IFLA Section
Library Services to
Multicultural Populations

IFLA Section
Local History and
Genealogy

Submitted on: 01.08.2025

The Language of the Conquerors: Opening the Lost World of the Turkic Empires for Genealogical Research

Jonathan McCollum
FamilySearch International, Salt Lake City, UT
E-mail address: jonathan.mccollum@familysearch.org

Abstract:

The imperial records of the Turkic empires of the past several centuries, initially composed in such dead languages as Ottoman and Chagatai Turkish, remain opaque for the millions of people attempting to trace their ancestry into a past in which Turks ruled over much of Europe and Asia. While these empires have receded into the pages of history, their robust records remain in the hands of thousands of state archives, libraries, private repositories, and personal collections. Archivists and librarians struggle to properly catalog and index these orthographically complex Turkic collections. Likewise, researchers without proper training in these moribund and deceased languages overlook these rich resources. Aware of the potential of Turkic records for genealogical purposes, FamilySearch International has implemented an approach to make these records accessible to non-specialists. First, this paper documents the ongoing efforts to train teams of students to index Ottoman and Chagatai Turkish manuscripts. Given the sheer size and scope of these global collections, FamilySearch also leverages machine learning to develop handwritten text recognition to assist in the indexing of Turkic record collections. In sum, this paper proposes a strategy for making all historical Turkic records accessible and useable to local and global researchers and suggests a framework for approaching similar language problems that afflict libraries around the world.

Keywords: Ottoman Turkish, Chagatai Turkish, Arabic Script, Hadwritten Text Recognition

The Language of the Conquerors: Opening the Lost World of the Turkic Empires for Genealogical Research

When the Islamic armies reached Central Asia in the late seventh century, a meeting of languages and cultures occurred that shaped the history of the Eurasian continent. Over the next millennium, Turkic peoples conquered and founded empires from Europe and North Africa in the west to China and South Asia in the east. In the process, they adopted Islamic forms of state craft and the Arabo-Persian alphabet. Starting in the early twentieth century, reformers and bureaucrats in modernizing states imposed the Roman and Cyrillic alphabets on the lion's share of Turkic speakers. Today the imperial records of the Turkic empires, initially composed in such dead languages as Ottoman and Chagatai Turkish, remain opaque for the millions of people attempting to trace their ancestry into a past in which Turks ruled over much of Europe and Asia.

While these empires have receded into the pages of history, their robust records remain in the hands of thousands of state archives, libraries, private repositories, and personal collections. Archivists and librarians struggle to properly catalog and index these orthographically complex Turkic collections. Likewise, researchers without proper training in these moribund and deceased languages overlook these rich resources. This paper outlines FamilySearch's approach to solving the problem of this "language of the conquerors." Over the course of this paper, I will outline FamilySearch's efforts to make transparent these Turkic languages with special emphasis to its continuing developments in Ottoman Turkish and its more recent experience with other Turkic languages and the records of Central Asia.

For over 6 centuries the Ottoman Empire ruled over territories in 3 continents from which over 25 current countries emerged as successor states. A conservative estimate suggests that the records of the Ottoman Empire are of genealogical significance for over 500 million people living today. The majority of these records were kept in Ottoman Turkish, a

dead language, and can be found in archives in over 30 countries. Moreover, the language of Ottoman Turkish also had significant influence over Chagatai Turkish which competent Ottoman language specialists can read. Records in Chagatai Turkish extend beyond the boundaries of the Ottoman Empire and can be found in Russian and Central Asian archives all the way into the Xinjiang Province of China.

In order to make collections in Ottoman Turkish accessible to the broader general public, FamilySearch has trained teams of Arabic speakers since 2021 to provide the necessary data entry to make searchable indexes from digital historical images. FamilySearch began with a collection of digital images derived from 84 microfilms of an Ottoman population register for the territory of Palestine. The dataset primarily includes the 1881-1917 Ottoman *nüfus* registers of the Ottoman territories of Palestine—*nüfus* translates as "population" from Ottoman Turkish. The Ottoman state, to provide accurate data on the population of the empire, instated an imperial-wide population registration system in 1881.[i] Over 450 of these registers, which recorded robust vital statistics on the population, are today preserved in the Israel State Archives.[ii] This robust data collected by the Ottoman state is of crucial importance for demographic and social research on the historical population of Palestine in the final years of the Ottoman Empire. Completed in 2024, a full index of these nüfus registers will be available on the FamilySearch website and accessible to the general public later this year.

Recognizing the limitations of this approach, FamilySearch decided to invest in the acceleration of Ottoman Turkish index production through the development and use of Handwritten Text Recognition (HTR) models for the language. Recent advancements in document recognition have achieved results on a wide variety of documents across many languages. The ability to fine-tune a model on a specific domain has improved. To this end, we provide what we believe to be the largest Arabic script line-level handwriting dataset for

Ottoman Turkish to date. We have extracted at least 199,650 lines (388,369 words) of text from Ottoman Nüfus records. These text lines derive from roughly 700 images, and a human expert has manually reviewed or labeled each one. This scale of data is nearly double the number of words used to train Ottoman handwritten text, which is available publicly via Transkribus (https://www.transkribus.org/model/ottoman-turkish-generic). After 18 months of development, our current model has achieved a 30.5% Word Error Rate, 14.6% Character Error Rate. As more data becomes available, we have just added a collection from Albanian Ottoman records, we observe an incremental increase in accuracy.

Although the records of the Ottoman Empire are of particular interest to scholars and genealogists of the eastern Mediterranean area, FamilySearch is also currently exploring the ways in which the capabilities developed for Ottoman Turkish can be leveraged for the reading of other Turkic languages. A recent indexing project was initiated for late nineteenth-century Metrical Books of the Russian Empire from the Oblast of Astrakhan for the region's Muslim inhabitants. The Russian Empire mandated the collection of vital events, such as birth, death,and marriage, in these metrical registers by local religious organizations. In the case of the these particular registers, the imams of the Tatar and Kazakh minority community recorded these life events of the local Muslim population in what they referred to as *Turki*, a dialect of Chagatai Turkish, a literary language widespread throughout Central Asia. FamilySearch is completing the keying of these documents through use of teams trained in Ottoman Turkish to explore capabilities across Turkic languages in the Arabo-Persian script. The full index of this collection will be available on FamilySearch by thte end of this year.

Preliminary assessments by FamilySearch has revealed the significance of these Ottoman and other Turkic historical documents for family history research. Individuals with ancestry from Palestine have been able to locate over two hundred ancestors from these rich collections. As these indexes become more readily available to the general public later this

year, it is our hope that the living descendants of the peoples of that lands of the great Turkic conquests will finally be able to learn about their heirtage and ancestry in these rarely investigated records. This process for accelerated index production through the use of AI language models can provide the basis for turther investigation into making documents in dead and dying languages accessible to researchers and general publice.

Thank you

Acknowledgments

Stephen Filios, FamilySearch; Chedi Hachana, InfoScribe; Ahlem Ellafi, InfoScribe

References

[i] Karpat, Kemal. "Ottoman Population Records and the Census of 1881/82-1893." *International Journal of Middle East History* 9, no. 2 (May 1978): 251. https://doi.org/10.1017/S0020743800000088

[ii] Campos, Michelle. "Placing Jerusalemites in the History of Jerusalem: The Ottoman Census (*sicil-i nüfus*) as a Historical Source." Angelos Dalachantis and Vincent Lemire, eds. *Ordinary Jerusalem, 1840-1940.* Brill, 2018. 15-28. https://doi.org/10.1163/9789004375741_004