Transliterating Kurdish texts in Latin into Persian-Arabic script

Hossein Hassani
University of Kurdistan Hewlêr
Kurdistan Region - Iraq
hosseinh@ukh.edu.krd

Abstract

Kurdish is written in different scripts. The two most popular scripts are Latin and Persian-Arabic. However, not all Kurdish readers are familiar with both mentioned scripts that could be resolved by automatic transliterators. So far, the developed tools mostly transliterate Persian-Arabic scripts into Latin. We present a transliterator to transliterate Kurdish texts in Latin into Persian-Arabic script. We also discuss the issues that should be considered in the transliteration process. The tool is a part of Kurdish BLARK, and it is publicly available for non-commercial use[1].

1 Introduction

Kurdish is a multi-dialect that is written in different scripts (Hassani et al., 2016). The two most popular scripts are Latin and Persian-Arabic. However, not all Kurdish readers are familiar with both mentioned scripts that could be resolved by automatic transliterators. So far, the developed tools mostly transliterate Persian-Arabic scripts into Latin (Hassani, 2018; Ahmadi, 2019). We present a transliterator to transliterate Kurdish texts in Latin into Persian-Arabic script. Kurdish language processing requires endeavor by interested researchers and scholars to overcome the resource and tool scarcity to eliminate the obstacles in front of its tasks. The areas that need attention and the efforts required have been addressed in (Hassani, 2018).

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents different parts of the dataset, such as the dictionary, phoneset, transcriptions, corpus, and language model. Finally, Section 4 concludes the paper and suggests some areas for future work.

2 Related work

Several scholars have addressed transliterators for Kurdish scripts (Esmaili et al., 2014; Hassani, 2018; Ahmadi, 2019). Those studies mostly focused on transliteration from Persian-Arabic into Latin script. Particularly, Ahmadi (2019) addressed several important issues in transliterating Kurdish texts in Persian-Arabic into Latin script and proposed and implemented appropriate resolutions for them.

Also, some online tools exist for transliteration of Kurdish scripts into each other, such as https://www.lexilogos.com/keyboard/kurdish_conversion.htm and http://www.transliteration.kpr.eu/ku/en.html. Those are mainly based on standard Latin scripts, and therefore, they miss some of the Persian-Arabic scripts such as ح, ع, and غ. Other issues also exist in those transliterators, particularly when the text is in Kurmanji Kurdish. For example, both mentioned tools transliterate the Kurmuanji word "diînine" as " دینە", while the correct

[1] https://kurdishblark.github.io/

transliteration is "دئننه".

Furthermore, to use the mentioned tools, the users must copy and paste their texts to the online tool that not only has limited space but also might not be desirable to the users from the copyright perspective. Our translator addresses those issues, and its script is publicly available under the GNU license.

3 The Transliterator

Our transliterator is a script in Python that could be used standalone. It receives an input file in the text that should be saved as UTF-8, and it provides the transliterated version of the text also in UTF-8. It resolves the mentioned issues that we addressed in Section 2. Particularly, it works for both Kurmanji and Sorani texts in Latin.

Figure 1 shows a text in Latin script, and Figure 2 illustrates its equivalent in Persian-Arabic script that our suggested script has transliterated.

```
Wan pêncane dibêjinê "fasiqêt xemse", ye'nî pênc fasiq û xirabîker.
Derheq wan hatiye gotin, "fasiqêt xemse, diyar e mişk û dûpişk û mar e,
keşkele ye, paşî kolare." Her pênc rojekê çûne ḫuzûra şeytanî, gotin,
"em hatîne me xelat bikey. Me gelek xizmet kiriye w noke jî dikeyn û ji
noke paş we jî dê her keyn." Şeytanî got, "xizmeta hingo çi ye w hingo
çiyê weto kiriye ku ez pê keyfxoş bibim, da hingo ji bedel wê çakîyê
xelat bikem?" Gotin, "ma tu dujminê benî-ademî nîy [nînî]?" Got, "belê,
ez zor dujminê wan im." Gotin, "çak e. Bes eger kesek zerera dujminê miro
biket, miro wî kesî nakete dostê xo w çakîyê digel wî kesî naket?"
Şeytanî got, "belê, bo çî?" Gotin, "bes lazim e ku me xelat bikey, çunko
em daymî zerera benî-ademî dikeyn."

Şeytanî got, "xizmeta xo w zerera xo derheq benî-ademî bo min beyan ken
da xelatê her kesekî bi qeder xizmeta wî bihête dan, çunko di her tiştekî
da inṣaf û 'edalet çak e." Gotin, "belê, dê noke 'erz-i ḫuzûra aẋayê xo
keyn."
```

Figure 1: A Kurmanji text in Latin script.[2]

The transliterator is rule-based. We have recognized one-to-one one-to-many, two-to-one, two-to-many, and three-to-many letters transliterations. The *many* equivalent is maximum three. The order that the transliterator applies those mapping is important. The mapping method and the mapping order were designed based on studying various texts and writing styles. For example, the case of *bizroke* that (Ahmadi, 2019) has addressed it transliteration from Persian-Arabic into Latin must also be considered in transliteration form Latin into Persian-Arabic. That is "min" must be written as "من". However, the users might still find issues when they transliterate using the tool. We would be grateful to report the possible issues to enhance the script.

The users of the transliterator may notice an issue with the full stop at the end of a paragraph when they use editors such as Microsoft Word or Libre Writer. When they open the text in those editors, the full stop at the end of the paragraphs might flip to the left. That case happens if the

[2]The text is from: Öpengin, Ergin. 2021. Bazeber: Nivîsarên Mela Se'îd Şemdînanî li ser çand û dîroka Kurdistana navendî [The texts of Mulla Said Shamdinani about the history and culture of Central Kurdistan]. Istanbul: Avesta.

وان پێنجانه دبێژنێ "فاسقێت خەمسه"، یەعنی پێنج فاسق و خرابیکەر. دەرهەق وان هاتیه گۆتن، "فاسقێت خەمسه، دیاره مشك و دووپشك و ماره، کەشکەله یه، پاشی کۆلاره." هەر پێنج رۆژەکێ چوونه حوزووڕا شەیتانی، گۆتن، "ئەم هاتینه مه خەلات بکەی. مه گەلەك خزمەت کریه و نۆکه ژی دکەین و ژ نۆکه پاش وه ژی دێ هەر کەین." شەیتانی گۆت، "خزمەتا هنگۆ چ یه و هنگۆ چیێ وەتۆ کریه کو ئەز پێ کەیفخۆش ببم، دا هنگۆ ژ بەدەل وێ چاکییێ خەلات بکەم؟" گۆتن، "ما تو دوژمنێ بەنی ئادەمی نیی [نینی]؟" گۆت، "بەلێ، ئەز زۆر دوژمنێ وانم." گۆتن، "چاکه. بەس ئەگەر کەسەك زەرەرا دوژمنێ مرۆ بکەت، مرۆ وی کەسی ناکەته دۆستێ خۆ و چاکییێ دگەل وی کەسی ناکەت؟" شەیتانی گۆت، "بەلێ، بۆ چی؟" گۆتن، "بەس لازمه کو مه خەلات بکەی، چونکۆ ئەم دایمی زەرەرا بەنی ئادەمی دکەین."

شەیتانی گۆت، "خزمەتا خۆ و زەرەرا خۆ دەرهەق بەنی ئادەمی بۆ من بەیان کەن دا خەلاتێ هەر کەسەکی ب قەدەر خزمەتا وی بهێتە دان، چونکۆ د هەر تشتەکی دانصاف و عەدالەت چاکه." گۆتن، "بەلێ، دێ نۆکه عەرزی حوزووڕا ئاغایێ خۆ کەین."

Figure 2: A Persian-Arabic version of the text in Figure 1 transliterated using our suggested script.

default direction for the input text in the editor is set to LTR. Using an appropriate command, according to the editor, the issue is resolved. For example, in Libre Writer, the user can select the entire text and then press Ctrl+Righ Shift.

Also, it is usually necessary to change the font to the Unicode fonts, for example, UnikurdWeb, to view and print the output correctly.

4 Conclusion

We presented a script that transliterates Kurdish texts in Latin script into Persian-Arabic texts. The script could be used standalone, and it is publicly available. The script resolves some issues that available online transliterators have not considered. As it is standalone, it could be used with requiring the Internet connection, and it does not have size limitations for its input document. In its current form, the transliterator provides proper output for Kurmanji and Sorani Kurdish.

In the future, we like to receive feedback [3] from the transliterator concerning any possible issues to enhance the script. We also want to test it for Zazaki and Hawramit texts as they have special letters that are not used in Sorani and Kurmanji.

Acknowledgement

We are grateful to Dr. Ergin Öpengin for his feedback and assistance in checking the accuracy of the transliterator.

References

Ahmadi, S. (2019). A rule-based kurdish text transliteration system. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(2):1–8.

Esmaili, K. S., Salavati, S., and Datta, A. (2014). Towards kurdish information retrieval. *ACM Transactions on Asian Language Information Processing (TALIP)*, 13(2):7.

Hassani, H., Medjedovic, D., et al. (2016). Automatic Kurdish Dialects Identification. *Computer Science & Information Technology*, 6(2):61–78.

[3] Please kindly report any issues via Kurdish BLARK (`https://kurdishblark.github.io/`)

Hassani, H. (2018). BLARK for multi-dialect languages: towards the Kurdish BLARK. *Language Resources and Evaluation*, 52:625–644.