# Transformers for Bridging Persian Dialects: Transliteration Model for Tajiki and Iranian Scripts

**MohammadAli SadraeiJavaheri*, Ehsaneddin Asgari♢, Hamid Reza Rabiee***
* Sharif University of Technology
♢Qatar Computing Research Institute, Doha, Qatar
{m.sadraei, rabiee}@sharif.edu and easgari@hbku.edu.qa

**Abstract**

In this study, we address the linguistic challenges posed by Tajiki Persian, a distinct variant of the Persian language that utilizes the Cyrillic script due to historical "Russification". This distinguishes it from other Persian dialects that adopt the Arabic script. Despite its profound linguistic and cultural significance, Tajiki Persian remains a low-resource language with scant digitized datasets for computational applications. To address this deficiency, we created a parallel corpus using Shahnameh, a seminal Persian epic poem. Employing optical character recognition, we extracted Tajiki Persian verses from primary sources and applied a heuristic method to align them with their Iranian Persian counterparts. We then trained and assessed transliteration models using two prominent sequence-to-sequence architectures: GRU with attention and transformer. Our results underscore the enhanced performance of our models, particularly in contrast to pre-trained large multilingual models like GPT-3.5, emphasizing the value of dedicated datasets in advancing computational approaches for underrepresented languages. With the publication of this work, we are disseminating, for the first time, a vast collection of Persian poetry spanning 1000 years, transcribed in Tajiki scripts for the benefit of the Tajiki-speaking communities. The dataset, along with the model's code and checkpoints, is accessible at https://github.com/language-ml/Tajiki-Shahname, marking a significant contribution to computational linguistic resources for Tajiki Persian.

**Keywords:** Tajiki Persian, Iranian Persian, Transliteration

## 1. Introduction

Persian language, also known as Farsi, is an Indo-European language spoken by nearly 130 million people worldwide. It is the official language of Iran, Afghanistan, and Tajikistan. This language has three mutually intelligible standard varieties: Iranian Persian, Dari Persian, and Tajiki Persian (Windfuhr, 2009). Tajiki Persian is primarily spoken in Tajikistan but also has a significant number of speakers in parts of Uzbekistan. Despite the high degree of similarity among the Persian variations, there are discernible differences in vocabulary, pronunciation, and specific syntactic structures between Tajiki Persian and Iranian Persian. A prominent distinction between Tajiki Persian and other varieties arises from their scripts. The "Russification" of central Asia by the Soviet Union in the late 1930s led to the introduction of the Cyrillic script in Tajikistan, replacing the traditional Persian alphabet (Keller, 2001) (Figure 1). This change has since impeded written communication between Tajikistan and other Persian-speaking regions.

Today, only a limited number of individuals in Tajikistan can read the Persian script. Consequently, the majority are unable to access Iranian Persian written content, including literature and most online materials. This limitation has deepened the cultural divide between these groups. Due to the scarce availability of Tajiki Persian materials, many in Tajikistan prefer using Russian, which offers a broader range of scientific and literary texts compared to Tajiki Persian (Khudoikulova, 2015). It's worth noting that Persian itself is a language rich in literature (Storey, 1972).

Despite the close resemblance between these two Persian language variations, Iranian Persian boasts extensive resources, while Tajiki Persian is markedly resource-deprived. For instance, in the FLORES-101 dataset of Wiki sentences (Goyal et al., 2021), Iranian Persian comprises 620M data points, whereas Tajiki Persian accounts for a mere 0.5M. Moreover, Iranian Persian ranks as the $11^{th}$ most used language on the web and benefits from robust support by language technologies, including machine translation. Conversely, Tajiki Persian occupies the $79^{th}$ spot in terms of web resources [1].

Transliteration is the process of converting text from one script to another. It's vital for representing foreign words from different writing systems within a single language (Knight and Graehl, 1997). This method is also advantageous for languages with more than one standard script. Transliteration systems cannot simply be replaced by translation systems, as in some contexts, every word is crucial. They are particularly essential for literary texts, such as poems, where each word and its

[1] $w3techs.com/technologies/overview/content\_language$



*LREC-COLING 2024*, pages 16770–16775

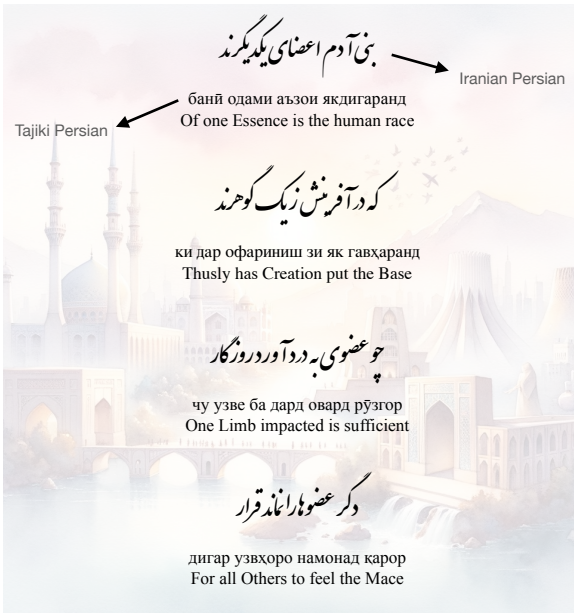


Figure 1: Four hemistiches from Saadi (1210-1291), presented in Iranian Persian and Tajiki Persian scripts, along with their translations.

placement are significant. Using a translation system might disrupt the poem's rhythm and rhyme (Friar, 1983).

Transliteration between Iranian and Tajiki is an essential language processing task for various reasons. Specifically, (i) transliteration can act as a bridge between Tajiki and Iranian cultures, and (ii) the automatic generation of Tajiki from Iranian can help mitigate the resource limitations associated with Tajiki. However, the development of a transliteration system necessitates a parallel corpus. To the best of our knowledge, no adequate parallel resources for this purpose have been presented in previous work.

In summary, our contributions are as follows: **(i)** we introduce a transformer-based transliteration system between Iranian and Tajiki dialects of Persian, which outperformed few-shot learning on GPT 3.5. **(ii)** For the first time, we are releasing an extensive collection of Persian poetry, covering 1000 years, transcribed in Tajiki scripts, to benefit the Tajiki-speaking communities and bridges the gap between Persian-speaking nations.

## 2. Related Works

### 2.1. Sequence-to-sequence Models

Sequence-to-sequence models, also known as encoder-decoder models, transform a text input into another form through encoding and decoding processes (Sutskever et al., 2014). Prominent applications of such encoding-decoding scenarios include machine translation, transliteration, and summarization. These models are trained on parallel corpora to generate a target sequence from a provided source sequence. Over the past decade, Recurrent Neural Networks (RNNs) initially made a significant mark on machine translation (Sutskever et al., 2014), with notable architectures like the Gated Recurrent Unit (GRU) (Chung et al., 2014) and Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997). A major limitation of this approach was the bottleneck between the encoder and decoder components. To mitigate this, the attention mechanism was introduced Bahdanau et al. (2015). This mechanism was so effective that it led to the inception of a novel model architecture known as the transformer (Vaswani et al., 2017).

### 2.2. Persian Transliteration

Multilingual sequence-to-sequence transliteration models have been proposed (Firat et al., 2016; Kundu et al., 2018). Existing models for Persian-English transliteration are also available (Mahdi Mahsuli and Safabakhsh, 2017). A study by Davis (2012) delved into Tajik-Iranian transliteration using statistical models, whose model was not available anymore. Although there is a website providing transliteration between Tajik and Iranian through a rule-based method[2], no publicly available models currently offer this service. We also did not obtain permission from the owner of $persian - tajik.ir$ to systematically access their model for comparison purposes.

## 3. Problem Description

Given the similarity in spoken forms of the two languages, it would be tempting to assume that the translation task is straightforward, potentially resolved with a mere letter substitution. However, as depicted in Figure 1, their written forms differ significantly in appearance. Thus, this assumption is incorrect due to various challenges, some of which we detail below:

### 3.1. Homophone Consonants

The Iranian Persian writing system includes several consonants that sound alike. This poses challenges when transliterating from Tajiki to Iranian Persian because ensuring the correct spelling is difficult. Sometimes, two words can be homophones, and incorrect spelling can drastically alter the intended meaning in Iranian Persian. For instance, the word /hæjɔːt/ is written in Tajiki Persian as "ҳаёт" but in Iranian Persian, it has

[2]www.persian-tajik.ir

two written forms with different meanings: "حیات" (meaning life) and "حیاط" (meaning yard).

### 3.2. Short Vowels

The Iranian Persian script, similar to the Arabic script, does not explicitly write short vowels. This becomes problematic when transliterating from Iranian to Tajiki, as the system must predict the correct short vowels to insert between consonants. In Persian noun groups, a short vowel, "Ezâfe", is added to the end of a noun, pronounced as /-e/. Though this vowel isn't written in the Iranian Persian script, it appears in the Tajiki script. This poses challenges when transliterating from Iranian Persian to Tajiki, as noun groups must be identified.

### 3.3. Direct Object Marker

Persian uses the postposition marker /râ/ for definite direct objects. In Iranian Persian, this marker, "را", is a separate token placed after the direct object with a space between the noun and the token. However, in Tajiki, it attaches directly to the noun without any space, acting as a suffix "-ро" (Davis, 2012). Due to this difference, when transliterating from Tajiki to Iranian Persian, the system must decide if the /râ/ at the end of a Tajik word is an object marker or just part of the word. If recognized as an object marker, a space should be inserted during transliteration.

### 3.4. Pronunciation Shift

Because of geographical distance, some pronunciations of words have changed over time. When the writing system changed, the updated pronunciation influenced the spelling. For instance, in Iranian Persian, the word "تاریخ" is pronounced as /tɔːrix/ and means "history." It derives from a similar word in Arabic. However, in Tajiki Persian, this word is written as "таърих" and pronounced as /tæʔrix/. Thus, the transliteration model needs to account for such edge-case words.

## 4. Dataset

The *Shahnameh*, meaning "The Book of Kings", is a lengthy epic poem penned by the Persian poet Ferdowsi between c. 977 and 1010 CE, standing as the greatest Persian epic resource. Comprising nearly 50,000 verses, the Shahnameh ranks among the world's longest epic poems. It predominantly chronicles the mythical, and to a lesser extent, the historical events of the Persian Empire from the world's creation up to the Muslim conquest in the seventh century (Davis, 1995). The work holds immense significance in Persian culture and the Persian language, symbolizing a literary masterpiece and defining the ethno-national cultural identity of the Persian people (Mousavi, 2021).

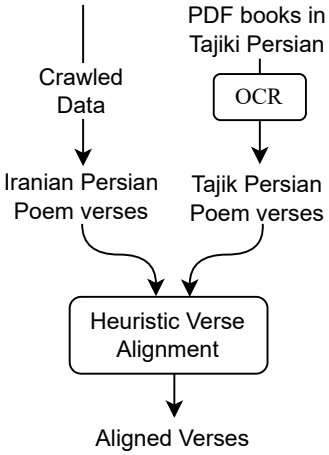


Figure 2: A workflow diagram illustrating the process of extracting and aligning Persian and Tajik verses.

We built our dataset by taking the Iranian Persian version of the *Shahnameh* poems from Ganjoor[3]. Ganjoor is an Iranian Persian site that offers a vast collection of poems in text format, encompassing more than a million verses that trace back to a thousand years ago. We got the poem from Ravshanfikr[4] in ten PDF files for the Tajik version. Ravshanfikr is an online library with many books in Tajiki Persian. We used Optical Character Recognition (OCR) to get text from these PDFs and relied on Tesseract (Smith, 2007) for the extraction.

We attempted to find and align the same verse from both Tajiki and Iranian versions to create a parallel corpus. This was challenging because there isn't a single, standard version of Shahnameh. Several versions exist, each with variations in verse order, word choices, and even the total verse count. Aligning the Iranian Persian version with the Tajiki one proved difficult. To tackle this, we used a heuristic method to remove the vowels from the verses. If two verses, without the vowels, were similar enough, we paired them in our dataset. Figure 2 illustrates the workflow diagram of our procedure, while Table 1 offers insights into the statistics of the final dataset.

[3] ganjoor.net
[4] ravshanfikr.tj

| Corpus | Verses Count |
|---|---|
| Tajiki Persian | 52,146 |
| Iranian Persian | 49,609 |
| Aligned | 34,105 |

Table 1: Statistics of the constructed dataset. Due to variations in the verses across different versions, not all verses could be perfectly aligned.

## 5. Approach and Evaluation

To assess the effectiveness of our dataset, we trained various models for both transliteration directions (Tajiki to Iranian and Iranian to Tajiki) employing two distinct sequence-to-sequence architectures: GRU with attention (Bahdanau et al., 2015) and the transformer encoder-decoder (Vaswani et al., 2017). These models were trained at the character level. We partitioned our dataset into 80%, 10%, and 10% splits for training, validation, and testing, respectively, utilizing the validation set for early stopping.

Concerning hyperparameters, we used a two-layer configuration for both the transformer and GRU. The embedding layer size was set to 128 in both architectures. We used a hidden size of 512 for the GRU and a feedforward size of 512 for the transformer. Additionally, we used four heads in the multi-head attention mechanism of the transformer

We selected the mean edit distance as our evaluation metric. Although metrics like the BLEU score (Papineni et al., 2002) or ROUGE score (Lin, 2004) are valuable for translation tasks, they are not well-suited for transliteration. In transliteration, unlike translation, there exists a gold label. Therefore, we can determine the edit distance between the target and the predicted sequences for each sequence and subsequently average these distances to arrive at this metric. Notably, a lower value for this metric indicates better performance.

| Method | Mean Edit Distance |
|---|---|
| **Tajiki to Iranian** | |
| GRU | 0.99 |
| Transformer | **0.88** |
| gpt-3.5-turbo | 5.44 |
| **Iranian to Tajiki** | |
| GRU | 1.11 |
| Transformer | **1.05** |
| gpt-3.5-turbo | 6.42 |

Table 2: Results of the tests on various transliteration methods. The table shows the average edit distance for each method, with a lower score indicating better performance.

As a point of comparison, we evaluated an LLM using a 3-shot prompt. Our tests of different LLMs in simple conversation showed that the only generative LLM with a good Pesian capability is ChatGPT (Brown et al., 2020). Consequently, we employed the ChatGPT API for transliteration.

## 6. Result

Results from the ChatGPT model, as well as our trained models, are presented in Table 2. We evaluated methods using hemistiches. Each verse is composed of two hemistiches. The results indicate that LLM models find this task to be challenging. Of the techniques we tested, the transformer proved more effective than the GRU.

It is important to note that each hemistich in Iranian Persian has an average length of approximately 25 characters. In comparison, Tajiki Persian hemistiches average around 30 characters. As detailed in §3.2, this length disparity largely stems from the omission of short vowels in Iranian Persian script.

Using an accurate Persian transliterator, we converted 2,620,477 Persian poem hemistiches from the Ganjoor dataset (outlined in §4) to the Tajiki script, encompassing over 1000 years of poetry. Upon publication of this paper, this resource will be made available for both digital humanities endeavors and the enrichment of Tajiki-speaking communities.

## 7. Conclusion

In this paper, we focused on Tajiki Persian, a low-resource language, and introduced a new parallel corpus derived from the $Shahnameh$, or "Book of Kings" an epic poetry work written between the late 10th and early 11th century. Utilizing this dataset, we trained and evaluated models based on GRU and transformer architectures for transliteration tasks. As anticipated, the Transformer-based transliterator surpassed the recurrent neural network solution and the GPT 3.5 model in few-shot learning scenarios. Our findings showed that even pre-trained LLMs, such as GPT 3.5, fell short in this task due to the limited data available for the Tajiki Persian language. We remain optimistic that future efforts will produce more datasets and technologies tailored for these underrepresented languages.

One limitation of our model was its reliance on a historical poetic dataset. For optimal performance in transliterating contemporary texts, utilizing datasets that reflect current language practices is crucial. We also suggested that the integration of monolingual data from both languages might enhance the model's adaptability and performance

across various text domains.

## 8. Bibliographical References


Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tom B. Brown, Benjamin Mann, Nick Ryder, and Melanie Subbiah et al. 2020. Language models are few-shot learners.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.

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).

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Kimon Friar. 1983. How a poem was translated. *Translation review*, 11(1):15–19.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Shoshana Keller. 2001. *To Moscow, Not Mecca: The Soviet Campaign Against Islam in Central Asia, 1917-1941*. Praeger, Westport, Conn.

Noora Khudoikulova. 2015. Linguistic situation in tajikistan: language use in public space. *Russian Journal of Communication*, 7(2):164–178.

Kevin Knight and Jonathan Graehl. 1997. Machine transliteration. *arXiv preprint cmp-lg/9704003*.

Soumyadeep Kundu, Sayantan Paul, and Santanu Pal. 2018. A deep learning based approach to transliteration. In *Proceedings of the Seventh Named Entities Workshop*, pages 79–83, Melbourne, Australia. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Mohammad Mahdi Mahsuli and Reza Safabakhsh. 2017. English to persian transliteration using attention-based approach in deep learning. In *2017 Iranian Conference on Electrical Engineering (ICEE)*, pages 174–178.

SeyyedMehdi and Mousavi. 2021. Canonizing shahnameh in early modern iran: A historical-semiotic approach [in english].
, (8):185 – 202.

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.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 629–633, Washington, DC, USA. IEEE Computer Society.

Charles Ambrose Storey. 1972. *Persian literature: a bio-bibliographical survey*. Psychology Press.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Gernot Windfuhr. 2009. *The Iranian Languages*. Taylor and Francis.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.


## 9. Language Resource References


Dick Davis. 1995. The shahnameh.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.