DHQ: Digital Humanities Quarterly
2022
Volume 16 Number 2

Automated Transcription of Non-Latin Script Periodicals: A Case Study in the Ottoman Turkish Print Archive

Suphan Kirmizialtin <suphan_at_nyu_dot_edu>, NYU Abu Dhabi
David Joseph Wrisley <djw12_at_nyu_dot_edu>, NYU Abu Dhabi

Abstract

Our study discusses the automated transcription with deep learning methods of a digital newspaper collection printed in a historical language, Arabic-script Ottoman Turkish (OT), dating to the late nineteenth- and early twentieth-century. We situate OT text collections within a larger history of digitization of periodicals, underscoring special challenges faced by Arabic script languages. Our paper approaches the question of automated transcription of non-Latin script languages, such as OT, from the broader perspective of debates surrounding OCR use for historical archives. In our study with OT, we have opted for training handwritten text recognition (HTR) models that generate transcriptions in the left-to-right, Latin writing system familiar to contemporary readers of Turkish, and not, as some scholars may expect, in right-to-left Arabic script text. As a one-to-one correspondence between the writing systems of OT and modern Turkish does not exist, we also discuss approaches to transcription and the creation of ground truth and argue that the challenges faced in the training of HTR models also draw into question straightforward notions of transcription, especially where divergent writing systems are involved. Finally, we reflect on potential domain bias of HTR models in other historical languages exhibiting spatio-temporal variance as well as the significance of working between writing systems for language communities that also have experienced language reform and script change.

Introduction

Imagine that you are a researcher starting a Ph.D. project or a new book. In addition to identifying a topic, conducting a literature review, and fine-tuning your research question, you will also need to navigate an extensive archive of millions of documents, large parts of which are yet to be cataloged; some collections have been opened to researchers only in recent decades and the digitization of the archive is still at its earliest stages, making a keyword searchable corpus unavailable.[1] This situation is the reality of historical research in the Ottoman archives at the time of writing this article. In addition to the challenges of accessibility of the archive, there is also a fundamental question of script literacy; Ottoman Turkish (OT) is written in Arabic script that most Turkish speakers today cannot read. To what scholarly apparatus would historians of differing skill levels turn for doing their work — annotating and transcribing for recall, or simply combing through the most relevant sections of the archive — when these documents are printed in a writing system that is foreign to their contemporary script habitus? It might seem that we are speaking about a many century's old historical situation, but in the case of OT, archival documents created as little as one hundred years ago are not accessible to speakers of modern Turkish who have not received a specialized education in OT. 1

In Istanbul, where the central archive of the Ottoman State and several of the most important research libraries in the field of Ottoman studies are located, there is an informal market of hireable transcribers to meet the demand for transcription of OT historical documents to Latin script modern Turkish. Consisting mainly of students and graduates of departments of history and Turkish literature, the informal labor of this market offers its services to scholars who do not have the time required to transcribe their documents. In the absence of extensive digitization of archives, these transcribers also provide on-site research services for their individual clients, many of whom do not live, or work regularly, in Istanbul.[2] 2

As the digitization of Ottoman-era archives is underway,[3] our article concerns one approach to increasing searchable (and computational) access to them. We discuss experiments that we carried out by automatically transcribing a small amount of the holdings of the larger Ottoman archive: two Ottoman Turkish periodical series from the Hakkı Tarık Us (HTU) Digital Repository. With some 1,500 periodicals and an estimated number of 400,000 pages, the HTU online archive is perhaps the most comprehensive digital collection of the Ottoman Turkish press from the 1840s to the 1920s. Originally the private library of Mr. Hakkı Tarık Us (1889-1956), the digitization (2003-2010) of this collection was a collaboration between Beyazıt State Library in Istanbul and Tokyo University of Foreign Studies. Despite being online for a little more than a decade, searchability of the HTU documents is still unavailable. 3

Whereas some newspaper mass digitization projects in the world were “launched under circumstances in which searchability was not a priority” [Salmi 2021, 45],[4] it is not known to us if this was the case with the HTU corpus or whether OCR-based text creation was simply technically impossible at the time, or both. In the mid- to late-2000s, the HTU platform digitized images in DjVu format — a format that offers a corresponding text layer for each image — the choice of which presumably was to leverage the linguistic skills of 4 literate Ottomanists to transcribe the documents into modern Turkish.[5] Unfortunately, this format was already outdated at the time [Baykal 2011, 206–8]. Although the file format would have allowed for the creation of a text layer for the HTU as early as 2010, the promise of a text searchable OT corpus remains to this day an unfulfilled one.[6]

You do not need to be an Ottomanist, of course, to confront such a problem with unsearchable digitized archives; the same could be said of other digital collections in many other world languages: medieval manuscripts, letter collections, miscellanea archives. It is important to note that both digitization and OCR are historical processes with differing degrees of technical development and variable success rates across world languages. In the case of mass digitization of periodicals in the nineteenth century US context, we have been encouraged to “[take] dirty OCR seriously” [Cordell 2017]. This is not only because early digitization and OCR methods did not produce high quality text and it is unlikely that either process will be carried out again, but also because no matter what methods of studying imperfect corpora are devised in the future, understanding the layered histories of digitization and text creation remains a key element of digital historical inquiry. 5

For languages written in Arabic script from right to left, OCR is a complex process in which research has made progress, but is yet to become fully efficient [Romanov et al. 2017].[7] As OCR is an algorithmic process in computer vision that transforms images of letters and words — by many iterations of guessing, ranking, and even human intervention and correction — into digital text, the more intervention that is required the more labor-intensive the process of text creation becomes. Additionally, the more the features of the language that one works with differ from the Latin-based scripts on which OCR was first developed, the more development and research into new systems are required. At the time of writing this article, however, the situation of historical OCR has already started to change radically. This development is mainly thanks to the increased accessibility of user-centered neural systems for transcribing handwritten archives, and by extension for printed cursive scripts, such as Arabic or Ottoman Turkish, in which it can be difficult for a computer to determine where one letter ends and another begins. Some of the current generation neural systems’ text recognition abilities go beyond simple character recognition, which has been the main challenge OCR faces with handwritten texts and cursive scripts such as Arabic — due to the fact that some letters are connected to each other — to allow for larger chunks of the text, such as words or whole lines, to be recognized as units. In this paper, we adopt one such system for automating the transcription of printed OT texts, the platform known as Transkribus, and we argue that not only are new approaches to OCR beneficial in opening up historical archives, but also that with those approaches comes an evolving role for human effort and new, concomitant challenges for working with such machine-created text. 6

Our choice to apply handwritten text recognition (HTR) technologies to non-Latin script digitized historical periodicals is an intentional one, with the purpose of creating a workable access point to Ottoman historical materials. The long term goal of our research is to generate general HTR models for text creation from this periodicals collection that will facilitate its use in computational research. In this article we discuss some of the initial challenges we have faced, as well as some of the promising results that we have achieved. 7

As interest turns to a multilingual DH, the accessibility of archives and modes of participatory text creation, we find the Ottoman Turkish example to be a thought-provoking case study since it links problems which the GLAM sector has traditionally confronted with other problems not typically found in Western language archives: text directionality and script, in particular. 8

Part 1 - Background & Historical Context

Historical Context

How we encode Ottoman Turkish (OT), the Arabic-script language of administration and high culture in the Ottoman Empire between the fourteenth and early twentieth centuries, presents a significant challenge for OCR, not only from a computer vision perspective, but also importantly from a linguistic one.[8] While Turkish syntax and vocabulary formed the main framework of the language, OT borrowed extensively from Arabic and Farsi in terms of both loan words and grammatical structures. With the political upheaval that ensued the demise of the Ottoman state and the creation of the Turkish Republic in the wake of WWI, OT was outlawed for public use in 1928 and supplanted by a “Turkified”[9] version of the language written from left to right (LTR) in Latin script [Lewis 2002]. 9

It would be incorrect to say, however, that the Ottoman archive was sealed in 1928. Physical access to it was possible, although limited [Aktaş and Halaçoğlu 1992], but within a generation or so, the Arabic-script literacy required to use the archives declined. The first generation of historians who had been educated in Republican schools no longer had the linguistic skills of native speakers of OT to carry out research in the archives. Coupled with the ideological choice of the political leadership to shun the Islamic imperial heritage of the newly created nation-state of Turkey, the so-called Turkish language reform created deep gaps in the scholarly infrastructure of the early Republican period, with most serious academic discussions on Ottoman history and language shifting to scholarly circles outside of Turkey.[10] 10

Alphabet change had deeper cultural implications that went well beyond its ramifications for the scholarly establishment. Today, any Turkish citizen who wants to access their pre-1928 cultural heritage needs either to receive specialized education in Ottoman Turkish or to read the transcribed and edited versions of the select OT texts available in modern Turkish. Reading OT documents and books transcribed into Latin script has become the norm. There is, therefore, a real demand for transcription, a demand that, to this day, has been only partially satisfied by human transcription.[11] Furthermore, the language reform has not only created a barrier between one nation and its historical, cultural documentation but also, as OT was the former administrative language of the successor states of the Ottoman Empire, both language and script persist as barriers in accessing this larger imperial heritage by the broader scholarly 11 community.

Archives and Access

For many languages of the world that have not traditionally been represented in the textual digital humanities, in particular historical languages, the obstacles of being “at the table” are very real.[12] Research materials and infrastructure for these languages are absent or in outdated formats, digitization is either partial or piecemeal, and accessing those digital resources can be a daunting task.[13] In the larger case of Ottoman Turkish, there is a lack not only of sufficient machine-readable corpora but also of natural language processing (NLP) tools designed to work with the multi-layered, multi-epoch complexities of this historical language.[14] When the archive is so abundant, yet still so inaccessible, the eagerness to open up research fields to exploration with digital methods grows every year.[15] 12

We believe that pursuing digital methods in some non-Latin script research fields, such as Ottoman studies, is not so much a question of resistance to, or distrust of, the methods as it is of the need for workable starting points, given the historical and linguistic complexity of these languages.[16] Furthermore, the desire for high-quality text creation at scale can stand at odds with the capacity and affordances of the platforms at our disposal to do so. Deferring the question of what kind of corpus might be representative of late Ottoman society for now, and building on the decades-long efforts to create open digital collections for reuse and high order analysis, what has come to be known as a “collections as data” approach [Padilla et al. 2019], we turn to automatic transcription of printed Ottoman periodicals with the *Transkribus* platform using deep learning methods. Whereas traditional Ottomanist communities have relied on workarounds for research that depend on informal labor for gaining access to archival materials, our case study argues that technical workarounds within developing humanities research infrastructure can, and do, spark innovation in the interest of linguistic inclusion. 13

New Modes of Text Creation

At the intersection of archives and the textual digital humanities is the central question of access. Initiatives such as the EEBO-Text Creation Partnership have opened exciting windows for a wide range of scholarly projects grounded in access to historical print sources [Martin 2007]. Its presence can be felt all over the scholarship in pre-modern English, and yet its mode of creation — outsourced double typing of texts in Asian countries — casts an uncomfortable, but often overlooked, shadow over the material conditions of its creation. Scholarly communities working on languages printed in the Arabic script (Arabic, Persian, Urdu, etc.) have been slower to arrive at robust textual digital humanities for multiple reasons, not the least of which has been the difficulties of converting the cursive-like printing of these languages into a machine-readable format [Ghorbaninejad et al. 2022]. We anticipate that examples in crowd-transcription, such as Transcribe Bentham (UCL), or more recent initiatives spearheaded within the GLAM sector (Library of Congress, Europeana, the Newberry Library, the Getty Museum, Zooinverse), and the democratized platforms for engaging a larger public in doing so, will slowly make their way to cultural heritage institutions in societies using RTL (right to left) scripts. What can be done in situations, however, such as OT, when the script of the archive is no longer native for the majority of the public? Our challenge has been not only that of automatic transcription but also rendering the output of this process sufficiently legible for the LTR (left to right) Turkish reader today 14

Whereas some formal modeling has been done with OT (take for example, the markup of multilingual versions of the Ottoman Constitution in TEI XML including OT and Arabic versions [Grallert 2019], our intervention focuses at a more rudimentary stage of new text creation; it is ground research that aims to produce someone of the first, if not the first, searchable text corpus of OT. While this situation might be surprising to digital humanists working in European, or even large non-Western languages, for whom the existence of some kind of corpus has been a given for many decades, the lack of basic keyword searchability is the reality of text-based research in Ottoman studies. 15

One key technological development that opened a promising pathway for the creation of full-text searchable corpora for both printed and handwritten documents in Arabic script is the recent advances in pattern recognition with HTR. Unlike many OCR systems, which operate at the level of individual characters, HTR works at line level and, as a result, yields higher accuracy rates in recognizing cursive scripts such as Arabic.[17] Taking advantage of this particular characteristic of the HTR technology, we have carried out our automated transcription experiments for printed OT periodicals with HTR. The larger question of the full Ottoman archive, most of which is manuscript, however, still sits in the background of this article. Turn of the century printed materials in Ottoman Turkish (or in Arabic or Urdu, for matter), do exist on a continuum with handwritten materials, due to the fact that they are printed in a “cursive-like” manner. It is too simplistic, therefore, to oppose print and manuscript cultures in OT, especially because Ottoman master printers did, in fact,strive to reproduce the calligraphic traditions of the manuscript culture in the printed page [Özkal 2018, 71]. 16

Why Periodicals

After decades of digitization efforts in libraries and archives worldwide, the number of periodicals that are available to today’s researchers has increased dramatically. For the historian, the serial periodical offers special insight into cultural debates as they unfold throughout time. Nineteenth-century periodicals are the place of remarkable emergence of public discourse in many regions of the world, including the multi-ethnic and multi-religious Ottoman Empire.[18] They also provide a valuable opportunity to reconsider this cultural sphere in a period marked by rapidly evolving linguistic usage in the face of political change [Mussell 2012]. Significant advances have been made globally in the accessibility of cultural collections through digitization and the implementation of full-text 17 searchability and vibrant debates have opened up in digital history on how to use this digitized archive to the fullest extent.

There are a number of features of historical periodicals that make them complex for digital analysis in any language: OCR quality, the commercial nature of digitization collections, complex page layouts, the presence of images within the text, and particularities of their digital remediation [Cordell 2016] [Nicholson 2013] [Clifford et al. 2019]. Yet again, the challenge of dealing with Arabic-script periodicals is an even greater one. Whereas many periodicals have been scanned and brought together in digital archives in these languages, allowing researchers to access them at a distance, the texts of such periodicals are still messy and not full text searchable, limiting users to slow reading and requiring computer-assisted modes of analysis to adapt to the text quality.[19] Unsurprisingly, the few studies that have used various Ottoman periodicals collections as their main primary source have so far been qualitative and of limited scope [Baykal 2019] [Çakır 1994]. The sheer volume of the available material, as well as the outdated formats and piecemeal nature of their digitization, renders distant reading of these sources unfeasible. These collections are not yet data. 18

Part 2 - Transcription and Its Discontents

Audience & A Pragmatic Approach

Mastering OT today is a demanding task. People who undertake learning the language/script nowadays do so in order to read historical manuscripts and documents. Among those who are proficient in OT, it is not common practice to take notes in Arabic script, but rather in what is the contemporary script habitus of Turkish historians as well as that of the majority of non-Turkish Ottomanists: the LTR Latin alphabet [Ghorbaninejad et al. 2022]; researchers in the field simply tend to annotate their documents in Latin script. Owing to script differences and language change, annotation style can vary from person to person, with some researchers even transcribing their documents fully into modern Turkish. Furthermore, in Turkey, editions of OT works are rarely published in Arabic script, or if they are, they are accompanied by a facing-page transcription. As such, the OT script in contemporary Turkey could be considered a “dead” script. Being able to automate the transcription of OT to Latin writing system of modern Turkish, therefore, is also a question of creating a knowledge infrastructure that corresponds to a practical issue of contemporary script literacy. 19

The usage of terms for discussing the passages between languages and scripts, from OT to MT, have varied significantly. Our Appendix (Key Concepts) details how we use the terms *transliteration*, *transcription* and *romanization*. These three terms all imply some kind of passage between writing systems and they require careful consideration if we are to adopt a critical approach to text creation of OT printed material. We have come to realize that our approach to automated transcription with deep learning methods needs to be a pragmatic one that balances a number of concerns, such as paying enough attention to the requirements of the HTR engine so that the model is effective while also being consistent enough so that the digital reuse of the resultant data is the highest quality possible, at the same time keeping in mind the needs of the human reader. We are aware that, due to the particularities of OT, which we discuss in the following section about the encoding of grammatical information, it will probably never be possible to build an automated transcription system that flawlessly bridges the gap between the original OT documents and their transcriptions. Our goal in applying neural models of automated transcription to OT printed materials is, instead, the creation of a “decent” text, a workable, “good-enough” approach to transcription, a goal echoed by other historians working in digitally under-resourced languages [Rabus 2020]. 20

No doubt this will be disappointing to our philologically inclined colleagues, but it is the scale and the modes of downstream analysis that encourage us to adopt this position. We do not see the text creation process as a failure from the outset. Instead, working with automatic transcription has encouraged us to think about the transcription process itself as an initial access point to the historical materials. Our goal is to produce a result that is sufficiently accurate and usable by our target reader (the historian scholar engaged in professional reading), and we recognize that the end result will never look fully like human-transcribed texts [Siemens et al. 2009]. From this starting point, it follows that what is necessary to use HTR methods to bring new kinds of access to a larger number of OT texts is a critical conversation about what is commonly called transcription. 21

We are particularly drawn to the idea of crowd transcription among Turkish speakers that could transform the existing informal economy of OT transcribers into a participatory community of shareable, open knowledge production. This will not be possible, however, without a larger scholarly debate about transcription and transliteration that departs from contemporary practices of inconsistent romanization in favor of one that is more harmonized with the ways that algorithmic systems work. Pilot studies such as ours with automated transcription technologies are in a position to launch this debate. The question resembles one that is asked regularly in human-computer interaction research: how do we design interfaces and input mechanisms for human knowledge so that we can optimize the results of computational processes? The answers to such questions can only emerge, we believe, through the development of a user community that tests, evaluates, and validates such transcription norms in different textual domains. 22

Challenges of Transcribing OT and Its Implications for HTR Processing

Ottoman Turkish, in its classical form, is a patchwork of Turkish, Arabic, and Farsi vocabulary and grammar. The complexity of the language is compounded by the challenges of the romanization of Arabic script in which OT was written [Halpern 2007]. Elezar Birnbaum summarizes the complex character of OT orthography as such: 23

> the Ottoman Turkish writing system is only an *indication* of pronunciation rather than a representation of it. It

> incorporates two quite different methods of indicating sounds, which are ill-joined into one system... On the one hand, Arabic and Persian spelling conventions are preserved almost intact for all Ottoman words derived from those languages, while completely different conventions, rarely explicitly formulated and still more rarely consistently applied, even in a single piece of writing, hold the field for words of Turkish and other origins, and for Turkish affixes to Arabic and Persian loan words. [Birnbaum 1967, 123]

In contrast to the complicated nature of OT orthography, where several sounds in the language are omitted in writing and the correct pronunciation of a word is contingent upon the reader’s literacy and linguistic background, Modern Turkish (MT) spelling is unequivocal and highly phonetic. This crucial difference between the two writing systems renders a one-to-one, diplomatic transliteration scheme from OT to MT unattainable.[20] This situation also partially accounts for the fact that to this day there is no scholarly consensus on how to, if at all, transcribe OT to MT.[21] However, if we remember the anecdote with which we began this article about the informal labor market of transcribers, despite the lack of a transcription scheme based on consensus,[22] OT text transcribed to MT is often preferred over original material as a matter of practicality. 24

To further break down the factors that hamper character-accurate, diplomatic transcription from OT script to MT: There are only three vowel signs in Arabic (ى , و ,ا) , which are often not represented but only implied in writing. MT, on the other hand, has eight vowels and the written word is always fully vocalized. Moreover, letters (ى , و ,ا) are polyphonic, i.e., they correspond to more than one sound in MT script. For example, Arabic letter (و) may correspond to any of the following characters in MT: (v, o , ö, u , ü). Several of the OT consonants are polyphonic as well. Depending on the textual context, they may be substituted by several different letters in the MT alphabet (Table 1). 25

Ottoman Turkish	Modern Turkish
ا	a, e
ض	d, z
ك	k, g, ğ, n
و	v, o, u, ő,
ه	h, e, a
ى	y, a, ı, i

Table 1. Character correspondence chart of polyphonic OT letters

Finally, the phonological developments of Turkish language over the centuries and the wide variation in its pronunciation across the vast geography of the Ottoman Empire as well as across its socio-economic and ethnic groups at any given point in time, created a gap between its written and spoken forms. This means that many OT words may be pronounced in and transcribed to modern Turkish in several different forms, all of which might be considered accurate. Any endeavor to produce a “correct” transcription of OT in MT script is, in reality, an attempt to recreate what we imagine was the pronunciation of the Istanbul elite at the time, which, of course, is lost to us in its spoken form [Anhegger 1988]. 26

One of the key steps in HTR is the creation of ground truth transcriptions corresponding to segmented parts of the digitized image. Therefore, the above three points have important practical implications for how we carry out transcriptions. As an Ottoman Turkish text is transcribed into Latin script, linguistic information is added to its romanized version by way of vocalizing it and rendering it in one of its several alternative pronunciations, which might not necessarily correspond to the written form. The presence of polyphonic letters necessitates another layer of choices to be made by the transcriber. It is not unreasonable to regard the reading and transcribing of OT as an art rather than a science, a quality that does match the exigencies of the current generation of computational tools, including HTR. The unavoidable addition of linguistic information to the transcribed material is especially problematic when training a neural-network based transcription system such as Transkribus, where one-to-one correspondence between the original text and its transcription is essential. It also results in the introduction of a significant degree of bias to the language data that is provided by the transcriber for the neural network. We will discuss these issues in more detail in Part 3. 27

Previous Work on the Automated Transcription of OT

There are currently two other projects that focus on the automated transcription of Ottoman Turkish [Korkut 2019] [Ergişi and Şahin 2017]. Both are similar in methodology and scope, employing morphological analysis and lexicon-based approaches to romanize OT script. Korkut and Ergiş et al.'s studies take advantage of the fact that Turkish is an agglutinative language, and that the stems of Turkish words do not inflect when combined with affixes. Once an OT word is stripped of its affixes through morphological parsing, the stem can be looked up in a dictionary for possible matches in modern Turkish.[23] At the next step, these two romanization models reconstruct the stripped affixes in modern Turkish according to its spelling and pronunciation rules and offer potential transcriptions for the original OT word in Latin script. Both projects work with nineteenth and early twentieth century OT vocabulary, presumably because dictionaries with relatively extensive content are more readily available for this period than for previous centuries. 28

An important distinction between these romanization schemes and our approach using Transkribus is their rule-based approach to 29 automated transcription. Their systems depend on conventional linguistic knowledge (dictionaries and grammatical rules) whereas the HTR offers a statistical, brute force technique that utilizes large data sets for pattern recognition without any explicit reliance on linguistic information. While rule-based approaches might conceivably produce a higher precision output, Korkut and Ergişi et al.'s methods still have important shortcomings. In the absence of a rigorous OCR system for OT, it is not feasible to transcribe longer texts in their interfaces. They can also only romanize those words that are already in their databases. The scalability of these platforms depends on the creation of comprehensive OT to MT word-databases/dictionaries, which need to be significantly more substantial in content than anything that is currently available,[24] and the development of effective OCR systems for the Arabic script to convert OT texts to machine-readable format. Even if these two conditions are met, however, the automated transcription of large corpora in these platforms might still be prohibitively slow due to the time needed for dictionary searches to match the OT stems to their MT counterparts.[25]

HTR for Automated Transcription

In principle, an HTR engine can be taught to recognize any writing style from any language. To start, the system needs to be provided with training data to identify the specific patterns in a given body of text. In order to produce a reliable transcription model for the rest of the corpus, it is imperative to create an accurate "ground truth," a literal and exact transcription of a small body of sample text, that is representative of the corpus in question. 30

While the absence of linguistic rules[26] in HTR training does create certain shortcomings in the quality of the transcription, we prefer the approach for its practicality, time-efficiency, and scalability. HTR has proven to be most effective with large corpora and after the initial investment of time for creating the ground truth, it is possible to automatically transcribe hundreds of pages within a matter of hours. The system is also flexible and accepts changes in parameters. That is, with a large and diverse enough training set,[27] it is possible to generate general HTR models that will work for corpora produced with different handwriting styles or typefaces.[28] However, it is also important to highlight here that what the creation of "general models" entail is not yet well defined within the Transkribus environment. What combination of materials or which order or re-training would produce best results appears to depend on the size and the nature of the corpora and requires, at the moment, a trial and error approach. 31

One advantage of using HTR for OT is its relatively higher accuracy rate at segmenting and recognizing cursive and connected writing styles. Unlike most OCR based systems, which operate at letter level (for both segmentation and recognition) and are, therefore, not very efficient for connected scripts such as Arabic, HTR works at line level and recognizes characters in the context of words and lines rather than as individual units. 32

This renders it an excellent tool for languages written with the Arabic alphabet. 33

Part 3 - Our Experiment

The Procedure

The *Transkribus* platform produces automated transcriptions of text collections by using customized HTR models trained on partial transcriptions of the corpus. As we have argued at the beginning of Part 2, tools are often not a perfect fit for the historical source material we possess and the results we obtain teach us a significant amount about our objects of study. In this study, after significant trial and error, we ultimately reverse engineered our research workflow to address critical questions of both the object of study (the periodicals and their language) and the method itself. 34

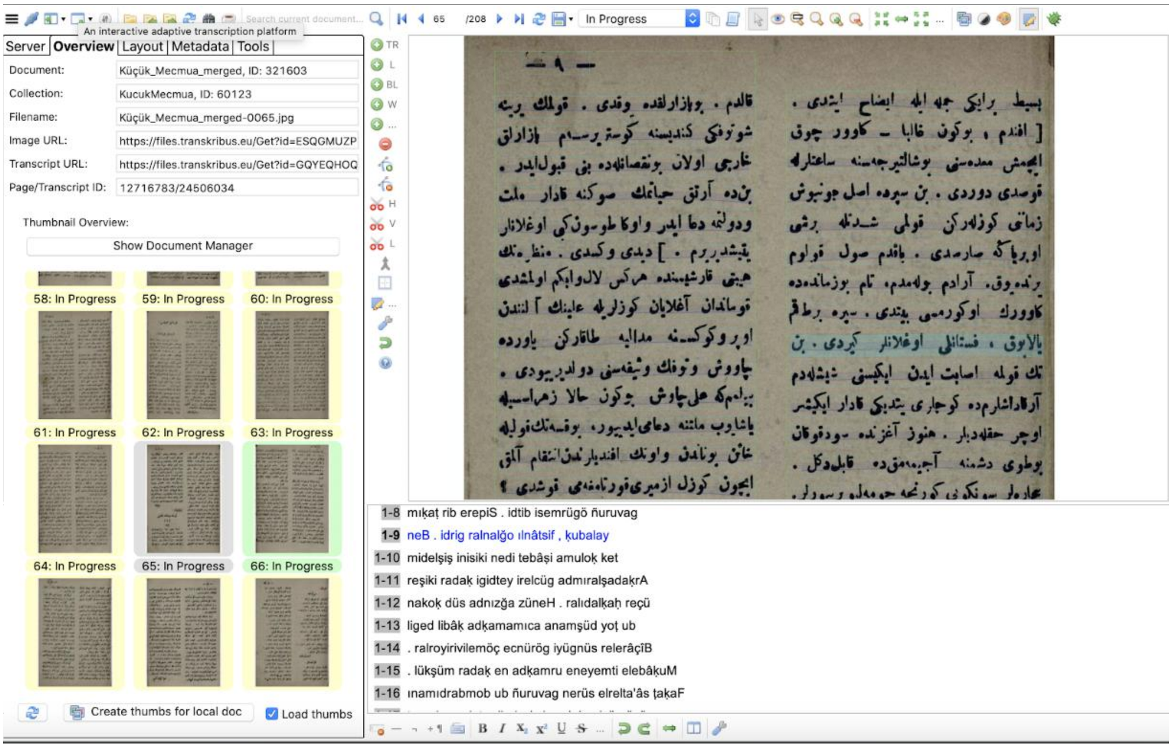
We created two sets of training data and corresponding HTR models for two periodicals[29] from the HTU online collection, the above-mentioned digitized collection of OT periodicals. The first one of these publications is the *Ahali* newspaper, printed in Bulgaria between 1906 and 1907. For this publication, we adhered to a "loose transcription"[30] scheme that only indicates long OT vowels and does not use any other diacritical marks. The second periodical for which we generated a training set is *Küçük Mecmua*, published by the leading ideologue of Turkism at the time, Ziya Gökalp, between 1922 and 1923 in Diyarbakir, Turkey. This publication has a significantly "Turkified" vocabulary and a relatively standardized orthography compared to other OT publications. We applied the *Islam Ansiklopedisi* (IA) transcription scheme[31] for *Küçük Mecmua*, which uses diacritical marks to differentiate between polyphonic characters in the OT and MT alphabets. Our assumption was that the regularized orthography and the detailed transcription system might help to reduce ambiguity in the transcription text for this publication, providing us with a "cleaner" starting point for the HTR. 35

The HTR neural network employed "learns" to recognize a script by matching the characters, strings or groups of words in the image files with their counterparts in the transcription. Adhering to a diplomatic transcription scheme, therefore, is recommended for creating reliable HTR models. This basic principle complicates the workflow for OT in *Transkribus* significantly for two reasons; first, as we mentioned above, the lack of one-to-one mapping between OT and MT scripts, and second, the opposite directionality of Ottoman Turkish documents and their Latin script transcriptions. 36

To address the first issue and minimize the inconsistencies in the training data, in our transcriptions we prioritized character accuracy over both the correct pronunciation of the OT words and some of the grammatical rules of modern Turkish.[32] We also opted for maintaining spelling mistakes and typos in the transcription text exactly as they appeared in the original pages, rather than correcting them. In other words, although these decisions seem counterintuitive to the experienced transcriber, working with an HTR system for 37 a RTL historical language required us to unlearn some of the conventions of our “practical” research habits for the purposes of transcription for creating training data. By avoiding usual scholarly practices of transcription and giving the algorithm what it needs to do its task, we hoped to optimize its performance.

The first step of the HTR workflow involved transferring the images of the documents into the platform, followed by the automated segmentation and layout analysis for the recognition of text regions and baselines in these documents. 38

The opposing directionalities of OT and MT come into play at the next step of the workflow, in which images of the documents are linked to corresponding transcriptions. While the *Transkribus* platform does support RTL languages, it does not allow connecting RTL images to LTR transcription text. To bypass this problem, we devised a workaround and reversed the direction of our transcription text in modern Turkish from LTR to RTL (Fig 1).[33] To this end, we wrote a short script in Python and ran it for the plain text files of our transcriptions.[34] 39



Document: Küçük_Mecmua_merged, ID: 321603
Collection: KucukMecmua, ID: 60123
Filename: Küçük_Mecmua_merged-0065.jpg
Image URL: https://files.transkribus.eu/Get?id=ESQGMUZP
Transcript URL: https://files.transkribus.eu/Get?id=GQYEQHOQ
Page/Transcript ID: 12716783/24506034

1-8 mıḳaṭ rib erepiS . idtib isemrügö ñuruvag
1-9 neB . idrig ralnalğo ılnâtsif , ḳubalay
1-10 midelşiş inisiki nedi tebâşi amuloḳ ket
1-11 reşiki radaḳ igidtey irelcüg admıralşadaḳrA
1-12 nakoḳ düs adnızğa züneH . ralıdalḳaḥ reçü
1-13 liged libâḳ adḳamamıca anamşüd yoṭ ub
1-14 . ralroyirivilemöç ecnürög iyügnüs relerâçîB
1-15 . lükşüm radaḳ en adḳamru eneyemti elebâḳuM
1-16 ınamıdrabmob ub ñuruvag nerüs elrelta'âs ṭaḳaF

Figure 1. A snapshot from *Transkribus* interface demonstrating the transcription process. Note the left-justified, yet reversed, Latin-alphabet transcription (center bottom) of the OT text (top right). The OT text displayed in the canvas tool is from *Küçük Mecmua*.

After the careful preparation of the two sets of ground truth, sixty pages each, we generated our first HTR models for *Ahali* and *Küçük Mecmua*, yielding 9.84% and 9.69% Character Error Rates (CER) respectively. In the second phase of the process, we retrained the models by manually correcting errors in the automated transcription and expanding each set of ground truth by an additional twenty pages. We also used our initial HTR models as “base models” in the second round of training. The manual correction of the automated transcription output, the expansion of the training set by twenty pages, and the use of base models to boost the accuracy rate of the later models have dropped the CER for *Ahali* newspaper to 5.15 % and for *Küçük Mecmua* to 7.86%. (Table 1) As it is clear from these numbers, the IA transcription scheme did not provide an advantage over the less detailed transcription system in terms of CER. This might, however, be due to the relatively modest size of the training data and a larger sample set might offer a better comparison amongst the various transcription models. 40

Finally, for text analysis purposes or human reading, the automatically transcribed texts are exported from the platform and their directionality is reversed to LTR. 41

	First HTR Training Experiment			**Second HTR Training Experiment**		
	Number of lines	**Number of words**	**CER**	**Number of lines**	**Number of words**	**CER**
Ahali	3268	20572	9.84%	4931	31390	5.15%
Küçük Mecmua	1928	12323	9.69%	2580	16326	7.86%

Table 2. The ground truths and the CERs for the first two HTR experiments with *Ahali* and *Küçük Mecmua*.

Cross-Domain Applicability

42 The *Transkribus* platform allows the repurposing of HTR models to re-train the neural network for recognizing different corpora as well as community sharing of those models. In future work, we hope to take advantage of this function to develop general HTR model(s) and evaluate to what extent they might work for the entirety of the HTU digital repository. To this end, we will focus on creating additional training sets from the holdings of this digital collection.

43 For the HTR technology we have at our disposal, the HTU collection of late Ottoman periodicals is an ideal corpus. By the late 1870s the typeface for printing had been standardized and a set of printing conventions for OT periodicals established.[35] Furthermore, the OT press from this period tended to cover similar topics and promote comparable agendas; as a result, they contain similar vocabulary, terminology, and named entities. All of these factors, along with the possibility of using a “language model” function,[36] contribute to the reduction of error rate in text recognition. It is this uniformity of this printed corpus that, we hope, will allow the HTR model to generalize across the entire HTU corpus with acceptable accuracy. It should be said though that for scholars in the field who intend to expand HTR to manuscript archives, finding a comparable corpus uniformity will no doubt be a challenge.

44 To test how the HTR model would function across different textual domains, we ran the *Ahali* HTR model for three other periodicals from the HTU collection. Attention was paid to include in the sample pool publications from different time periods with distinct content and agendas. The CER for random pages from these periodicals are as follows (Table 2):

Name of Publication	Subject	Date of Publication	CER
Tasvir-i Efkar	politics	1863	8.54%
Mekteb	education and literature	1891	10.26%
Kadınlar Dünyası	feminism	1914	6.22%

Table 3. The newspapers used in the cross-domain applicability experiment

45 It is important to note here that the CERs listed above were attained without any training data from these publications or any correction and retraining of the model. This is an encouraging indication suggesting that with the expansion of our ground truth, it is, in fact, achievable to create viable general HTR model(s) for the entire online collection of the HTU library.

Discussion of the Results

46 The primary challenge HTR faces when working with OT seems to be not character recognition, but rather identifying OT words in their proper Modern Turkish (MT) forms. As discussed earlier, the absence of one-to-one correspondence between OT and MT alphabets leads to multiple possible transcription outputs. Even when the HTR correctly identifies the characters in an OT word, accurate reading still depends on precise vocalization as well as context-appropriate pronunciation of the polyphonic letters. These, in turn, necessitate prior linguistic information, which is not taken into account during HTR training. In other words, no deep learning tool would be able to complete the desired task of a flawless MT transcription.

47 The recently implemented Language Models (LM)[37] function in *Transkribus* appears to partially compensate for the absence of rule-based text recognition in the platform. When supported with LM, HTR renders Ottoman Turkish words in their context-accurate form as they are defined in the training data. This, in turn, reduces Character and Word Error Rates.[38] For example, the OT word (عمله), which appears in our training set for the *Kucuk Mecmua* periodical, has two possible readings, both of which are character-accurate, “imle” and “amele” whereas only the latter is a meaningful word unit in OT, the equivalent of the English word “worker.” In our experiment, text recognition without Language Model transcribed the word as “imle.” When we ran the HTR with the LM, however, the word was correctly recognized as “amele,” the version that was defined in the training data. In addition, the LM appears to be able to identify character sequences that are not in the training data and even “assign a high probability to an inflected form without ever seeing it.” [Strauss et al. 2018, 6]. This contributes to the accurate transcription of words that are not in the training set. While this is far from being the perfect solution for working around the ambiguities of OT, and a significant departure from rule-based NLP systems, it is still a step towards HTR-created, human-readable text.

48 Moreover, the neural network appears to be able to easily detect oft-repeated words and word pairs (such as Farsi possessive constructions, which are particularly challenging to identify in OT texts), presumably as a result of line-level operations of HTR. Consequently, it is reasonable to expect that the system will produce higher accuracy rates for document collections with formulaic language or corpora dealing with similar subjects and vocabulary. It is tempting to imagine a future of not only “language-blind” HTR, but also one that is language agnostic, or even multilingual adaptive, that can learn the particular linguistic strata or multilingual usage in a given corpus. It is still important to underline, however, that a more comprehensive solution to the problem of multiple possible outputs in OT transcription might be the integration of dictionaries/word indices into the system.[39] In the absence of such an option, the best alternative would seem to be providing the HTR and LM with a larger sample of language data to account for a greater degree of orthographical variance.

49 Stemming from our experience with generating HTR models for different publications, we infer that the best approach to extensive collections, such as HTU, is to create date/period specific HTR models for subsets of those corpora. The late nineteenth century was a time of accelerated linguistic development and experimentation for Ottoman Turkish, both in its spoken and written forms. Therefore, temporal (and/or genre) proximity of publications in sub-groups is likely to contribute to the creation of more accurate HTR

models. This would, presumably, also be the most practical approach to the automated transcription of the Ottoman manuscript archive, which is considerably more sizable than the print documents and publications in this language.

Finally, our experiment with HTR for OT print material affirmed the widely accepted conceptualization of transcription as a biased process and that machine facilitated transcription is no exception to this. Both HTR and LM rely on the language data provided by the transcriber; they, in turn, reproduce the transcriber's linguistic bias in the output. As aptly described by Alpert-Abrams: 50

> The machine-recognition of printed characters is a historically charged event, in which the system and its data conspire to embed cultural biases in the output, or to affix them as supplementary information hidden behind the screen. [Alpert-Abrams 2016]

In the case of the automated transcription of OT, the imposition of a uniform transcription scheme on the language obscures regional, temporal, and ethnic varieties in its pronunciations and creates an artificially homogeneous outlook for OT which does not reflect the historical reality of the language. For our project with OT periodicals, however, this point is less of a concern because these publications tended to, indeed, adhere to the standards of "high Ottoman," that is, to an OT spoken and written by the educated elites of the Empire. Still, the language information we, as the users of the automated transcription platform, impose on the neural network does not only affect the quality and readability of the output but also has important downstream implications from keyword searches and Named Entity tagging to other NLP applications to OT corpora. Corpora creation has for a long time been viewed as a complex political process, and with deep learning and computationally intensive models, this is no less the case.

Conclusion

In our paper we have discussed ongoing research into text creation via automatic transcription in order to bring OT into dialogue with analytical modes of the textual digital humanities. The limited development and application of OCR methods compatible with Arabic script languages has no doubt been one of the rate determining steps, not only in the development of textual corpora, but also in the attendant language technologies that support the use of such corpora. To wish — in 2022 — for basic keyword search capacity within digitized media of the nineteenth and early twentieth century might seem to some as a rather old fashioned request, and yet for Ottomanists, it is an actual one. Limited access to digitized versions of the archive, a lack of language- and script-specific OCR in addition to a lack of scholarly infrastructure in the OT situation have meant that scholars have been slow to adopt digital methods. 51

The landscape of Arabic script languages is changing rapidly, especially with neural automatic transcription systems, as they can accommodate cursive-like scripts and typefaces. In order for these systems to perform well, however, they will require training on many more domains and spatio-temporal variants of the language and handwriting styles. In the case of printed materials, training on different typefaces and printing conventions will also be necessary. 52

A move to new transcription methods for OT will not mean the total automatization of the transcription process, removing it from scholarly labor; instead it will require a reorganization of those efforts from informal to somewhat organized one. We do not expect that people will stop transcribing small amounts by hand for specific purposes, but we do suspect that it will change the way that large text creation projects work, as well as how archives create finding aids as digitized material becomes available. 53

In the specific case of printed periodicals in OT, in order to achieve the goal of enabling keyword searching and more complex forms of modeling and analysis of OT textual content via HTR produced text, a multi-pronged approach will be required. First, a team-based approach to transcription for the purposes of the creation of ground truth will be required (or even some crowdsourced, community-based approaches that have yet to be defined). Such a community effort might be achieved by sharing the model generated for the HTU print collection publicly, although how the endeavor would scale remains to be seen. Second, even though the platform-as-service model of the READ COOP may provide scholars with a variety of public models, the computing resources that allow these to run might exclude scholars and archives without the resources to buy in. It may be that only open models provide a more sustainable solution. Third, research is needed not only to recognize the Arabic script that is used in the specific genre discussed in this article — periodicals — but also, as in many other language situations, for structure recognition and image extraction that would allow for formal modeling of the resultant periodicals as research objects. 54

We are not experts in the histories of all writing systems, but we believe our results may be applied to other language situations that have changed, or have used multiple, scripts over time: Aljamiado, Azeri, Bosnian, Crimean, Judeo-Spanish, Malay and Wolof, just to name a few. Our research with automatic transcription of periodicals printed in the Ottoman Empire during the mid nineteenth to early twentieth century has important implications for multi-script archives of nation-states located in the borderlands of empires and religions. After all, language reforms forged new modernized versions of languages such as Hebrew and Arabic while maintaining their original scripts, but in other places writing systems were changed altogether. As we described in depth above, in Turkey the Arabic script of OT was abandoned in favor of a modified Latin alphabet. The political disintegration of states at other moments of the twentieth century, in particular the former Yugoslavia and Soviet Union, led to fragmentation of similar languages into script-divergent varieties that both straddle current political borders and co-exist with older variance found in archives. If script change comes about in moments of radical political change enacting ruptures in archives, it also seems to map onto a slow globalizing of the Latin script — and not without significant political debate — in different countries around the world around the turn of the twenty-first century. Turkmenistan adopted a Latin alphabet in the late 1990s as well as Kazakhstan as recently as 2017. This geopolitical context raises the question of what will become of digitized archives in such locales — printed and handwritten — and adds totally new contours to the notion of the "great unread" in nations already cut off from the scripts of their past, as well as those that may be facing a phasing 55 out of old alphabets. Future developments in digitization and text creation from borderland archives — if they are to happen — need to take into account not only the automatic transcription of written language, but also the various writing systems in which those languages will be expressed.

Key Concepts: Transcription, Transliteration, Romanization

There are a number of terms that are intrinsically related, and some of which we use interchangeably in this paper, that are central to our automated recognition of OT project. 56

Transliteration is the substitution of the characters of a script by those of another alphabet. In an ideal transliteration, each character of the source script will be represented with only one and always the same symbol in the target writing system, creating a one-to-one mapping, or graphemic correspondence, between scripts. 57

Transcription, on the other hand, is the representation of the sounds of an alphabet with the characters of another script. A transcription attempts to offer an accurate phonetic rendering of the source language even when this violates the graphemic correspondence between scripts. This means that, depending on the textual context, a polyphonic character- a letter that represents more than one sounds- may be substituted by several different letters in the target alphabet. In some cases, the reverse might be true. In the instance of OT transcription to modern Turkish, transcription also involves representing sounds, vowels to be more specific, which are only implied but are not written in the original text. Precisely because of this, the transliteration of OT inevitably becomes transcription. 58

A popular, or "loose," transcription, is an approximation of the conventional orthography and popular pronunciation of a word in a different script. It will produce a transcription that is easy to follow by contemporary audiences while inevitably forgoing important linguistic information in the original text. 59

Finally, romanization is the substitution of the characters of a non-Roman script by those of the Roman alphabet [Halpern 2007]. 60

Notes

[1] In a 1960 article, Stanford Show, a leading scholar of Ottoman history, described the challenges of working in Ottoman archives as follows: "Months of searching into the catalogs is necessary to locate all available materials concerning each subject, and much longer time is required to gather these materials and mold them into an intelligible unit of study" [Shaw 1960, 12]. For a more recent evaluation of the state-of-the-field, which reveals that there has not been any significant improvement in the accessibility of the archive since the 1960s, see Gratien et al. (2014).

[2] Such was the experience of one of the authors of this article, a native speaker of Turkish, while conducting her doctoral research on the emergence of a modern public education network for women in the Ottoman Empire. The examination of the undigitized and uncataloged documents of the Ottoman Ministry of Public Education and the relevant Ottoman Turkish periodicals took over a year. She enlisted the help of a transcriber to annotate and transcribe the documents she collected. The hired transcriber also conducted research in the Hakkı Tarık Us Periodicals Collection, which, at the time, had not yet been fully digitized.

[3] The two most notable projects in this vein are the digitizations of the Hakkı Tarık Us (HTU) print periodicals repository and the manuscript collections of the Presidential Ottoman State Archives. Even though it has been a decade since the completion of the digitization phase of the project, at present the web interface for the HTU is still in Beta version. Catalogue information about the contents of the collection is meager and search options are non-existent. For a detailed report on the contents and accessibility of the HTU digital repository, see Baykal (2011). Work on the digitization of the central Ottoman archive continues to this day. So far, some 40 million documents from the collections of the Ottoman and Republican archives are estimated to have been digitized and made available online.

[4] Salmi draws upon the argument made by [Prescott 2018] which explains that many digitization projects were actually concerned with the remediation of microforms rather than searchable text creation.

[5] We know that it was during roughly the same period of the digitization of the HTU that the Arabic-language plain text corpus known as *Al-Maktaba al-Shāmila* came into being, probably via manual keying of the texts [Verkinderen 2020].

[6] It is important to note that one of the most famous European crowd-transcription projects Transcribe Bentham, was not set up until 2012.

[7] The authors of this article worked with an OCR system (*Transkribus*) that was developed in the context of an EU-funded research initiative for transcribing historical archives which in recent years has evolved into a scholarly cooperative (READ COOP). Other methods for neural-based automatic transcription have been under development for some time, including Kraken and E-Scripta, which require a significant investment in digital research infrastructure that was not feasible for the authors in the institution to which they belong.

[8] As the small Ottoman principality from western Anatolia evolved into a multi-lingual, multi-ethnic empire that came to rule over most of the Middle East and South Eastern Europe the syntax and the writing system of OT grew in their complexity while its vocabulary expanded exponentially. For an evaluation of the development of Ottoman language in its written form, see Darling (2012).

[9] By "Turkified" we mean that many of the Arabic and Persian words in the language were removed in favor of words of Turkish origin. For an in depth analysis of Turkish language reform, see Lewis (2002).

[10] As a case in point, two of the most prominent Turkish Ottomanists of the early Republican era, Kemal Karpat and Halil Inacik, received their higher education and/or worked in Western universities. A sizable portion of both of these scholars' publications are in English, i.e. their intended audience was Anglophone and not necessarily Turkish. For a further example of a key academic discussion taking place outside of Turkey, see Birnbaum's 1967 article on the transliteration of OT, where he discusses in detail the indifferent attitude in Turkey toward developing a coherent transcription scheme for OT [Birnbaum 1967].

[11] To this end, a team of professional transcribers is employed by the Presidential State Archives in Istanbul to transcribe, edit, and publish selections of documents from their vast holdings. These are made available online for free for public use. Turkish publishing houses also commission experts in the field to transcribe and annotate important cultural texts in Ottoman Turkish or to produce adaptations of these in modern Turkish for popular consumption.

[12] The conversation is not a new one, but recent years have seen growth in scholarly discussion about them. Researchers have recently stressed not only the urgency of access to digital tools but also the inequities of digital material and knowledge infrastructures across languages as a question of access. Notable efforts include a full-day workshop at the 2019 ADHO conference entitled Towards Multilingualism In Digital Humanities: Achievements, Failures And Good Practices In DH Projects With Non-latin Scripts organized by Martin Lee and Cosima Wagner; the Multilingual DH Group and the efforts of Quinn Dombrowski; the RTL (Right to Left) workshop at the Digital Humanities Summer Institute organized by Kasra Ghorbaninejad and David Wrisley; and one-and-a-half day workshop Disrupting Digital Monolingualism organized by the Language Acts and Worldmaking Project.

[13] To this question of infrastructure, one really needs to add contemporary forms of inaccessibility that researchers — local and global — experience in times of reduced funding and mobility.

[14] While there is a small community of Turkish developers working on NLP tools for modern Turkish in open source environments, these are still far from being fully workable platforms. Neither are they sufficiently comprehensive to accommodate the complicated orthography of Ottoman Turkish. To our knowledge, the most advanced NLP tool for Turkish so far developed is Zenberek. Also see Sezer et al. (2008) and Sak et al. (2008).

[15] There have been several initiatives in recent years to introduce digital tools and methods to the field of Ottoman studies. The most notable is OpenOttoman, an online platform to foster collaborative DH scholarship among Ottomanists. For an overview of the objectives of the platform, see Singer (2016). The Baki Project and the Ottoman Text Archive Project are the two other noteworthy projects for creating textual DH platforms for Ottoman studies.

[16] An excellent argument for devising computational tools for such one such language, nineteenth century Hebrew, is Soffer et al. (2020). For a pilot study on the automated text recognition of nineteenth century printed books in Bengali, see Derrick (2019). Although not in a non-Western language, Ventresque et al.'s work on transcribing Foucault's reading notes with Handwritten Text Recognition technology is another important case study that seeks a workable starting point for linguistically complex archival material [Ventresque et al. 2019].

[17] For an OCR system that is based on the same "line-level" recognition principle as HTR, see Romanov (2017).

[18] See, for example Baykal (2019), Karakaya-Stump (2003), Gülaçar (2018).

[19] This is no less true of OT periodicals collection at HTU and Arabic periodicals in large collections such as Arabic Collections Online (ACO) and the Qatar Digital Library (QDL).

[20] Birnbaum discusses this issue in detail in his paper [Birnbaum 1967].

[21] In the context of debates surrounding a canonical seventeenth century OT text, H.E. Boeschton declared that "a fully consistent transcription is impossible" and "a transcription is a medium ill-suited for the presentation of linguistic results" [Boeschoten 1988, 23–6]. The position we take in this study, which is arguably the commonly accepted practice today, is formulated by Robert Anhegger as a response to Boeschton: "The guiding principle should be to produce a transcription which is as easy as possible to follow" [Anhegger 1988, 14].

[22] Currently, there are four transcriptions schemes widely used in scholarly publications: Deutsche Morgenländische Gesellschaft (DMG) system, Encyclopaedia of Islam (EI, 2nd edition) system, International Journal of Middle Eastern Studies (IJMES) system, and and Islam Ansiklopedisi (IA) system. For a comparative chart of these transcription systems, see Bugday (2014). Turkish scholars tend to prefer the IA system. Modern Turkish versions of OT texts that are produced for the general population, on the other hand, often do not adhere to a strict system but rather follow a "loose" transcription. As a case in point, the Turkish Historical Society — a government agency for the study and promotion of Turkish history — endorses a "loose" rather than "scientific" transcription of post-1830s printed works. For archival material from the same period, however, they recommend "scientific transcription," a term they use interchangeably with "transliteration." Türk Tarih Kurumu, "Eser Yayın Süreç: Çeviriyazı Metinlerde Uyulacak Esaslar" (https://www.ttk.gov.tr/eser-surec/eser-yayin-surec/).

[23] The Arabic loanwords in OT complicates this scheme significantly as Arabic stems do inflect. Therefore, a comprehensive dictionary for such a system needs to include all inflected forms of the Arabic loanwords in OT.

[24] Korkut's system currently has only about 43,000 words while Dervaze boasts 72,400 OT words in its database.

[25] Korkut points out a few long term solutions to this problem in his article [Korkut 2019, 5].

[26] Transkribus does allow integrating custom prepared dictionaries into HTR training, but at present the computational cost is too high to justify the miniscule improvement in the accuracy rate of the automated transcription. Based on a correspondence with the Transkribus team, however, we understand that there has been a recent improvement in the recognition-with-dictionary workflow which speeds up the process. In expectation of further developments in this vein, we have started compiling a comprehensive digital Ottoman Turkish dictionary based on Ferit Devellioğlu's seminal work [Devellioğlu 1998]. Once the dictionary is ready, it will be publicly shared on the Transkribus platform.

[27] By "diverse," we mean a training set that includes samples from various subdomains of the corpus in question. These subdomains will ideally present variations in typeface, page layout, and content, thus, in vocabulary and named entities.

[28] As a case in point, the number of general HTR models made available to the public in the Transkribus interface have increased significantly over the last year, expanding not only the language, but the domain and genre applicability of the method. For example, the National Archives of the Netherlands has published a composite model for Dutch-language documents of the seventeenth, eighteenth and nineteenth centuries called Ijsberg, trained on dozens of handwriting styles in commercial letters and notarial deeds, that has achieved a CRE rate of 5.15% [Transkribus 2019].

[29] We take these two publications to be typical OT periodicals in terms of page layout, typeface, and content. Transcriptions for the training data were repurposed partially from the research material of Suphan Kirmizialtin's dissertation study and partially from the website Osmanlıca Mahalli Gazeteler. Existing transcription texts were modified to normalize spellings, correct transcription errors, and standardize the transcription schemes.

[30] See Appendix-Key concepts.

[31] See note #22.

[32] This most frequently occured in cases that involve the conventionalized affixes of OT which violate the vowel harmony of MT. In such cases, while character-accurate transcription results in archaic-sounding pronunciation of those words (such as الدی : “oldı” - “happened” ; اوجونجی : “üçünci” - “third”) they are still easily recognizable by modern readers; therefore, in those occasions, we opted to go with diplomatic transcription. However, in other instances, where literal transcription, i.e. transliteration, produces a pronunciation that renders the word unrecognizable to the speakers of the Turkish language (for example, خواجه : “havace” ; كوكرجن : “kükercin”) we preferred the conventional pronunciation over character accuracy (خواجه : hoca ; كوكرجن : “güvercin”). This decision did, inevitably, introduce a certain degree of inconsistency to our HTR models.

[33] Here, we had to take into account the bi-directionality of OT. In OT, as in Arabic, alphabetical characters are written RTL while numerical characters are written LTR.

[34] As necessity can be the mother of invention, our “hack” to the system and our initial results with OT prompted the *Transkribus* team to develop and integrate a new functionality into the interface that automates reversing of the direction of the transcription text.

[35] By this time, *naskh* style typeface, cut by the revered punch cutter and master printer Ohannes Muhendisyan, had emerged as the standard continuous text typeface for OT publications [Yazıcıgil 2015].

[36] For Language Models, see *Discussion of the Results* section.

[37] “Language Models (LM) estimate the probability of a specific word *w* given the history of words *w*1, *w*2,...using external language resources. Assuming that these probabilities model the language of the current document well, we output the transcription which maximizes a combination of the HTR probability and the LM probability” [Strauss et al. 2018].

[38] For example, in our experiment with *Küçük Mecmua*, the CER was reduced from 5.23% to 3.63% when the automated transcription was implemented with the Language Model option.

[39] See note #26.

Works Cited

Ahali n.d. Ahali (HTU no. 0341/1) http://www.tufs.ac.jp/common/fs/asw/tur/htu/list1.html#a

Aktaş and Halaçoğlu 1992 Aktaş, N., Halaçoğlu, Y. (1992). *Başbakanlık Osmanlı Arşivi*. TDV İslâm Ansiklopedisi.Retrieved from https://islamansiklopedisi.org.tr/basbakanlik-osmanli-arsivi

Alpert-Abrams 2016 Alpert-Abrams, H. (2016). “Machine Reading the Primeros Libros.” *Digital Humanities Quarterly*, 10(4). http://www.digitalhumanities.org/dhq/vol/10/4/000268/000268.html

Andres et al. 2008 Andres, W., Inan, M., Kebeli, S., Waters, S. (2008). *Rethinking the Transcription of Ottoman Texts: The Case for Reversible Transcription*. Retrieved from http://otap.bilkent.edu.tr/reverse/reverse/o_Reverse_trans_article728.html

Anhegger 1988 Anhegger, R. (1988). “On Transcribing Ottoman Texts.” *Manuscripts of the Middle East* 3, 12-15. http://www.islamicmanuscripts.info/reference/articles/Anhegger-1988-Transcribing-MME3.PDF

Baki Project n.d. *Baki Project* (http://www.thebakiproject.org/main/)

Baykal 2011 Baykal, E. (2011). Periodicals of the Hakkı Tarık Us Collection. *Turkish Historical Review* 2, 205–212.

Baykal 2019 Baykal, E. (2019). *The Ottoman Press (1908-1923)*. Leiden, Boston: Brill.

Birnbaum 1967 Birnbaum, E. (1967). “The Transliteration of Ottoman Turkish for Library and General Purposes.” *Journal of the American Oriental Society* 87, 122–156.

Boeschoten 1988 Boeschoten, H.E., 1988. “Why Transcribe Ottoman Turkish Texts?” *Manuscripts of the Middle East* 3, 23–26. http://www.islamicmanuscripts.info/reference/articles/Boeschoten-1988-Why_transcr-MME3.PDF

Bugday 2014 Bugday, K. (2014). *An Introduction to Literary Ottoman*. London, New York: Routledge.

Çakır 1994 Çakır, S. (1994). *Osmanli Kadin Hareketi*. İstanbul: Metis Yayınları.

Clifford et al. 2019 Clifford, E., Rusinek, S., Segal, Z., Rißler-Pipka, N., Ketchley, S., Roeder, T., Bunout, E., Düring, M. (2019). “Complexities in the Use, Analysis, and Representation of Historical Digital Periodicals,” *Proceedings of the ADHO2019 Digital Humanities Conference*. Utrecht. https://dev.clariah.nl/files/dh2019/boa/0447.html

Cordell 2016 Cordell, R. (2016). “What Has the Digital Meant to American Periodicals Scholarship?.” *American Periodicals: A Journal of History & Criticism* 26 (1), 2-7.

Cordell 2017 Cordell, R. (2017). “‘Qi-jtb the Raven’: Taking the Dirty OCR Seriously.” *Book History*, 20, 188-225.

Darling 2012 Darling, L.T. (2012). “Ottoman Turkish: Written Language and Scribal Practice, 13th to 20th Centuries.” In W. Hanaway & B. Spooner (Eds.), *Literacy in the Persianate World: Writing and the Social Order,* 171–195. Philadelphia: University of Pennsylvania Press.

Derrick 2019 Derrick, T. (2019). “Using Transkribus For Automated Text Recognition of Historical Bengali Books,” British Library Digital Scholarship Blog. Retrieved from https://blogs.bl.uk/digital-scholarship/2019/08/using-transkribus-for-automated-text-recognition-of-historical-bengali-books.html

Devellioğlu 1998 Devellioğlu, F. (1998). *Osmanlıca-Türkçe Ansiklopedik Lugat*. Ankara: Aydın Kitabevi Yayınları.

Ergişi and Şahin 2017 Ergişi, A., Şahin, E. (2017). "Osmanlıca Belgelerin Elektronik Çeviri Uygulamaları İçin Bir İmla Kılavuzu Örneği: Dervaze." *International Journal of Languages Education* 1, 78–84.

Ghorbaninejad et al. 2022 Ghorbaninejad, M., Gibson, N., & Wrisley, D.J. *RTL Debates in Digital Humanities* 2022 (forthcoming).

Grallert 2019 Grallert, Till. (2019) *Kanun-i-Esasi*. Retrieved from https://github.com/tillgrallert/kanun-i-esasi

Gratien et al. 2014 Gratien, C., Polczyński, M., Shafir, N. (2014). "Digital Frontiers of Ottoman Studies." *Journal of the Ottoman and Turkish Studies Association* 1, 37–51. Retrieved from https://doi.org/10.2979/jottturstuass.1.1-2.37

Gülaçar 2018 Gülaçar, A.G. (2018). "II. Meşrutiyet Dönemi İktidar Oluşumu Sürecinde Basının Rolü." *VAKANÜVİS- Uluslararası Tarih Araştırmaları Dergisi* 3, 105–128.

Hakkı Tarık Us n.d. Hakkı Tarık Us Collection (HTU) http://www.tufs.ac.jp/common/fs/asw/tur/htu/

Halpern 2007 Halpern, J. (2007). "The Challenges and Pitfalls of Arabic Romanization and Arabization." [Semantic Scholar] Retrieved from https://www.semanticscholar.org/paper/The-Challenges-and-Pitfalls-of-Arabic-Romanization-Halpern /8744e1a7aa3637387331fcc56973f6e7b409695c

Kadınlar Dünyası n.d. *Kadınlar Dünyası* (HTU no. 1262) http://www.tufs.ac.jp/common/fs/asw/tur/htu/list1.html#k

Karakaya-Stump 2003 Karakaya-Stump, A. (2003). "Debating Progress in a 'Serious Newspaper for Muslim Women:' The Periodical 'Kadin' of the Post-Revolutionary Salonica, 1908-1909." *British Journal of Middle Eastern Studies* 30, 155–181.

Korkut 2019 Korkut, J. (2019). "Morphology and Lexicon-Based Machine Translation of Ottoman Turkish to Modern Turkish." Retrieved from https://www.cs.princeton.edu/~ckorkut/papers/ottoman.pdf

Küçük Mecmua n.d. Küçük Mecmua (HTU no. 0171) http://www.tufs.ac.jp/common/fs/asw/tur/htu/list1.html#k

Lewis 2002 Lewis, G. (2002). *The Turkish Language Reform: A Catastrophic Success*. New York: Oxford University Press.

Martin 2007 Martin, S. (2007). "Digital Scholarship and Cyberinfrastructure in the Humanities: Lessons from the Text Creation Partnership." *Journal of Electronic Publishing* 10.

Mekteb n.d. *Mekteb* (HTU no. 0138) http://www.tufs.ac.jp/common/fs/asw/tur/htu/list2.html#m

Mussell 2012 Mussell, J. (2012). *The Nineteenth Century Press in the Digital Age*. London: Palgrave Macmillan.

Nicholson 2013 Nicholson, B. (2013). "The Digital Turn: Exploring the Methodological Possibilities of Digital Newspaper Archives." *Media History* 19(1), 59-73.

OTAP n.d. Ottoman Text Archive Project http://otap.bilkent.edu.tr/

Osmanlıca Mahalli n.d. Osmanlıca Mahalli Gazeteler https://www.osmanlicagazeteler.org/index.php

Özkal 2018 Özkal, Ö. (2018). "Ottoman Foundations of Turkish Typography: A Field Theory Approach." *Design Issues* 34, 59–75.

Padilla et al. 2019 Padilla, T. et al. (2019). "Final Report - Always Already Computational: Collections as Data." [Zenodo]. Retrieved from https://zenodo.org/record/3152935/

Prescott 2018 Prescott, A. (2018). "Searching for Dr Johnson: The Digitisation of the Burney Newspaper Collection." *Travelling Chronicles: News and Newspapers from the Early Modern Period to the Eighteenth Century.* Leiden: Brill.

Rabus 2020 Rabus, A. (2020). *Training of Large Models*. Presented at the Transkribus User Conference, University of Innsbruck.

Romanov et al. 2017 Romanov, M., Miller, M.T., Savant, S.B., Kiessling, B. (2017). *Important New Developments in Arabographic Optical Character Recognition (OCR)*. arXiv:1703.09550 [cs]. Retrieved from http://arxiv.org/abs/1703.09550

Sak et al. 2008 Sak, H., Güngör, T., Saraçlar, M. (2008). "Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus." *Advances in Natural Language Processing* 5221, 417–427.

Salmi 2021 Salmi, H. (2021). *What is Digital History?* Cambridge, UK: Polity Press.

Sezer n.d. Sezer, T. TS Corpus - The Turkish Corpora and NLP Project. [TS Corpus]. https://tscorpus.com/

Shaw 1960 Shaw, S.J. (1960).“ Archival Sources for Ottoman History: The Archives of Turkey." *Journal of the American Oriental Society* 80, 1–12.

Siemens et al. 2009 Siemens, R., Leitch, C., Blake, A., Armstrong, K., Willinsky, J. (2009). "'It May Change My Understanding of the Field': Understanding Reading Tools for Scholars and Professional Readers." *Digital Humanities Quarterly* 3.4.

Singer 2016 Singer, A. (2016). "Introducing the Ottoman Gazetteer and OpenOttoman." *Journal of the Ottoman and Turkish Studies Association* 3, 407–412.

Soffer et al. 2020 Soffer, O., Segal, Z., Greidinger, N., Rusinek, S., Silber-Varod. (2020). "Computational Analysis of Historical Hebrew Newspapers: Proof of Concept." *Zutot* 17, 97–110.

Strauss et al. 2018 Strauss, T., Weidemann, M., Labahn, R. (2018). *D7.12 Language Models; Improving Transcriptions by External Language Resources.* Retrieved from https://readcoop.eu/wp-content/uploads/2018/12/D7.12_LMs.pdf

T.C.a n.d. T.C. Cumhurbaşkanlığı Devlet Arşivleri Başkanlığı Yayınlar. Retrieved fromYayınlar - TC Cumhurbaşkanlığı Devlet Arşivleri Başkanlığı

T.C.b n.d. T.C. Cumhurbaşkanlığı Devlet Arşivleri Başkanlığı. (2019, September 2). 60 Milyon Tarihi Belge Dijital Ortama Aktarıldı. Retrieved from https://www.devletarsivleri.gov.tr/Sayfalar/Haberler/Haber/2142

TTK n.d. Türk Tarih Kurumu, n.d**.** *Eser Yayın Süreç: Çeviriyazı Metinlerde Uyulacak Esaslar.* Retrieved from https://www.ttk.gov.tr/eser-surec/eser-yayin-surec

Tasvir-i Efkar n.d. *Tasvir-i Efkar* (HTU no. 2267) http://www.tufs.ac.jp/common/fs/asw/tur/htu/list2.html#t

Transkribus 2019 Transkribus. (2019). *Public Models in Transkribus*. Retrieved from https://transkribus.eu/wiki/images/archive/d/d6/20200317101228%21Public_Models_in_Transkribus.pdf

Ventresque et al. 2019 Ventresque, V., Sforzini, A., Massot, M.-L. (2019). “Transcribing Foucault’s handwriting with Transkribus.” *Journal of Data Mining & Digital Humanities*. Retrieved from https://jdmdh.episciences.org/5218

Verkinderen 2020 Verkinderen, P. (2020). “Al-Maktaba al-Shāmila: A Short History.” Kitab Project [blog] Retrieved from http://kitab-project.org/2020/12/03/al-maktaba-al-shamila-a-short-history/

Yazıcıgil 2015 Yazıcıgil, O. (2015). *TYPO Talks: Continuous text typefaces versus display typefaces in the Ottoman Empire*. Retrieved from https://www.typotalks.com/videos/continuous-text-typefaces-versus-display-typefaces-in-the-ottoman-empire/

Zenberek n.d. Zenberek https://github.com/ahmetaa/zemberek-nlp