filtered out as well. A few sample sentences from ParsText are available in Appendix B.

4.2. Data Processing

As the corpus that we compiled was not aligned on a sentence-to-sentence basis, we aligned each individual source document with GaChalign[11], a Python implementation of the Gale-Church alignment algorithm (Tan and Bond, 2014; Gale and Church, 1993). We note that our corpus presents some inconsistencies due to differences in the authors' word choice, and has not undergone in-depth analysis from native speakers to be corrected. Experimentation with the data uploaded on Github by *stibiumghost* revealed that those present similar inconsistencies. Creation of a digraphic corpus rigorously checked by native Persian speakers therefore presents another avenue for further research.

5. Statistics and Observations

In the absence of lemmatization tools for Tajik, token - rather than lemma - statistics of the corpus are presented in this paper. Table 2 lists corpus statistics, while Table 3 provides the top ten most frequent tokens in both scripts.

Statistics	**Farsi**	**Tajik**
# of sentences	2,813	2,813
# of word tokens	43,846	42,226
# of characters	186,414	222,986
Avg. # of tokens in a sentence	15.57	15.00
Avg. # of characters in a token	66.15	79.13

Table 2: ParsText Statistics. Note that any character statistic does not include whitespace characters.

In accordance with the fact that Farsi does not have a phonetic orthography, the Farsi character statistics are lower than the Tajik character statistics. However, the token measures are larger, likely reflecting how several Persian affixes and function words are written attached to the preceding word in Tajik, but separately in Farsi.

From Table 3, several observations can be made. First, the top 10 most frequent tokens in Tajik and Farsi are the exact same Persian words. Furthermore, the order of these tokens is also mostly shared with the exceptions of ва / و (English: 'and') and аст / است (English: 'is'). These likely differ in frequency as both words are expressed in multiple ways in both orthographies. For example, аст / است can be attached to the previous word in Tajik but is always written separately in Farsi. Meanwhile, و can be written either separately or attached to the previous word in Farsi. Its two Tajik equivalents, ва and у, are written separately or attached, respectively.

We also note that all but one of the top ten most frequent Tajik and Farsi tokens in ParsText are stop words, with the ninth most frequent token being тоҷикистон/تاجیکستان (English: 'Tajikistan'). While stop words typically do not indicate alignment, in the case of our digraphic, word-to-word corpus, the fact that the Farsi token frequencies are generally very close to their Tajik equivalents

[6] http://dariussthoughtland.blogspot.com/

[7] http://jaamjam.blogspot.com/

[8] https://www.bbc.com/tajik

[9] https://www.bbc.com/persian/indepth/cluster_tajikistan_page

[10] https://www.bbc.com/persian

[11] https://github.com/alvations/gachalign

Farsi				Tajik			
Token	Transliteration	Translation	Frequency	Token	Transliteration	Translation	Frequency
و	va, u	‘and’	2,096	дар	dar	‘in’	1,495
در	dar	‘in’	1,498	ба	ba	‘to’	1,323
به	ba	‘to’	1,307	ва	va	‘and’	1,230
که	ki	‘that’ (Conj.)	1,212	ки	ki	‘that’ (Conj.)	1,216
از	az	‘from	1,154	аз	az	‘from’	1,149
این	in	‘this’	883	ин	in	‘this’	910
است	ast	‘is’	636	бо	bo	‘with’	448
با	bo	‘with’	458	аст	ast	‘is’	394
تاجیکستان	tojikiston	‘Tajikistan’	428	тоҷикистон	tojikiston	‘Tajikistan’	393
بود	bud	‘was’	290	буд	bud	‘was’	285

Table 3: Top 10 Most Frequent Farsi and Tajik Tokens in ParsText

indicate that the same wording is being used. As under-resourced languages, few Farsi stop word lists exist, and we know of only one for Tajik. Ensuring removal of the exact same stop words requires a digraphic list of Tajik/Farsi stop words. This task is best left to native speakers.

Additionally, although one would expect the frequency of ’Tajikistan’ to indicate abnormalities, further manual inspection revealed no unnatural occurrences of the word. As such, we believe this is a natural reflection of the provenance of these texts. As the sources all focus on Tajikistan, it appears that the frequency of this proper noun has become similar to that of a pronoun.

To analyze ParsText on a character level, the most frequent trigraphs (character trigrams) were also calculated. To ensure that three-letter tokens did not overpopulate this list, trigraph frequency was conducted over all word types, rather than tokens. The character ’#’ is inserted at word-initial and word-final postion. These data are presented in Tables 4 and 5.

Trigraph	Transliteration	Frequency
а н д	a n d	534
р о #	r o #	519
# т а	# t a	506
н и #	n i #	463
# м а	# m a	410
о и #	o i #	403
# н а	# n a	351
о н и	o n i	319
# м у	# m u	296
и и #	i i #	217

Table 4: Top 10 Most Frequent Tajik Trigraphs in ParsText (Calculated Over Word Types)

Based on the above trigraphs, several more observations can be made. First, the Persian subject marker ’ро’ / ’را’ (’ro’) appears to be well represented, being present as (’r o #’) among the Tajik trigraphs and (’r a #’) and ’r a _’) among the Farsi trigraphs. The two Farsi forms demonstrate that the ZWNJ is not always used by native speakers. The Ezafe also appears to be present in the Tajik list as (’n i #’, ’o n i’, and ’i i #’) and the Farsi list as (’a i #’ and ’h a i’). Altogther, these trigraph frequencies greatly differ, demonstrating the large contrasts between the Tajik-Cyrillic and Perso-Arabic script. To provide a different picture, we conduct a different set of trigraph frequencies without any vowels, included Appendix C. However, as several Tajik consonants have multiple (up to four) equivalents in Farsi, this does necessarily result in a clearer picture.

Trigraph	Transliteration	Frequency
ا ن #	a n #	528
را#	r a #	528
_ را	_ r a	431
ای#	a i #	424
ه ا ی	h a i	420
ه ا _	h a _	392
ن د #	n d #	392
م ی _	m i _	384
# م ی	# m i	356
ی - #	i - #	343

Table 5: Top 10 Most Frequent Farsi Trigraphs in ParsText (Calculated over Word Types)
Note: The Perso-Arabic text must be read right-to-left and the ZWNJ is denoted with ’_’.

6. Conclusion and Future Work

This paper presented ParsText, a corpus of 2,813 digraphic Persian sentences written by native speakers in the Tajik-Cyrillic and Perso-Arabic orthographies. ParsText contains manually-collected data from blog pages and BBC news articles. Based on manual inspection by a non-native speaker and analysis of most frequent tokens, we confirmed ParsText exhibits word-to-word alignment, a crucial requirement for direct evaluation of Tajik-Farsi transliteration systems that is unavailable in other parallel corpora. As such, it enables direct evaluation of Tajik-Farsi machine transliteration efforts. The corpus is available on OSF.[12]

[12] https://doi.org/10.17605/OSF.IO/37GZX

7. Bibliographical References

Sina Ahmadi, Hossein Hassani, and Daban Q. Jaff. 2022. Leveraging Multilingual News Websites for Building a Kurdish Parallel Corpus. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(5).

Habibollah Asghari, Jalal Maleki, and Heshaam Faili. 2014. A probabilistic approach to Persian ezafe recognition. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 138–142, Gothenburg, Sweden. Association for Computational Linguistics.

BBC. 2015. Поёни фаъъолияти сафҳаи сириллики бахши форсии Би-би-сӣ.

Chris Irwin Davis. 2012. Tajik-Farsi Persian transliteration using statistical machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3988–3995, Istanbul, Turkey. European Language Resources Association (ELRA).

Ehsan Doostmohammadi, Minoo Nassajian, and Adel Rahimi. 2020. Persian Ezafe Recognition Using Transformers and Its Role in Part-Of-Speech Tagging. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 961–971, Online. Association for Computational Linguistics.

Tomasz Gacek. 2015. Some comments on a parallel text in Dari, Tojiki and Farsi. In Anna Krasnowolska and Renata Rusek-Kowalska, editors, *Studies on the Iranian World*, volume 2, chapter Medieval and Modern, pages 23–34. Jagiellonian University Press, Kraków.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Daler Ghufronov. 2008. Shifting Tajik writing system to Arabic script takes a lot of time, says minister. *ASIA-Plus*.

Edward Gow-Smith, Mark McConville, William Gillies, Jade Scott, and Roibeard Ó Maolalaigh. 2022. Use of Transformer-Based Models for Word-Level Transliteration of the Book of the Dean of Lismore. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 94–98, Marseille, France. European Language Resources Association.

Jalal Maleki. 2008. A romanized transcription for persian.

Karine Megerdoomian and Dan Parvaz. 2008. Low-density language bootstrapping: the case of tajiki Persian. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

John R. Perry. 2005. *A Tajik Persian Reference Grammar*. Brill, Leiden, The Netherlands.

Bashar Talafha, Analle Abuammar, and Mahmoud Al-Ayyoub. 2021. Atar: Attention-based LSTM for Arabizi transliteration. *International Journal of Electrical and Computer Engineering (IJECE)*, 11(3):2327–2334. Number: 3.

Liling Tan and Francis Bond. 2014. NTU-MC Toolkit: Annotating a Linguistically Diverse Corpus. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 86–89, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Wikimedia Foundation. 2023a. List of wikipedias.

Wikimedia Foundation. 2023b. Statistics.

A. Challenges in Tajik-Farsi Transliteration

As previously described, Farsi and Tajik diverge in a number of ways which render one-to-one letter conversion largely ineffective. An example transliteration employing such a technique can be seen in Table 6.

Farsi	Farsi Translit.	Tajik	Tajik Translit.
من کتاب را خواندم	mn ktob ro xwondm	ман китобро хондам	man kitobro xondam

Table 6: One-to-one Transliteration of Farsi and Tajik

Owing to the incongruous natures of the two scripts, Perso-Arabic an imperfect abjad and Tajik-Cyrillic an alphabet, many characters map to a single character and vice versa.

A.1. Vowels

The character ا, known as *alef*, can represent several different vowels as demonstrated in Table 7.

The letter و, known as *vav*, can map to the vow-

Farsi	Farsi Translit.	Tajik	Tajik Translit.	English
انجمن	anjmn	анҷуман	anjuman	‘organization’
انتخاب	antxob	интихоб	intixob	‘choice’
امید	amyd	умед	umed	‘hope’
او	aw	ӯ	ü	‘(s)he’
آهنگ	ohng	оҳанг	ohang	‘song’
خاردن	xoridn	хоридан	xoridan	‘to itch’

Table 7: Examples of Alef mapping to various vowels

Farsi	Farsi Translit.	Tajik	Tajik Translit.	English
ولایت	wlayt	вилоят	viloyat	‘oblast’
آورد	owrd	овард	ovard	‘brought’
گاو	gaw	гов	gov	‘cow’
بود	bwd	буд	bud	‘was’
امروز	amrwz	имрӯз	imrüz	‘today’

Table 8: Examples of Vav mapping to vowels and consonants

els у and ӯ, or the consonant в, as shown in Table 8.

The letter ی, known as *ye*, maps to several different vowels, as seen in Table 9.

Farsi	Farsi Translit.	Tajik	Tajik Translit.	English
یراق	yroq	ярок̦	yaroq	‘weapon’
دریا	drya	дарё	daryo	‘river/sea’
چای	çay	чой	çoy	‘tea’
ایران	ayran	Эрон	Eron	‘Iran’
خیلی	xyly	хеле	xele	‘very’
عالی	’aly	олӣ	olī	‘great’
حتی	hty	ҳатто	hatto	‘even’

Table 9: Examples of Ye mappings

The consonant ه, known as *he (do cheshm)*, maps to either the consonant ҳ or to vowels when in word final position, as shown in Table 10.

The character ع, or *ayn*, can map to any vowel (see Table 11), and is also inconsistently written.

The ء, or *hamza*, exhibits similar behavior to *ayn* by mapping to both vowels and the Tajik glottal stop sign. It can be written as a standalone letter, or over *alif*, *vav*, or *ye*. However, this character is often replaced with a *ye* or simply removed from the letter it is written over.

Vowel diacritics are often unwritten in the Perso-Arabic script, further obfuscating short vowel determination. Without vowel diacritics, the word گرد may represent either гард /gard/ (’dust’), гирд /gird/ (’round’), or гурд /gurd/ (’hero’).

A.1.1. Consonants

As the Perso-Arabic script retains many redundant Arabic consonants from Arabic, some Cyrillic letters each have multiple Perso-Arabic letter equivalents. Table 12 provides an overview of these.

Outside of these redundant consonants and the vowels mentioned previously, consonant to consonant mapping between the two scripts is one-to-one and can be considered trivial.

Farsi	Farsi Translit.	Tajik	Tajik Translit.	English
به	bh	ба	ba	‘to’
که	kh	ки	ki	‘that (conj.)’
چه	çh	чи	çi	‘what’
قاعده	qa’dh	қоида	qoida	‘rule’
سیاه	syah	сиёҳ	siyoh	‘black’
ده	dh	даҳ	dah	‘ten’
فربه	frbh	фарбеҳ	farbeh	‘fat’

Table 10: Examples of He Mapping

Farsi	Farsi Translit.	Tajik	Tajik Translit.	English
عضو	’zw	узв	uzw	‘limb’
علامت	’lamt	аломат	alomat	‘sign’
فعالیت	f’alyt	фаъолият	fa’oliyat	‘activity’
ساعت	sa’t	соат	soat	‘hour’
تاریخ	taryx	таърих/торих	ta’rix/torix	‘history’
قرآن	qron	Қуръон	Qur’on	‘Quran’

Table 11: Examples of Ayn mapping

B. ParsText Sample Sentences

These example sentences have been lowercased.

(1) вай гуфтааст ки донишгоҳҳои табрез низ омодаи пазириши донишҷӯёни тоҷик ҳастанд

وی گفته است که دانشگاههای تبریز
نیز آماده پذیرش دانشجویان تاجیک هستند

(2) як сол пеш аз таваллуди мирзо фатҳъалӣ падараш аз ин мақом барканор шуда буд

یک سال پیش از تولد میرزا فتحعلی پدرش از
این مقام برکنار شده بود

(3) мутмаъинам ки мардуми тоҷикистон ҳам аз аҳволи эрон нигарон ҳастанд

مطمئنم که مردم تاجیکستان هم از احوال ایران
نگران هستند

C. Trigraphy Frequencies with Vowels Removed

We provide trigraph frequencies excluding all in Tajik and Farsi in this appendix. For Tajik, the letters removed were у, е, ҳ, ъ, а, о, э, я, и, й, ӣ and ю. For Farsi, the letters removed were ا, آ, و, ع, and ی. Tables 13 and 14 show the trigraph frequencies in Tajik and Farsi.

Phoneme	Tajik	Farsi
/z/	з	ز ذ ض ظ
/s/	с	س ص ث
/t/	т	ت ط
/h/	ҳ	ح ه

Table 12: One to many Mappings of Consonants from Tajik to Farsi

Trigraph	Transliteration	Frequency
н д #	n d #	329
с т #	s t #	219
т р #	t r #	140
# б р	# b r	134
р н #	r n #	124
т н #	t n #	107
с т н	s t n	94
# ф р	# f r	94
# с р	# c r	94
# м р	# m r	94
д н #	d n #	94

Table 13: Most Frequent Tajik Trigraphs in Pars-Text (Without Vowels)

Trigraph	Transliteration	Frequency
_ ر #	_ r #	395
ن د #	n d #	346
# م _	# m _	204
س ت #	s t #	140
# ب ر	# b r	139
ر ن #	r n #	118
# ن م	# n m	104
# ف ر	# f r	104
ت ن #	t n #	97
ت ر #	t r #	96

Table 14: Most Frequent Farsi Trigraphs in Pars-Text (Without Vowels)
Note: The Perso-Arabic text must be read right-to-left and the ZWNJ is denoted with ’_’.