Aysu Akcan

University of Vienna, Austria
aysu.akcan@univie.ac.at

HTRising Ottoman Manuscripts

In 1928, Ottoman Turkish (OT), written in Arabic and Persian (Arabo-Persian), was abrogated and supplanted by modern-day Turkish written in a Latin-based script. Being able to read OT and transcribing it into the Latin-based script requires knowledge of both scripts, the different types of calligraphic styles and the learning and implementation of one of the transcription systems. Transcribing OT material has always been an integral part of scientific research in the fields of Ottoman history and literature. However, with the amount of new material becoming available through digitised collections, researchers are increasingly spending more time on transcribing, a time-consuming task. As such, little time is left for addressing actual research questions.

The *HTRising Ottoman Manuscripts* project investigates whether it is possible to develop digital tools and models that can assist researchers in transcribing OT texts, thus leaving more room for actual scientific enquiries. HTROM is the name of the transcription model developed in the *Transkribus* software which is supported by Artificial Intelligence (AI), Machine Learning (ML) methods and Handwritten Text Recognition (HTR) algorithms. The HTROM model aims to create an OT handwritten manuscripts' transcription model which deciphers Arabo-Persian OT to Latin-based OT. This paper will introduce the workflow of the HTROM model, its content, and the challenges faced during the project.

1. Workflow of the HTROM Model

1.1. Preparation of Datasets

The six-month project was divided in three phases. The first phase focused on the preparation of datasets followed by the second phase, creating datasets, whereas the third phase consisted of training the model. Preparation of datasets: As a prerequisite for preparing the datasets, two challenges were resolved. First, *Transkribus* recommends creating datasets either written by one person or a set of similar types of writing. The problem with using Arabo-Persian OT manuscripts is that we cannot always say for certain who the scribe of the texts was, meaning that we are uncertain if these materials were produced by one person or by many. Based on *Transkribus*' stipulation for using a similar type of writing, it seemed that the calligraphic style *nesih (naskhī* in Arabic and Persian), which is the most legible, would be the most appropriate style to use. Even though other writing styles were employed in OT manuscripts, the one

Diyâr, 3. Jg., 1/2022, S. 131–135, DOI: 10.5771/2625-9842-2022-1-131

most widely used is *nesih*.[1] Second, the transcription standards were determined. At present, the transcription standards used in the field are still varied.[2] Being aware of these standards, I decided to use one of the most preferred transcription standards in the field, namely *İA*.[3] The work of Muharrem Ergin, *Turkish Grammar*, on the other hand, is my reference guide regarding the peculiarities of Turkish grammar in OT.

1.2. Creating Datasets

Datasets should consist of images (mostly in .jpeg or .png format) of handwritten material and their accurate transcription texts (mostly in .docx, .pdf, and .txt format). After the images and corresponding transcriptions were prepared, the workflow was carried out as follows: a collection was created in *Transkribus* named HTROM, images were uploaded into the collection, an automatic layout segmentation was run, the problems that occurred in the layout segmentation were eliminated manually, and the transcription texts that matched with the images were entered line-by-line.

1.3. Training the Model

Following the abovementioned process, Ground Truth (GT) needs to be created before training a model. GT forms the basis for the HTR models on which to run, meaning that the image and its corresponding transcription to create GT should be 100% correct. After the creation of GT, the dataset can be used for model training. The model training process consists of two datasets: training datasets and validation datasets. These two datasets are similar in the choice of material namely the same writing style and in their GT status. In the training datasets, *Transkribus* learns how to make a transcription, while in the validation sets one tests how much it has learned. The success rate of the learning and testing phases is evaluated through a Character Error Rate (CER) which indicates the amount of transcription that could not be read correctly by the applied HTR model. For instance, a CER of 10% means that ten characters (and these are not only letters but also punctuation, spaces, etcetera) out of a hundred were not correctly identified. A good HTR model should recognize 95% of handwriting correctly, the CER should not be more than 5%.

1 Schmidt 2019, 139; Karowalski; Köse 2016, 73.

2 In the period from October 2020 to January 2021, a discursive survey on transcription standards in Arabo-Persian OT to Latin-based OT was conducted with participation of *Gesellschaft für Türkologie, Osmanistik und Türkeiforschung* (GTOT) members. Please see an analysis of the survey results at: https://www.menalib.de/community/umfragen/welche-umschrift-fuer-osmanische-handschriften/ (accessed 18 February 2022).

3 *MEB İslâm Ansiklopedisi* 1978, XVIII-XXII.

2. HTROM Datasets and Results

The datasets I have created for the model are threefold. The dataset at the centre of this model is the *Ġazavātnāme*[4]. It is written in clear and undifferentiated *nesih* style and the dataset is just below a 50,000-word count. The second is called the 'Court Register'[5] dataset with just under 9,000 words. The writing style of this dataset is also *nesih* but involves a mixture of *taʿlīḳ* and *dīvānī* styles. Although the base writing style of the HTROM model is *nesih*, this dataset aimed to set an example for different writing styles within a small amount of data. And lastly, although my aim was only to create a handwritten model, a "printed" dataset[6] was also added and trained in the model. The reason for this is that this dataset, printed with the *nesih* typography, has the potential to constitute an important visual *nesih* dataset for the model. The word count of the "printed" dataset is just under 16,000 words. The CER results of the HTROM model are as follows: the *Ġazavātnāme* dataset was 11.88%, the "Court Register" dataset was 28.16%, and the "Printed" dataset was 4.82%. The CER of the HTROM model in total was 28.32%. As mentioned above, models with a CER above 5% will not be perceived as prospering models and, in this case, only the 'printed' dataset can be considered successful. In order to understand the reasons for these high rates of the CER in the HTROM model, it is important to look at the challenges in the preparation and training process.

3. The Challenges

Within the framework of this project, the challenges encountered can be divided in two main categories: layout and navigation problems on the one hand and textual peculiarities on the other.

4 This dataset is based on *Ġazavāt-ı Hayreddin Pāşā* which is located at the Real Biblioteca del Monasterio San Lorenzo de El Escorial Madrid, cod. 1663. Aldo Gallota's article, which includes the facsimiles of the manuscript, was used: Gallota, Aldo. 1982. 'Il «Ġazavāt-ı Ḫayreddin Paša» Di Seyyid Murād'. *Studi Magrebini* Vol. XIII. Napoli: Istituto Universitario Orientale. 1–49. Besides, I owe many thanks to Ercan Akyol (Turkish Studies, University of Vienna), Hülya Çelik (Oriental and Islamic Studies, University of Ruhr, Bochum), Gisela Prochazka-Eisl (Turkish Studies, University of Vienna), Nazlı Vatansever (Turkish Studies, University of Vienna) for sharing the transcription with me.

5 This dataset is based on *İstanbul Kadı Sicilleri Galata Mahkemesi 5 Numaralı Sicil* (H. 983–984 / M. 1575– 1576). The facsimiles are taken from the open-access project of the Centre for Islamic Studies' (İSAM) Qadı Register Project [İstanbul Kadı Sicilleri]: http://www.kadisicilleri.org/ (accessed 18 February 2022).

6 This dataset is based on a novel by Suad Derviş, *Fatma'nın Günahı: Millî Roman*, printed in 1924 by Orhaniye Matbaası. This dataset was created by Julia Fröhlich (Turkish Studies, University of Vienna).

3.1. Layout and Navigation Problems

At this point, I should stress that in Digital Humanities (DH) which nowadays turn to multilingual DH, OT is a challenging case study, because it has a problem not typically found in European languages, namely text directionality. While Arabo-Persian OT is written from Right-to-Left (RTL), Latin-based OT is written from Left-to-Right (LTR). Under these circumstances, the Latin-based OT needs to be reversed in order to provide direction compatibility between Latin and Arabo-Persian scripts. To this end, *Transkribus* has a new feature[7] that solves the problem of direction and allows users to transcribe their material in an ideal direction.[8] Another challenge is that manuscripts and archival materials have different kinds of layouts, including marginal text fields, columns, frames and boxes, and superscripts. These layout structures also show different navigation properties: bottom-up, top-down, loop, etcetera. For instance, when the image of a "Court Register" or *Ġazavātnāme* is uploaded and the automatic layout segmentation has run, it is always necessary to make manual corrections because *Transkribus* encounters these layout features for the first time and has never worked with that kind of data before. For instance, for marginal text fields, the image has to be split into separate text regions; for columns, frames and boxes, the layout has to be segmented as a "table" via the *Transkribus* features. Since you cannot teach all these layout structures to *Transkribus* at once, you have to train it by increasingly adding more data. Since *Transkribus* uses AI and ML to segment and transcribe documents, it can automatically segment and transcribe if layout structures have already been seen and learnt sufficiently.

3.2. Textual Peculiarities

Another set of challenges can be evaluated with regards to textual homogeneity. For example, Latin-based OT includes 20 consonants and 8 vowel sounds, whereas the Arabo-Persian OT has 31 consonants and 3 vowels.[9] Thus, in this scenario, Latin-based OT has fewer consonants and more vowels than Arabo-Persian OT which causes many equivocal readings in Latin-based OT.[10] At this point, the main challenges arise with connection to "data". We need to train *Transkribus* with enough data to teach the complex structures of the language that it needs to transcribe. Hence training *Transkribus* with these peculiarities in mind could be compelling, but here we have

7 Thanks to the efforts of Yavuz Köse (Turkish Studies, University of Vienna) and Achim Rabus (Slavic Philology, University of Freiburg).

8 To use the 'reverse text feature for OT manuscripts, archival sources and printed texts' in *Transkribus*, please follow the steps: 1) In the main menu bar select 'install a specific version of the tool'. 2) Install the 'Snapshot 1.15.0.1. 3. And download complete package.

9 The three main long vowels in Arabic "ي ,و ,ا" correspond to the Turkish long "a, e, o, ö, u, ü, ı and i". Furthermore, و and ي can also indicate "y" and "v". (Schmidt 2019, 137)

10 For example, the word اولو can be read as *ölü, evlü, ævlu, ulu* in Turkish, and it may also be read as *ūlū* in Arabic.

to remember that *Transkribus* uses AI and ML to automatically transcribe data. Therefore, it is crucial to train *Transkribus* with more data.

4. Conclusion

In conclusion, the HTROM model has provided an investigation of the operability of OT manuscripts (and prints) with AI and ML supported by HTR technologies. At this point, important issues emerge within the scope of this project. *Transkribus* and similar text recognition platforms are designed to succeed when a consistent and vast amount of data is provided, and specified transcription standards are applied. For this reason, in the absence of a vast number of datasets and common standards and methods, it has been concluded that such technologies cannot be effective for OT manuscripts (and prints) at this point in time. For future research, however, it is very encouraging to know that with the current limited data in the HTROM model promising results are already being achieved.

Bibliography

Karolewski, Janina and Köse, Yavuz (eds.). 2018. *Wunder der erschaffenen Dinge: Osmanische Manuskripte in Hamburger Sammlungen. Wonders of Creation: Ottoman Manuscripts from Hamburg Collections*. (= *manuscript cultures 9*). Hamburg (2nd revised edition).

Schmidt, Jan. 2019. 'How to write Turkish? The Vagaries of the Arabo-Persian Script in Ottoman-Turkish Texts'. In Bondarev, Dmitry; Gori, Alessandro and Souag, Lameen (eds.). *Creating Standards: Interactions with Arabic script in 12 manuscript cultures*. Berlin, Boston: De Gruyter. 131–146.