OpenITI AOCP Phase I White Paper v. 1.1

Digitizing the Islamicate Written Traditions:
History, State of the Field, and Best Practices for Open-source Arabic-script OCR

The Open Islamicate Texts Initiative Arabic-script Catalyst Project
(OpenITI AOCP)

Introduction

This white paper consists of two distinct though interrelated parts: first, an overview and evaluation of the work we in the Open Islamicate Texts Initiative have undertaken to improve open-source Arabic-script OCR; second, a step-by-step discussion of the work-flow we developed as well as a discussion of outstanding issues and problems. In documenting the steps and missteps that we have taken over the course of this project, we hope that other teams and individuals pursuing related projects will be able to learn from our successes and failures. In addition to functioning as a work of documentation, the second half of this paper serves as a user's guide to our OCR workflow. We will close by highlighting the improvements we have made in Arabic-script OCR and present the outstanding issues in our OCR workflow and user-friendly digital text production pipeline that we will address in Phase II of OpenITI AOCP.

The first section of our paper will explore the process whereby we developed our current OCR models for Arabic script, one of the key promised deliverables for this phase of our work. We will also examine the development process for our digital text production pipeline, eScriptorium, which has been the main engine for generating training data for OCR model improvement. eScriptorium is itself an important deliverable that will be released to the wider public in the near future. The reader who is primarily interested in producing digital texts using eScriptorium can pass over this first section of the paper and advance to the second which consists of a step-by-step overview.

We have provided what follows out of a desire to document our own process and to give suggestions for others pursuing similar ends. As with any project, we learned much through trial and error. Like many others working over the past few years, our work was significantly interrupted by the COVID-19 pandemic and its cascading effects upon virtually every aspect of academic life and work. In keeping with other recent work on digitization in the humanities, our discussion below will serve in part to underline the centrality of physical material and of human labor to produce end products that 'live' online in a seemingly ethereal world.

This white paper primarily covers the humanities side of OpenITI AOCP work, i.e., the physical work, articulation of ideas, and other interventions primarily carried out by those team members with backgrounds in traditional humanities scholarship. In practice, the technical and humanistic aspects of this project grew closely intertwined, with each side of a hybrid team required to have some cognizance of the parameters and content of the other side's work and disciplinary norms. Readers interested in the details of the computer science side of our project should consult the technical publications produced by David Smith and Alejandro Toselli.[1] The information and insights contained here also are drawn from our interactions with users of eScriptorium from around the world over the last year, including many undergraduate and graduate students from various disciplinary backgrounds as well as faculty members at a range of

[1] A.H. Toselli, S. Wu, D. A. Smith, D.A., *Digital Editions as Distant Supervision for Layout Analysis of Printed Books*, in Lladós, J., Lopresti, D., Uchida, S. (eds), *Document Analysis and Recognition – ICDAR 2021* (Springer, Cham, 2021), https://doi.org/10.1007/978-3-030-86331-9_30. GitHub code and data release can be found here.

institutions. A number of individuals made use of eScriptorium as part of paid work in training data generation, while others used the platform as part of various pedagogical projects.

A Brief History of the OpenITI Arabic-Script OCR Catalyst Project (OpenITI AOCP)

The OpenITI AOCP project began in 2016 as a collaboration between the Persian Digital Library (PDL) project of the Roshan Initiative in Persian Digital Humanities (PersDig@UMD) of Roshan Institute for Persian Studies at the University of Maryland; the Knowledge, Information, Technology, and the Arabic Book (KITAB) project of Aga Khan University (London); and the OpenArabic project of Leipzig University.[2] The principal investigators of these respective projects–Matthew Thomas Miller, Sarah Bowen Savant, and Maxim Romanov–were brought together by a shared need for the development of international digital text standards and improved Arabic-script OCR. Their initial collaboration with computer scientist Benjamin Kiessling, creator of the open-source OCR system Kraken, yielded the best OCR character accuracy rates (CARs) ever achieved up to that time on printed editions of premodern Arabic texts.[3] The considerable promise of this work was recognized by the Mellon Foundation, who funded Phase I of OpenITI AOCP to advance open-source Arabic-script OCR. Below is a brief overview of the different components of the OpenITI AOCP Phase I work plan that together enabled us to bring OCR CARs on the most important Arabic and Persian typefaces to over 97.21% when averaged by line and 97.08% when averaged by typeface (so that typefaces with more annotated data count equally).

a. *Typographic Survey*

One of the first steps in the project was identifying a representative range of major typefaces used in Arabic and Persian from the advent of large-scale metal typographic printing[4]

[2] See the Persian Digital Library project homepage here, the KITAB project homepage here, and the OpenArabic homepage here.

[3] Benjamin Kiessling, Matthew Thomas Miller, Maxim G. Romanov, and Sarah Bowen Savant, "Important New Developments in Arabographic Optical Character Recognition (OCR)." *Al-ʿUṣūr al-Wusṭā: The Journal of Middle East Medievalists* 25 (2017): 1-13. Doi: http://dx.doi.org/10.17613/M6TZ4R. See previous baseline studies reported in: Mansoor Alghamdi and William Teahan, "Experimental evaluation of Arabic OCR systems," *PSU Research Review* 1/3 (2017): 229-241. doi: https://doi.org/10.1108/PRR-05-2017-0026

[4] As distinct historically from 1) medieval woodblock printing and 2) nineteenth and twentieth-century lithographic printing, which essentially reproduced the script conventions and extensive variability of the manuscript tradition.

(ca. 1830) up to the present.[5] As our goal in refining Arabic-script OCR was the production of a generalizable model—that is, one that can achieve reasonably high CARs for any given typeface—our training data set needed to be broadly representative of the Arabic-script print tradition. Since no truly comprehensive reviews or lists of Arabic-script typefaces and their features existed at the time, we determined that we would need to perform a manual review of a large number of Arabic and Persian works to ascertain the most important typefaces.[6] This sounds straightforward enough, but we encountered a number of unforeseen obstacles. While there is a plethora of Arabic-script texts available online in digitized format, this dispersed corpus has its own issues and limitations. Public domain repositories of print material are naturally limited to a chronologically restricted range of typefaces due to copyright concerns. Repository websites with less concern for copyright niceties invariably have their own biases in terms of genres covered, presses favored, and the like, which also limit the range of typefaces represented therein. As such, while online repositories have indeed been of immense help, particularly in the initial stages of generating training data, we realized that it would be preferable if we could access physical books held in a large research library.

Multiple attempts to coordinate a large-scale randomized selection of books from an academic library with significant Arabic and Persian holdings all came up short. Finally, after prolonged negotiations with the Library of Congress in early 2020, we secured an agreement with them to conduct a randomized survey of their substantial holdings in our target languages. In March of that year, we were on the cusp of initiating this study when virtually all institutions, the Library of Congress among them, shut down indefinitely in response to the COVID-19 pandemic. The Library did not reopen until 2021, and then only in limited capacity. We were eventually able to carry out a more limited survey using books made available on-site,[7] which was helpful and uncovered some new typefaces we had not previously encountered. While as of the writing of this paper we have been unable to carry out a comprehensive, genuinely randomized exploration of the Arabic-script print corpus from its origins to present, the reviews we have been able to make, including our semi-randomized forays at the Library of Congress, have indicated that our initial selections were accurate in representing the main typographic contours of Arabic script print. Our corpus pilot project has further reinforced this intuition. We are also working with our computer science colleagues to automatically analyze typographic diversity and coverage at scale using the extensive collections of New York University's *Arabic Collections Online*.

[5] It is certainly true that attempts at Arabic-script typographic printing predate the 1830s, but these were generally sporadic efforts concentrated in Western European printing centers, and as such were of relatively limited purchase in the Islamicate world.

[6] Since this project began, a handful of significant works on the history and parameters of Arabic typography have appeared, including: *Manuscript and Print in the Islamic Tradition,* ed. Scott Reese (Berlin: De Gruyter, 2022); *Arabic Typography History and Practice*, ed. Titus Nemeth (Salenstein: Niggli Verlag, 2023).

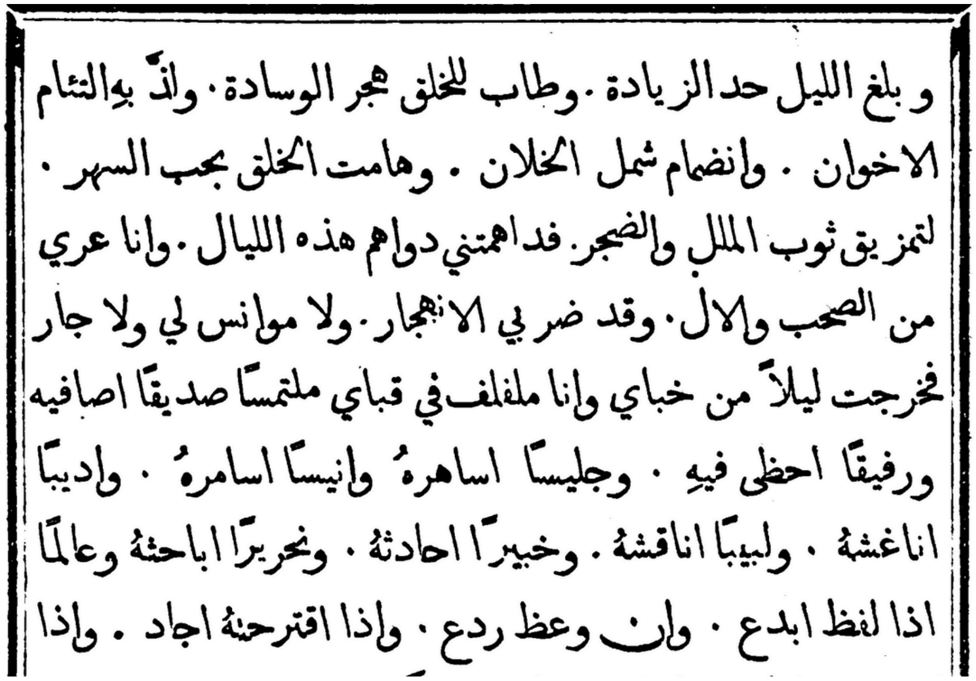
[7] The process was not genuinely random, however, as we had originally desired: the Library held aside books requested and used by other patrons, which meant our selection was contingent upon whatever other patrons in the past week had perused.

Despite these later positive developments, over the course of 2020 we had to make do with what was available to us, recognizing the limitations of our data collection. We remain open to the possibility that, with time, we will encounter unfamiliar typefaces and patterns, as has indeed been the case thus far. Fortunately, our generalized model is holding up well to new and even relatively divergent typefaces. The full results of our typeface study and continued work in further expanding our coverage will be reported in full in the forthcoming publication by OpenITI AOCP Phase I postdoctoral associate Jonathan Parkes Allen.[8] In general, however, we have benefited from Arabic script's lack of typographic diversity relative to Latin script print, with comparatively few typefaces predominating at any given period in time, including the last couple of decades. Instead, particularly for the sorts of texts in which we are most interested (i.e., editions of premodern Islamic works), a few typefaces have predominated, for reasons both technological and cultural, which has meant that our modest number of target typefaces for training data production has on the whole proven sufficient.

b. Producing Training Data:

In Phase I, we approached the work of improving Arabic and Persian OCR CARs by producing human-annotated training data for each of the typefaces we had previously identified as important for broad historical coverage. Tests completed shortly before the grant period began showed that highly accurate OCR transcription models could be produced using 800-1,000 lines of training data for most typefaces. Some of the more challenging typefaces (primarily nineteenth-century exemplars, subject to greater variability in terms of type production, paper and ink quality, and book preservation) have required training data in the range of 1,000-2,000 lines.

[8] Jonathan Parkes Allen, "From Metal Typography to Electronic Texts: A Contribution to the History of Arabic-Script and its Technological History (and Possible Futures)," forthcoming.



و بلغ الليل حد الزيادة .وطاب للخلق هجر الوسادة . واذّ بهِ التئام
الاخوان . وانضمام شمل الخلان . وهامت الخلق بحب السهر .
لتمزيق ثوب الملل والضجر. فداهمتني دواهم هذه الليال . وانا عري
من الصحب والال . وقد ضربي الانهجار . ولا موانس لي ولا جار
فخرجت ليلاً من خباي وانا ملفلف في قباي ملتمسًا صديقًا اصافيه
ورفيقًا احظى فيهِ . وجليسًا اساهرهُ وانيسًا اسامرهُ . واديبًا
اناغشهُ . ولبيبًا اناقشهُ . وخبيرًا احادثهُ . ونحريرًا اباحثهُ وعالمًا
اذا لفظ ابدع . وان وعظ ردع . واذا اقترحتهُ اجاد . واذا

Fig. 1: Examples of one potentially problematic typeface, the 19th-century Amrīkī typeface used primarily in the Levant.

This training data was manually generated and then double and triple checked by human reviewers. Our initial text collection process located scanned books containing the target typefaces as determined by our research into Arabic-script typography and print history. While our own team members generated much of the training data, we also experimented with recruiting a number of graduate student workers over the summer of 2020, providing a stipend for their work on the project. This approach was somewhat successful. In the end, much of our training data and the bulk of our corpus pipeline production work was done by our two graduate fellows at the University of Maryland working as part of our OpenITI team. A small team working over a sustained period of time seemed to be the best option, which subsequent work in the second phase of our project (which will be discussed in our future review of that work) has underlined further.

Besides the continual work of generating training data, we encountered a number of obstacles that have shaped our ultimate products and should be kept in mind by users. Some of these are internal to the Arabic-script tradition and its particularities that had in the past acted as impediments to the expansion of OCR into Arabic-script languages. One especially thorny issue is the question of vocalization and other optional orthographic marks. Printed Arabic-script texts on the whole either do not include vocalization (i.e., the ‘short vowels,’ written as what are effectively sub- and superscripts to the main letterforms), or they include them sporadically, such as in citations from the Qur’an or hadiths. We experimented with several different approaches. At first, we had hypothesized that including vocalization in our training data would boost overall

CARs, yet ultimately the opposite result proved true.[9] As such, the models that we ended up producing do not attempt to transcribe vocalization, which could be a limitation for some users depending on their use case scenarios.

From the very beginning, our stated goal has been the development and refinement of OCR for Arabic script and the eventual inclusion of Ottoman Turkish and Urdu typefaces into our generalized model. For the first phase of the project, we restricted ourselves to Arabic and Persian typefaces, only incidentally adding Urdu as we went along due to a collaboration with Forman Christian College in Pakistan. Even with this restricted focus, we have not always been successful in balancing the two languages, in part due to the simple fact that there are many more printed books in Arabic than in Persian. This issue is further compounded by the relative difficulty of finding downloadable typographically printed books in Persian, as well as the greater number of Arabic-centered workers in our larger project. For these and other reasons we ended up with a data set weighted towards Arabic (Fig. 2). This has been the cause of some drag on accuracy numbers for Persian typefaces, which we consequently had to control for in our data.

Typeface	**Training**		**Test**		**Total**	
	#Pages	**#Lines**	**#Pages**	**#Lines**	**#Pages**	**#Lines**
pers_typefaces	358	9405	218	5932	576	15337
arab_typefaces	529	13007	668	14154	1155	26299
pers_arab_typefaces	887	22412	886	20086	1731	41636

Fig. 2: Training and test set page and line totals for all Arabic and Persian typefaces used in this project.

Thanks to the work of our graduate fellow Mehdy Sedaghat Payam, we also discovered typefaces missing from our initial overview. These typefaces are used more or less exclusively in modern works of fiction and literary criticism and are by and large born digital.[10] Our sources have mainly been texts drawn from premodern Islamicate literature, due to both our corpus production goals as well as the interests and expertise of most of our participants. We were fortunate to have a team member well-versed in contemporary Persian literature who was willing to obtain, digitize, upload, and process large amounts of such literature. After obtaining these

[9] An internal experiment conducted in November 2021 examined CARs for two models, one trained on vocalized text (OCR-TrVoc) and one trained on unvocalized text (OCR-TrNVoc). The two models were, in turn, applied to vocalized and unvocalized test texts. We found that OCR-TrVoc transcribed vocalized and unvocalized texts equally well, while OCR-TrNVoc transcribed unvocalized texts at an average CAR 5.56% above that of its average CAR for vocalized texts. Overall, OCR-TrNVoc transcribed the vocalized test set at an average 0.85% CAR better than OCR-TrVoc, and it transcribed the unvocalized test set at an average 5.84% CAR better than OCR-TrVoc.

[10] See Sedaghat Payam's blog post on this issue here: https://openiti.org/2022/02/15/Challenge-of-an-unknown.docx.html

born-digital typefaces, our computer science postdoctoral associate at the time, Alejandro Toselli, was able to generate synthetic training data for these typefaces and integrate it into our training data repository. The subsequent general model performed excellently on these typefaces, achieving CARs of 99.64% (see Fig. 13 below and the adjacent discussion for a breakdown of these results).

c. *Corpus Pilot Creation:*

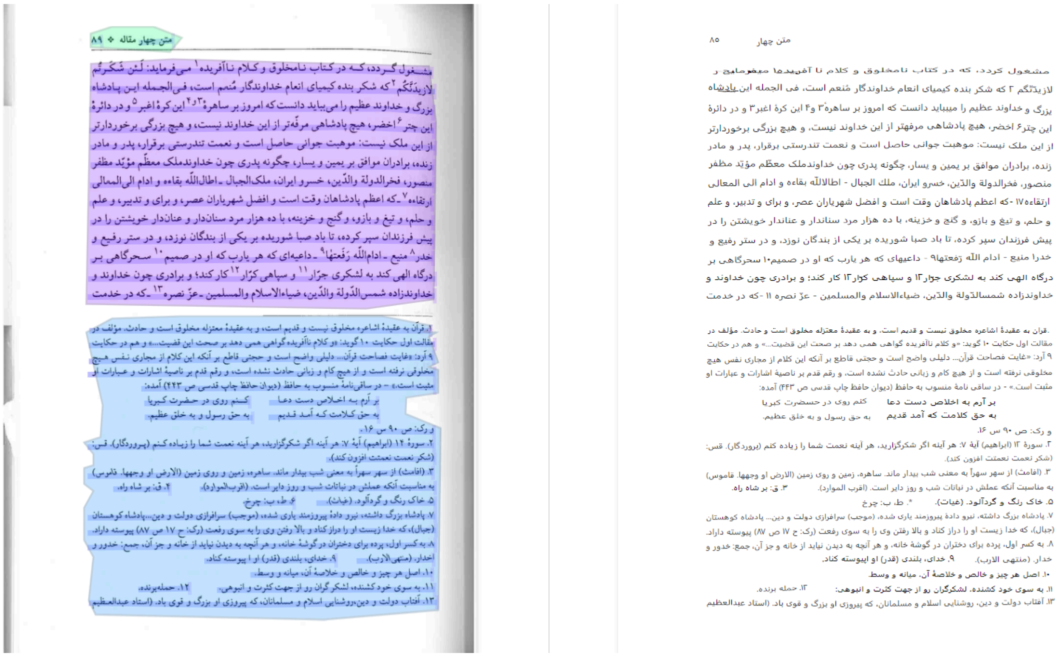
Our work over the last few months to produce and evaluate a corpus pilot of premodern Arabic and Persian literature has been an experimental process. We followed a work plan combining steps unique to corpus creation with the usual workflow of digital text creation using eScriptorium, namely:

1. Generating and refining list of works to be included in the corpus
2. Building a robust bibliography of the relevant works in their multiple editions
3. Locating and evaluating editions of individual works for inclusion in the corpus
4. Downloading and storing existing digital scans of books
5. Scanning and storing physical books
6. Uploading each work to eScriptorium and running the full range of processes on each document
7. Visually reviewing CARs for each document
8. Exporting and archiving the digitized documents

Our assembly of the corpus pilot was guided by a desire to include as much typographic diversity as possible while also expanding the OpenITI corpus, which at the time was made up of Arabic works which reflected many of the genre and chronological constraints of the source repositories. So-called ‘postclassical’ works, i.e., late medieval to early modern texts, were rather thinly represented, while entire categories and genres of premodern Islamicate literature were missing. For instance, there were few if any texts on Christian theology, philosophy, medicine, Sufism, and so forth. We populated our initial list of works, which contained approximately two hundred Arabic and two hundred Persian works, by surveying key stakeholders in Islamicate Digital Humanities for works they would like to see digitized as well as making our contributions based on our knowledge of the relevant bodies of literature.

The physical acquisition of books to be scanned for the corpus pilot was slow at first due to lingering COVID-19 restrictions, as we were almost entirely dependent upon Interlibrary Loan given the limited Arabic and premodern Persian coverage of our own institution’s physical library. We ended up obtaining copies of almost every work on our list. The actual work of scanning the books, saving the files, and organizing them proved less troublesome than we had anticipated. If anything, we realized that using a lower scan quality did not detract from OCR accuracy and meant smaller files for storage and manipulation. Perhaps the biggest bottleneck we encountered was periodic issues with eScriptorium, quite often due to issues related to server capacity. Finding ways to optimize server strength so as to ensure that eScriptorium runs quickly has been a recurring challenge, particularly in terms of making the platform open to a wider public. Despite slowdowns and supply issues, once we had scanned and uploaded the works in the corpus pilot, we were pleased by how well our generalized OCR model performed. We often got spot-check accuracy numbers higher than those given in our automatic CAR tests (on which see our discussion further below).

We did discover in this process, however, that our layout analysis model needed improvement, for reasons going back to the nature of our training data sources. Relatively few of the texts we used were published in recent decades, and as such often lacked the expansive footnote apparatus many Arabic (and to a lesser extent Persian) texts of recent decades possess. Having discovered shortcomings in our layout analysis models for detecting such footnote apparatuses, we trained a new model which reliably differentiates between main text and footnotes (Fig. 3). On the task of classifying pixels as belonging to footnotes, this model achieves 98% accuracy. This development facilitates export of premodern works suitable for textual analysis without modern critical commentary located on the margins of a page.



۸۵ متن چهار

مشغول گردد، که در کتاب نامخلوق و کلام نا آفریده۱ میفرماید ر
لازیدنّکم ۲ که شکر بنده کیمیای انعام خداوندگار مُنعم است، فی الجمله این پادشاه
بزرگ و خداوند عظیم را میبباید دانست که امروز بر ساهرهٔ۳ و۴ این کرهٔ اغبر۳ و در دائرهٔ
این چتر۶ اخضر، هیچ پادشاهی مرفهتر از این خداوند نیست، و هیچ بزرگی برخوردارتر
از این ملک نیست: موهبت جوانی حاصل است و نعمت تندرستی برقرار، پدر و مادر
زنده، برادران موافق بر یمین و یسار، چگونه پدری چون خداوندملک معظّم مؤیّد مظفر
منصور، فخرالدولة والدّین، خسرو ایران، ملك الجبال - اطالاللّه بقاءه و ادام الی المعالی
ارتقاءه۱۷ -که اعظم پادشاهان وقت است و افضل شهریاران عصر، و برای و تدبیر، و علم
و حلم، و تیغ و بازو، و گنج و خزینه، با ده هزار مرد سناندار و عناندار خویشتن را در
پیش فرزندان سپر کرده، تا باد صبا شوریده بر یکی از بندگان نوزد، و در ستر رفیع و
خدرا منیع - ادام اللّه رّفعتها۹ - داعیهای که هر یارب که او در صمیم۱۰ سحرگاهی بر
درگاه الهی کند به لشکری جرار۱۲ و سپاهی کرار۱۲ کار کند؛ و برادری چون خداوند و
خداوندزاده شمسالدّولة والدّین، ضیاءالاسلام والمسلمین - عزّ نصره ۱۱ -که در خدمت

.قران به عقیدهٔ اشاعره مخلوق نیست و قدیم است. و به عقیدهٔ معتزله مخلوق است و حادث. مؤلف در
مقالت اول حکایت ۱۰ گوید: «و کلام ناآفریده گواهی همی دهد بر صحت این قضیت...» و هم در حکایت
۹ آرد: «غایت فصاحت قرآن... دلیلی واضح است و حجتی قاطع بر آنکه این کلام از مجاری نفس هیچ
مخلوقی نرفته است و از هیچ کام و زبانی حادث نشده است، و رقم قدم بر ناصیهٔ اشارات و عبارات او
مثبت است.» - در ساقی نامهٔ منسوب به حافظ (دیوان حافظ چاپ قدسی ص ۴۴۳) آمده:

بر آرم به اخلاص دست دعا کنم روی در حسضرت کبریا
به حق کلامت که آمد قدیم به حق رسول و به خلق عظیم.

و رک: ص ۹۰ س ۱۶.
۲. سورهٔ ۱۴ (ابراهیم) آیهٔ ۷: هر آینه اگر شکرگزارید، هر آینه نعمت شما را زیاده کنم (پروردگار). قس:
(شکر نعمت نعمتت افزون کند).
۳. (افامث) از سهر سهراً به معنی شب بیدار ماند. ساهره، زمین و روی زمین (الارض او وجهها. قاموس)
به مناسبت آنکه عملش در نباتات شب و روز دایر است. (اقرب الموارد). ۳. ق: بر شاه راه.
۵. خاک رنگ و گردآلود. (غیات). *. ط، ب: چرخ
۷. پادشاه بزرگ داشته، نیرو دادهٔ پیروزمند یاری شده، (موجب) سرافرازی دولت و دین... پادشاه کوهستان
(جبال)، که خدا زیست او را دراز کناد و بالا رفتن وی را به سوی رفعت (رک: ح ۱۷ ص ۸۷) پیوسته داراد.
۸. به کسر اول، پرده برای دختران در گوشهٔ خانه، و هر آنچه به دیدن نیاید از خانه و جز آن، جمع: خدور و
خدار. (منتهی الارب). ۹. خدای، بلندی (قدر) او اپیوسته کناد.
۱۰. اصل هر چیز و خالص و خلاصهٔ آن. میانه و وسط.
۱۱. به سوی خود کشنده. لشکرگران رو از جهت کثرت و انبوهی: ۱۲. حمله برنده.
۱۳. آفتاب دولت و دین، روشنایی اسلام و مسلمانان، که پیروزی او بزرگ و قوی باد. (استاد عبدالعظیم

Fig. 3: *A sample text using our new layout analysis model; note the clean distinction between main text (purple) and footnotes (blue).*

More data annotation and modeling work needs to be done, however, to improve the model's accuracy on much smaller page regions, such as footnotes, catchwords, and running titles.

From Printed Book to Digital Text: A User's Guide to Digital Text Creation Workflow in eScriptorium

In the following sections, we will outline our workflow for digital text creation, drawing upon our insights from the last two years during which we have used eScriptorium for various tasks. eScriptorium can handle a wide range of use case scenarios and source materials, with the amount of user labor required depending on several factors. Potential issues in eScriptorium and their solutions will be addressed in more detail in the final section of this paper.

As noted above, our primary focus for this project has been Arabic and Persian texts, which, along with Urdu, represent the majority of Arabic-script materials in print. Ottoman Turkish is also represented, despite the termination of the use of Arabic script in the 1920s, though thus far it has only featured peripherally in our work (a situation to be rectified in Phase II). Where appropriate, we will note particularities of different languages in terms of their availability or print characteristics, but in general the workflow and issues laid out here apply across the spectrum of Arabic-script documents.

The following steps will take the user through a typical transcription process based on a typographically printed book (i.e., *not* a manuscript or lithograph):

1. *Select source material*: eScriptorium supports PDF import, individual image file import, and IIIF import. Our segmentation and transcription models are robust and can deal with varied DPI ranges, typographic quality and diversity, etc. The user should consult the discussion below for possible issues and limitations, along with a look at our own process of corpus creation (which entailed digital text production at scale, starting with the physical scanning of books themselves). For file naming, it is important that if the original file name contains diacritics or non-English characters they should be replaced, as their presence will interfere with the upload feature.

2. *Upload segmentation and transcription models*: At this stage in eScriptorium's development, users need to manually upload models for transcription and segmentation by clicking on the 'My Models' tab in the upper right-hand corner of the welcome screen and following the required steps. The models will then be available for use after creating a document and uploading images.

3. *Create a project and document*: A project contains individual documents and serves to organize them. Projects can be shared among users.

4. *Import images*: Individual image files can be dragged and dropped into the upload box demarcated by a dashed line. PDFs should be uploaded by clicking the 'Import' button and navigating to the appropriate file on the user's computer. This is the same path for IIIF and XML transcription import where applicable (PDFs should not be dragged and dropped in the box).

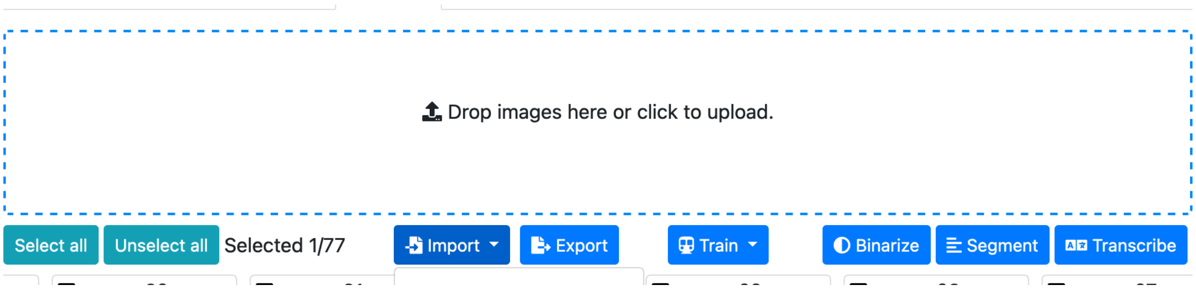
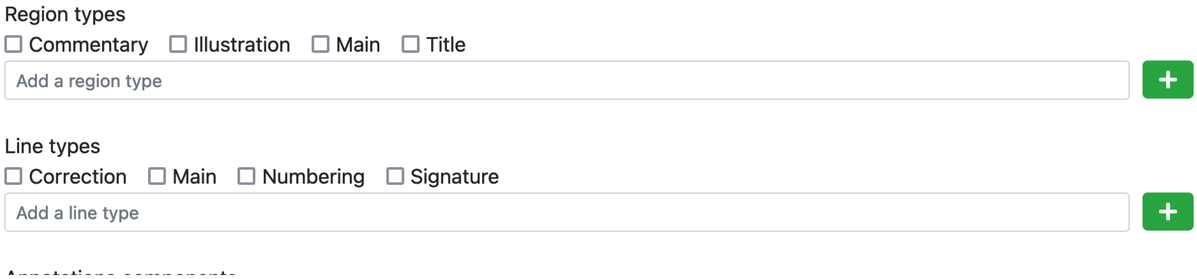


Fig. 4: Import options.

The time required for image import will vary depending on the size of the files. The user should wait until all images have successfully imported before running processes on them. It is important to note that at present there is an upward limit on PDF import size, as only files below 150 megabytes are supported. Larger files will need to be split into constitutive parts and then uploaded into the same document consecutively.

5. *Prepare regions analysis/ontology*: It is generally best to prepare regions analysis at this stage in the process. Regions analysis is used to identify physically discrete parts of the text on the page, such as main text, running titles, footnotes, and so forth, which will then allow the user to include or exclude those sections during the export process. This is an especially important function if one wishes to strip away the editorial apparatus (footnotes, indices, etc.) and only keep the original edited text. The specific regions to be added to a document can be specified and activated in the 'Ontology' section, which also permits designation of line types. These can be added by clicking on an individual segmented line, not via the regions analysis mode. It is important to click 'Update' after adding any new ontology so that the region and line types can be applied to the page images.



Region types

□ Commentary □ Illustration □ Main □ Title

Add a region type

Line types

□ Correction □ Main □ Numbering □ Signature

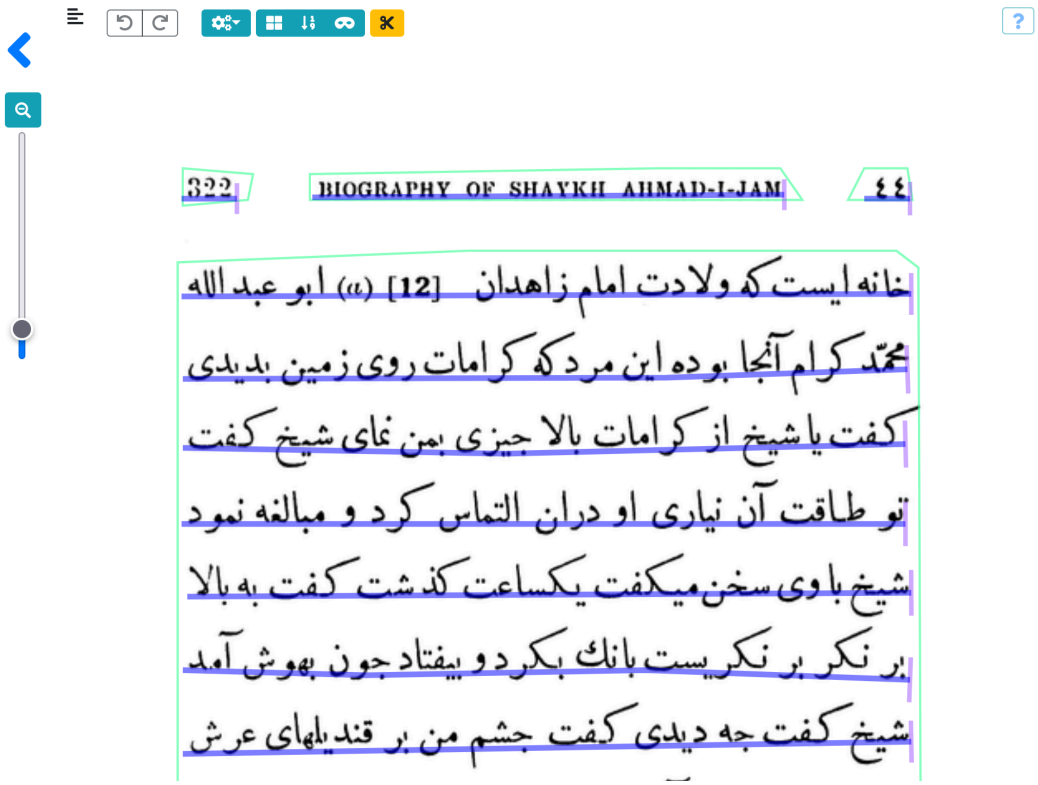
Add a line type

Fig. 5: The two types of ontologies: region level and line level.

6. *Begin segmentation*: Once the images have all been successfully imported, the user can select images to be segmented by clicking on 'Select all' or by clicking on the desired individual images, then clicking the 'Segment' button and selecting the desired model.

7. *Monitor segmentation progress*: Once segmentation has begun, each image will show a yellow flashing bar indicating that the process is still running. This bar will disappear once the process has completed. The user should wait until the segmentation processes have completely resolved before moving on to transcription. While waiting for this process to complete, it is possible to navigate away from the page entirely, or the user might choose to begin reviewing line segmentation quality.

8. *Review line segmentation quality*: It is always a good idea to check the overall quality of the segmentation before proceeding to the next step. Review of segmentation can be done by clicking through to 'Edit' or by clicking on the 'Edit' bar visible in each individual image icon. Segmentation lines can first be reviewed for line continuity as well as overlap of beginning and end points. Discontinuous lines can be joined by holding shift, clicking the relevant lines, then pressing 'J.' Lines can be deleted, redrawn, and manually dragged forwards or backwards to suit the line lengths of the document, should the automatic segmentation not fully capture each line in its entirety.

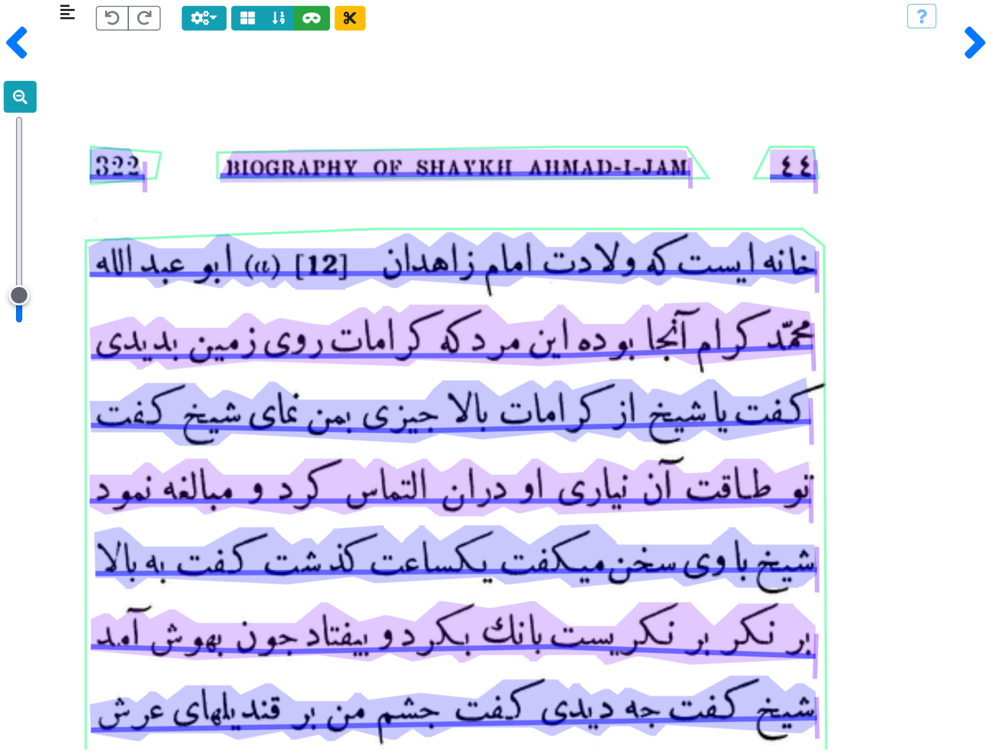


322 BIOGRAPHY OF SHAYKH AHMAD-I-JAM ٤٤

خانه ایست که ولادت امام زاهدان [12] (a) ابو عبد الله
محمّد کرام آنجا بوده این مرد که کرامات روی زمین بدیدی
کفت یا شیخ از کرامات بالا چیزی بمن نمای شیخ کفت
تو طاقت آن نیاری او دران التماس کرد و مبالغه نمود
شیخ با وی سخن میکفت یکساعت کذشت کفت به بالا
بر نکر بر نکریست بانك بکرد و بیفتاد چون بهوش آمد
شیخ کفت چه دیدی کفت چشم من بر قندیلهای عرش

Fig. 6: Segmentation lines after automatic transcription, displayed for review and editing.

9. *Review segmentation masking quality*: eScriptorium automatically generates masks around the letterforms associated with each segmentation line. These can be viewed by clicking the ‘Toggle line masks and stroke widths’ button in Edit view (see Fig. 7 below). Masks can then be evaluated for their coverage of letter forms and manually adjusted, either through adjusting the baseline or manually correcting individual points in the masking. Note that short vowels and orthographic features will in many cases be excluded by the masking.

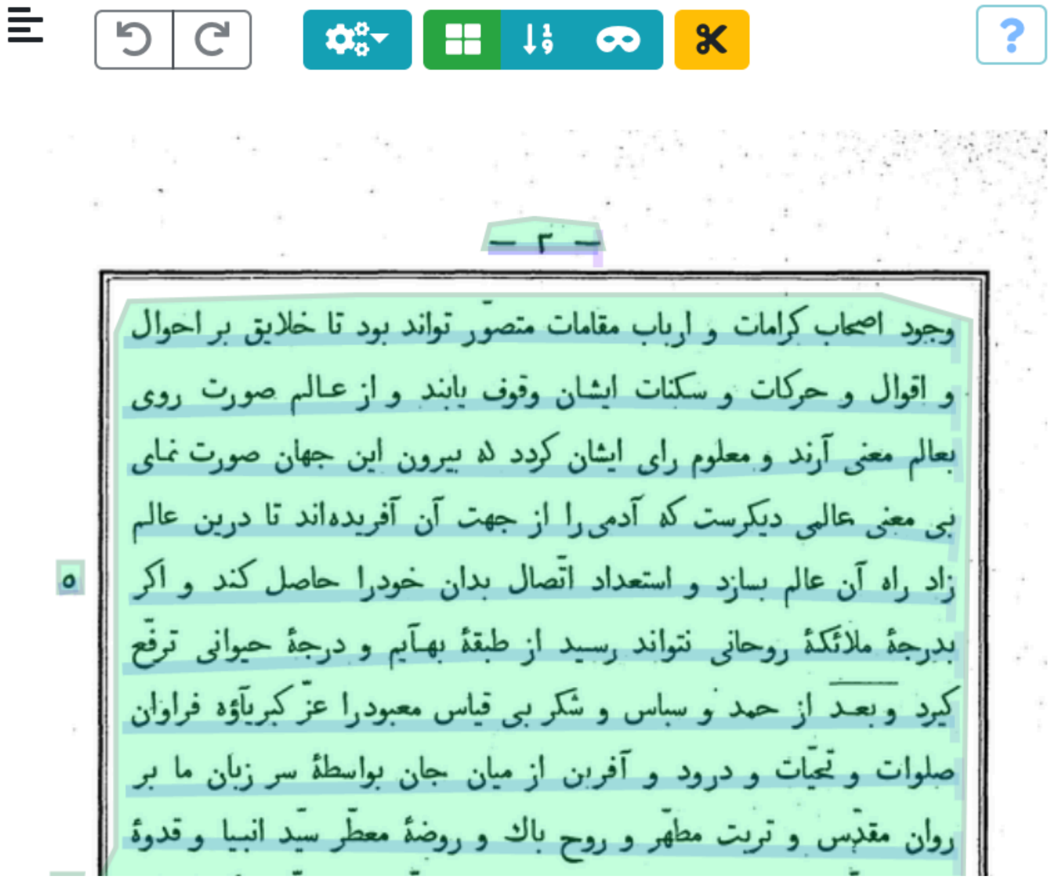


322 BIOGRAPHY OF SHAYKH AHMAD-I-JAM ٤٤

خانه ایست که ولادت امام زاهدان [12] (a) ابو عبد الله
محمّد کرام آنجا بوده این مرد که کرامات روی زمین بدیدی
کفت یا شیخ از کرامات بالا جیزی بمن نمای شیخ کفت
تو طاقت آن نیاری او دران التماس کرد و مبالغه نمود
شیخ با وی سخن میکفت یکساعت کذشت کفت به بالا
بر نکر بر نکریست بانك بکرد و بیفتاد جون بهوش آمد
شیخ کفت جه دیدی کفت جشم من بر قندیلهای عرش

Fig. 7: Masking display.

10. *Review regions analysis*: The quality of the regions analysis, both in terms of coverage (i.e., adequate distinction of regions) and in terms of ontology designation (i.e., adequate recognition of region types) can be assessed at this stage as well. The user may prefer to wait until after transcription is completed, as the accuracy or lack thereof of the regions analysis will not impact the transcription quality. By clicking on the 'Switch to region mode' icon in the 'Edit' view, the segmentation view will display automatically generated regions and allow the user to delete, add, modify, and tag regions.

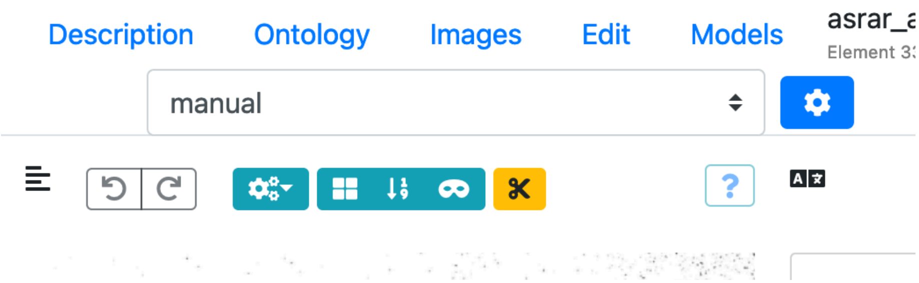


— ٣ —

وجود اصحاب کرامات و ارباب مقامات متصوّر تواند بود تا خلایق بر احوال
و اقوال و حرکات و سکنات ایشان وقوف یابند و از عالم صورت روی
بعالم معنی آرند و معلوم رای ایشان کردد که بیرون این جهان صورت نمای
بی معنی عالمی دیکرست که آدمی را از جهت آن آفریده‌اند تا درین عالم
٥ زاد راه آن عالم بسازد و استعداد اتّصال بدان خودرا حاصل کند و اکر
بدرجهٔ ملائکهٔ روحانی نتواند رسید از طبقهٔ بهآیم و درجهٔ حیوانی ترفّع
کیرد و بعد از حمد و سپاس و شکر بی قیاس معبودرا عزّ کبریآؤه فراوان
صلوات و تحیّات و درود و آفرین از میان جان بواسطهٔ سر زبان ما بر
روان مقدّس و تربت مطهّر و روح پاك و روضهٔ معطّر سیّد انبیا و قدوهٔ

Fig. 8: Regions analysis display.

11. *Run the transcription process*: Transcription is done in the same manner as segmentation. First, the desired pages should be selected. Then, select a transcription model by clicking on ‘Transcribe.’ The transcription will now appear in the edit section. The user may need to select the transcription model used in a drop-down menu visible at the center of the following figure if the program has defaulted to manual view.



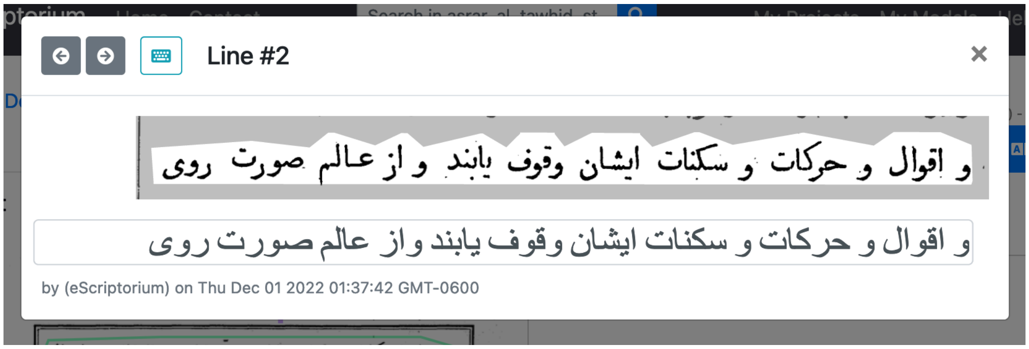
Description Ontology Images Edit Models

asrar_a
Element 3:

manual

Fig. 9: The drop-down menu for the transcription layer.

12. *Review transcription*: The quality of the transcription can be reviewed and corrected line-by-line in the transcription pane. Each edit to a transcription line will be registered with a timestamp and the username of the editor.

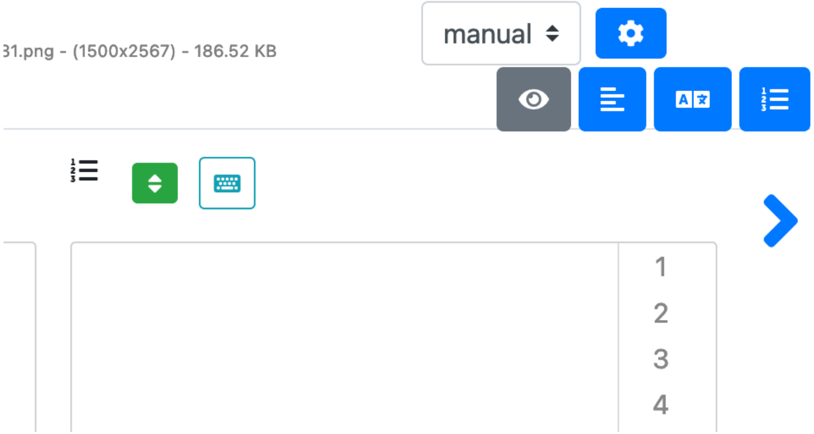


و اقوال و حرکات و سکنات ایشان وقوف یابند واز عالم صورت روی

by (eScriptorium) on Thu Dec 01 2022 01:37:42 GMT-0600

Fig. 10: Transcription review pane.

13. *Review line order*: While this step can be done after reviewing segmentation quality, it is easier to do once transcription has been completed, with line order not otherwise impacting transcription quality. The user should first click the 'Text' icon in the top right-hand corner of Edit view, then click the 'Toggle sorting mode' icon, which will allow the rearrangement of individual lines in the far right-hand pane (see Fig. 11 below). The user can keep track of the correct line arrangement using the other editing panes to the left. For some particular issues that can arise vis-a-vis line order and regions analysis, see the more detailed discussion below in the following section of this paper.



31.png - (1500x2567) - 186.52 KB

Fig. 11: Text pane opened in order to review and edit line order.

14. *Export the text:* Once the user has completed reviewing and editing the transcribed text, the text can be exported in the desired format by clicking on the desired images and then clicking on ‘Export.’ The user has several options of export format (plain text, ALTO, Page XML, OpenITI mARkdown, and OpenITI TEI XML). The transcription model used (if applicable) should also be selected, as well as the specific regions intended for export. If the ‘Include images’ option is selected, then PNG files of each individual page image will be included in the final export. This option is not available for plain text. The exported text is now ready for whatever end use case the user has in mind.

Issues and Fixes in the Digital Text Creation Process

In this section, we will explore some of the processes and possible issues of digital text creation in more detail, both at the level of source selection and preparation as well as at the level of using eScriptorium. Some of these processes have been important to our own digital text production and corpus pilot production (in particular, the question of sources and of book scanning). This section will also feature an overview of our current CARs. Lastly, we will highlight potential issues the user might encounter while working with typographic and layout features of certain texts and their possible solutions.

a. Issues in Locating and Evaluating Sources

The first step in creating any corpus is locating the required material, either in some preexisting digital format or in a physical book form. Even at this step, the user should consider an end goal for the corpus creation. If one’s end-use goal is the creation of a digital critical edition of a short text found in the manuscript tradition, the necessary corpus is likely to be small and dependent upon manual transcription. A user interested in the application of computational methods across a large corpus of genre-specific texts will require a different sort of corpus, and will be more dependent upon automatic transcription. Such a user will also need to be able to verify accuracy rates across texts.

While our primary focus in this white paper is the use of typographically printed texts, we will periodically touch upon lithographs and manuscripts given that eScriptorium is designed for use with both typographic print and handwritten texts. In the future, as part of *OpenITI AOCP Phase II*, we will improve HTR for Arabic-script languages and will produce a second white paper incorporating those developments into detailed digital manuscript corpus creation workflows.

For both typographically printed and handwritten texts in Arabic script, the internet is a vast and important source. A simple text search using a commercial search engine will often turn up texts that have been digitized by entities based out of the Middle East. The legality of these texts may be very much in question, raising issues for users that relate to end goals, in particular whether the resulting electronic texts will be released to the wider public, whether they will include editorial apparatuses, and so forth. Works in Arabic are the easiest to obtain, though it should be kept in mind that available editions online may not necessarily be of the highest quality. In fact, some of the most commonly represented publishers in online materials are not known for high degrees of critical accuracy. Once again, one's end goals are important here: large-scale computational questions can often be answered without reliance on one hundred percent accurate OCR or perfect critical editions. Small corpora of a more traditional philological orientation will require greater particularity in text selection, probably with recourse to physical sources that require digitization by the user.

Besides Arabic, other languages have their own special challenges. Repositories like Hathitrust have a large number of Ottoman Turkish texts, but most texts found 'in the wild' using a search engine are likely to be recent romanizations. Persian books in the public domain are almost all lithographed, and many Urdu texts available online are also either lithographs or photocopies of handwritten pages. Such texts, whether they are in *nasta'līq* script or highly vocalized *naskh*, will respond poorly to our current off-the-shelf generalized OCR models. They are prime targets for our ongoing refinement of generalizable HTR models for Arabic script, and are good candidates for individual model generation using manual transcriptions. That said, it is important to keep the distinction between typographic print and handwritten print in mind when collecting texts for a corpus. Unfortunately, while it is easy to tell the difference when one has a volume in hand, library catalogs do not always indicate the presence of lithography, and the user dependent upon Interlibrary Loan might end up with volumes of limited usability as a result. It is often possible to predict the likelihood that a volume is lithographed, but such a guess is rarely foolproof. Both typographic and lithographed print overlapped one another across the Islamicate world well into the twentieth century.

b. *Scanning Printed Books*

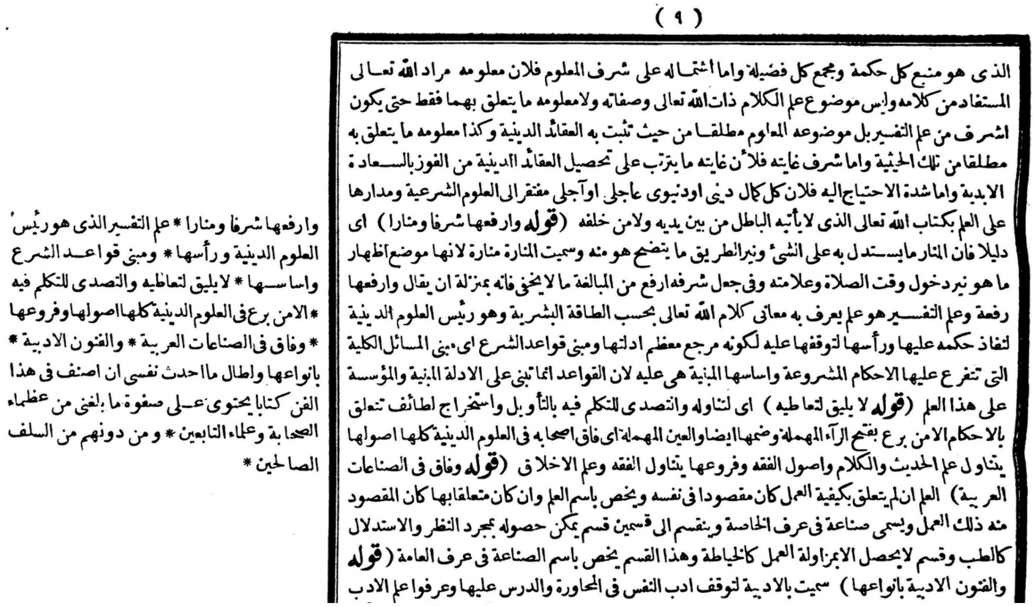
We adopted a simple set of standards for in-house scanning. All scans were made in color, they were not binarized, and they were captured at 300 DPI. In practice, our OCR model also worked well on scans at substantially lower resolutions, especially at later stages in its training. The model also processes black-and-white, greyscale, and color images equally well in terms of the resulting OCR accuracy. Print quality, especially for older, more fragile texts, proved to be a much greater factor than scan quality in determining the accuracy of the OCR transcription–see our discussion below for more detail regarding this problem and potential work-arounds.

As part of our workflow, we almost always scanned the entirety of books. Most of our scanned files contain editor's introductions, tables of contents, indices, etc. along with the main text of a book. We included all of the modern editorial apparatus in our scanning because we wanted the option in the future to use these scans to attempt to train models that can identify this copyrighted material for automatic exclusion (as we have done for footnotes, for example). Scanning large amounts of copyrighted material, however, does require manually exclusion of the copyrighted pages prior to exporting the text—a process which only requires you to uncheck individual page boxes when exporting your transcription. This process is not terribly time consuming for one text, but could become so for larger-scale corpus production.

c. Outstanding Issues and Areas of Ongoing Progress in Segmentation and Regions Analysis

Segmentation and transcription are among the more variable processes in the corpus production pipeline, as the accuracy of the models we have thus far developed depend upon a number of factors. While some manuscript hands may give surprisingly decent results using OCR models, most will not, and even those that do will require extensive manual correction for any use case. Automatic segmentation will sometimes work for manuscripts, but in most cases, manually segmenting a document is easier than correcting automatic segmentation. The remainder of this section will focus entirely on printed texts.

While our automatic segmentation is generally quite good, we are continuing to refine our automatic regions-analysis models. We faced several questions in articulating the parameters of regions analysis for Arabic-script books. How could we best capture the diversity of layout forms in a way that was generalizable across texts? For instance, many nineteenth-century books, emulating manuscript forms, possess different sorts of marginalia which might be additional discrete texts (e.g., discrete *sharḥs* or *ḥashiyyas*), notes, or other additions. These sometimes served as the functional equivalent of footnotes. Because commentary proper is visually indistinguishable from other marginalia features, we decided upon a single ‘Marginal material’ category to capture all of these layout features across different books.



(٩)

الذى هو منبع كل حكمة ومجمع كل فضيلة واما اشتماله على شرف المعلوم فلان معلومه مراد الله تعالى
المستفاد من كلامه وليس موضوع علم الكلام ذات الله تعالى وصفاته ولا معلومه ما يتعلق بهما فقط حتى يكون
اشرف من علم التفسير بل موضوعه المعاوم مطلقا من حيث تثبت به العقائد الدينية وكذا معلومه ما يتعلق به
مطلقا من تلك الحيثية واما شرف غايته فلأن غايته ما يترتب على تحصيل العقائد الدينية من الفوز بالسعادة
الابدية واما شدة الاحتياج اليه فلان كل كمال دينى اودنيوى عاجلى اوآجلى مفتقر الى العلوم الشرعية ومدارها
على العلم بكتاب الله تعالى الذى لا يأتيه الباطل من بين يديه ولا من خلفه (**قوله** وارفعها شرفا ومنارا) اى
دليلا فان المنار ما يستدل به على الشئ ونير الطريق ما يتضح هو منه وسميت المنارة منارة لانها موضع اظهار
ما هو نير دخول وقت الصلاة وعلامته وفى جعل شرفه ارفع من المبالغة ما لا يخفى فانه بمنزلة ان يقال وارفعها
رفعة وعلم التفسير هو علم يعرف به معانى كلام الله تعالى بحسب الطاقة البشرية وهو رئيس العلوم الدينية
لنفاذ حكمه عليها ورأسها لتوقفها عليه لكونه مرجع معظم ادلتها ومبنى قواعد الشرع اى مبنى المسائل الكلية
التى تتفرع عليها الاحكام المشروعة واساسها المبنية هى عليه لان القواعد انما تبنى على الادلة المبنية والمؤسسة
على هذا العلم (**قوله** لا يليق لتعاطيه) اى لتناوله والتصدى للتكلم فيه بالتأويل واستخراج لطائف تتعلق
بالاحكام الا من برع بفتح الراء المهملة وضمها ايضا والعين المهملة اى فاق اصحابه فى العلوم الدينية كلها اصولها
يتناول علم الحديث والكلام واصول الفقه وفروعها يتناول الفقه وعلم الاخلاق (**قوله** وفاق فى الصناعات
العربية) العلم ان لم يتعلق بكيفية العمل كان مقصودا فى نفسه ويخص باسم العلم وان كان متعلقا بها كان المقصود
منه ذلك العمل ويسمى صناعة فى عرف الخاصة وينقسم الى قسمين قسم يمكن حصوله بمجرد النظر والاستدلال
كالطب وقسم لا يحصل الا بمزاولة العمل كالخياطة وهذا القسم يخص باسم الصناعة فى عرف العامة (**قوله**
والفنون الادبية بانواعها) سميت بالادبية لتوقف ادب النفس فى المحاورة والدرس عليها وعرفوا علم الادب

وارفعها شرفا ومنارا * علم التفسير الذى هو رئيس
العلوم الدينية ورأسها * ومبنى قواعد الشرع
واساسها * لا يليق لتعاطيه والتصدى للتكلم فيه
* الا من برع فى العلوم الدينية كلها اصولها وفروعها
* وفاق فى الصناعات العربية * والفنون الادبية *
بانواعها واطال ما احدث نفسى ان اصنف فى هذا
الفن كتابا يحتوى على صفوة ما بلغنى من عظماء
الصحابة وعلماء التابعين * ومن دونهم من السلف
الصالحين *

Fig. 12: An example of ‘marginal material’ (left) often encountered in older print editions in which aspects of manuscript production were preserved to some extent.

Similar issues arose for features like titles, illustrations, decorations, and footnotes. We are now working to resolve an issue that arose due to the nature of our corpus selection for the initial training data. The works we chose were spread out across the last two hundred years of print, with only some coverage from the last few decades. However, in many recent works—a number of examples of which we have digitized and processed as part of our corpus pilot project—the editorial apparatus consists largely of abundant footnotes, which in extreme instances can take up an entire page of text. These sections are distinguished from the main text only by a thin right-justified line. As was mentioned previously, we are currently testing a new layout model which accurately distinguishes between different regions on a page. For now, however, in cases of such prolix footnotes, manual correction will almost certainly be required and is vital for users who want to export only the main text to the exclusion of footnotes.

d. Outstanding Issues and Areas of Ongoing Progress in Transcription

For the vast majority of typographically printed material, the OCR models which we have developed will have CARs in the high nineties (Fig. 4). Exceptions include some of the very oldest and poorly represented Arabic-script typefaces, as well as most Urdu typefaces (the subject of the next phase of our development plan). Accuracies are calculated using both the raw manual transcriptions and an ‘extranormalized’ version of the transcription that merges the many variant forms of Arabo-Persian characters, numerals, and punctuation available in Unicode and ignores tatweel and combining diacritics other than hamza.

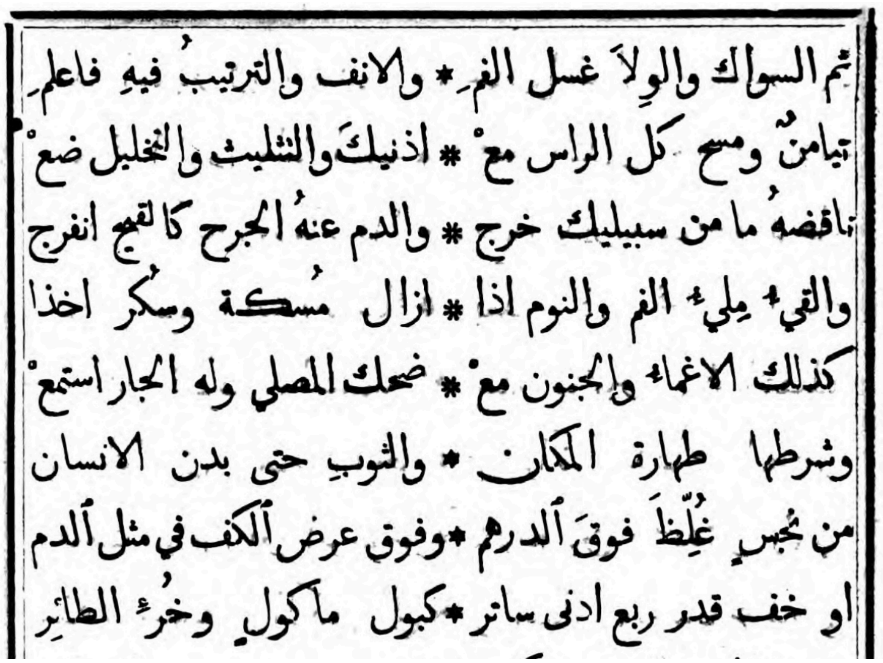
OCR transcription accuracy is dependent upon the quality of the segmentation, and in some cases the user would be well advised to check segmentation before proceeding with transcription, particularly for older texts which might have print quality and page layout issues.

	Typeface Microaverage	
Typeface	**Original**	**Extranormalized**
arabic_algerian_pseudo_maghribi	95.27%	97.70%
arabic_amriki_typeface	92.61%	96.24%
arabic_bulaq_ii_typeface	96.20%	99.10%
arabic_bulaq_i_typeface	90.87%	95.74%
arabic_decotype_typeface	94.56%	97.62%
arabic_lotus_linotype	95.85%	99.02%
arabic_simplified_typeface	89.88%	97.40%
arabic_traditional_typeface	94.82%	97.64%
backwards_ya_typeface	89.60%	94.78%
bulaq_nastaliq_typeface	84.19%	89.37%
eurarabic_i	95.93%	97.89%
ottoman_naskh_ii_typeface	95.36%	97.94%
ottoman_naskh_i_typeface	92.10%	96.25%
persian_amriki	91.73%	96.24%
persian_incunabula	86.48%	89.27%
persian_intertype	92.45%	98.21%
persian_lotus_linotype_typeface	93.14%	98.35%
persian_monotype	93.78%	97.16%
persian_watts_nastaliq_typeface	93.04%	95.38%
persian_watts_typeface	92.91%	97.14%
russian_persian_typeface	95.52%	98.64%
AVERAGE	92.68%	96.53%

Fig 13: Original and extranormalized CARs for Arabic and Persian typefaces as of September 29, 2022.

In some cases, peculiarities of layout could cause issues, such as divisions within poetry couplets. Some typefaces, such as the aforementioned Amrīkī typeface popular in the eastern

Levant up until the late nineteenth century (see Fig. 1 and Fig. 14), feature letterforms in which ascenders and descenders dip down below the baseline significantly.



ثم السواك والولاَ غسل الفمِ * والانف والترتيبُ فيهِ فاعلمِ
تيامنٌ ومسح كل الراس معْ * اذنيكَ والتثليث والتخليل ضعْ
ناقضهُ ما من سبيليك خرج * والدم عنهُ الجرح كالقيح انفرج
والقيء ملءِ الفم والنوم اذا * ازال مُسكة وسُكر اخذا
كذلك الاغماء والجنون معْ * ضحك المصلي وله الجار استمعْ
وشرطها طهارة المكان * والثوبِ حتى بدن الانسان
من نجسٍ غُلِّظَ فوقَ الدرهم * وفوق عرض الكف في مثل الدم
او خف قدر ربع ادنى ساتر * كبول مأكولٍ وخرءِ الطائرِ

Fig. 14: An example of an older typeface, as well as of lower quality print–note the smudges, variable line widths, and other irregularities.

Older texts are also more likely to display issues of print quality and of book preservation. Many nineteenth and early twentieth-century printers were operating in suboptimal conditions, with type being used well past its ‘expiration date,’ sometimes with lower quality ink and paper. As such, letterforms can be indistinct or smudged. The added component of age further obscures text for both human and mechanical readers.

Users might have several ways of dealing with low CARs on typefaces that are not represented in the original training data set. If the user can operate with dirtier-than-optimal data, errors within the OCR may not be worth correcting. If the user does wish to correct those errors, manual correction page-by-page might be the easiest way to do so depending on the length of the text. Finally, it is possible to train a new model, or series of models, improving the in-document CAR by deploying the underlying transcription model, training a new model using the corrected transcriptions, and then applying the new and improved model to another segment of text, continuing the process until the desired CAR is achieved. In the case of a physically degraded text, manual correction might be the best option. If the primary issue is the use of an unusual, poorly represented typeface, a boosted model using in-document training data might prove very useful and save human labor in the long run.

There are certain aspects that users should keep in mind when using our models. One of the major dilemmas we faced in this project was the question of what to do with vocalization and orthographic signs. Only a small percentage of printed books have full vocalization and the entire range of possible orthographic markers, with a somewhat larger percentage of books containing sporadic vocalization (e.g., Qur'an citations). To complicate matters further, the relationship between vowels and primary letter forms has varied greatly over the course of print history, due to the material exigencies of different print technologies. In some typefaces, the vowels are spaced evenly above the line. In others, they nest along the letterforms. In terms of end-use scenarios, some members of the team felt that inclusion of vowels was important, while others preferred stripping them out if possible. We went through several rounds of deliberation on the matter and tested different configurations of training data and output. It was the technical side of things that ultimately determined our non-inclusion of vocalization and the majority of orthographic symbols, as we found this to be the best approach to improving overall CAR. However, this decision on our part means that users should be aware of two things. Firstly, if inclusion of vocalization and full orthographic repertoire is important to one's end-case goals, manual correction and/or additional model training will be required. Secondly, in some typefaces the presence of extensive vocalization (such as in Qur'an citations) can generate errors that would not otherwise exist.

Finally, while we have achieved a very wide coverage of the last two hundred years of Arabic-script print, users should be aware that gaps in our coverage remain. We will fill some of these gaps in coming months and years, but others will probably remain for the foreseeable future due to the marginality and paucity of the relevant material. There are also rather marginal typefaces–mostly early, Western European 'Eurabic' typefaces–for which our accuracy rates remain in the low nineties. We had to make decisions about which of these more poorly represented typefaces to include, knowing from early on that all typefaces prior to the twentieth century posed real challenges both in terms of book quality, preservation, and design features. We focused on those with wider representation as they were more likely to be encountered by the average user.

e. Conclusions: Tackling an Edge Case and Its Lessons

For most users, the combination of our models and the eScriptorium platform will result in a smooth and relatively low-labor experience. The precise degree of work required will of course vary depending on desired end goals. That said, all users, particularly those working with editions from the late nineteenth and early twentieth centuries, should be aware of the issues we have noted. In such cases a little more clean-up might be required. The only area in which many users are likely to need page-by-page correction is in automatic regions analysis. Even in that case, only texts with the most extensive editorial apparatuses should pose any significant issue. We are currently working on improving our segmentation and layout models, and hope to see much improved results in places where accuracy has continued to lag.

There are cases in which a somewhat more complex operation will be required to ensure that the various sections of text remain in appropriate reading order. The following text is a good example. First of all, the lines of the marginal commentary require manual correction, as can be seen in the following figure:

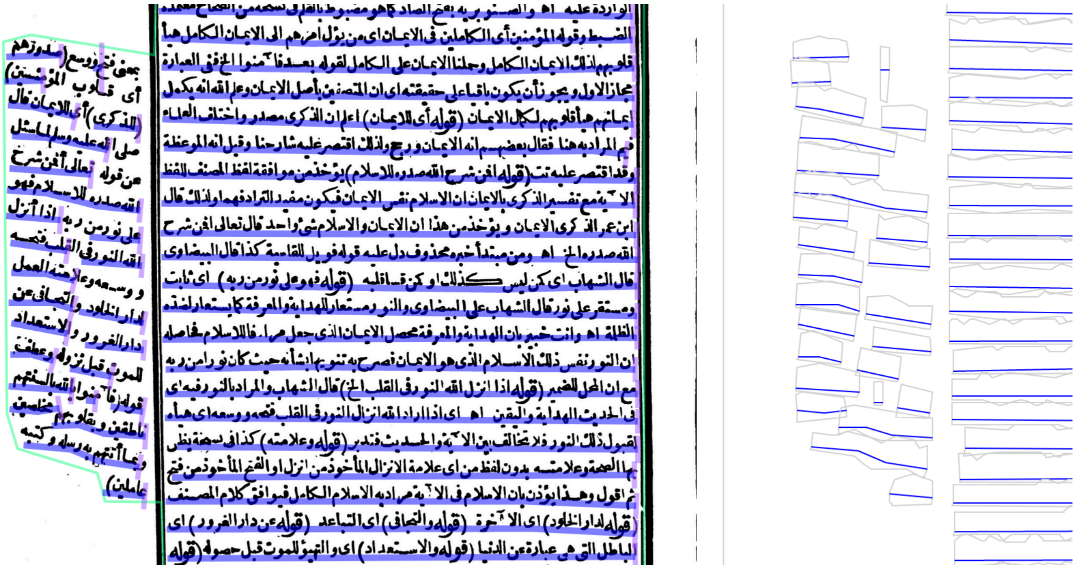


Fig. 15: A nineteenth-century printed book with slanting marginal material and resulting segmentation errors.

The presence of parallel marginal material also resulted in the marginal material region not being distinguished from the main text. This has caused inaccuracies in the line numbering order.

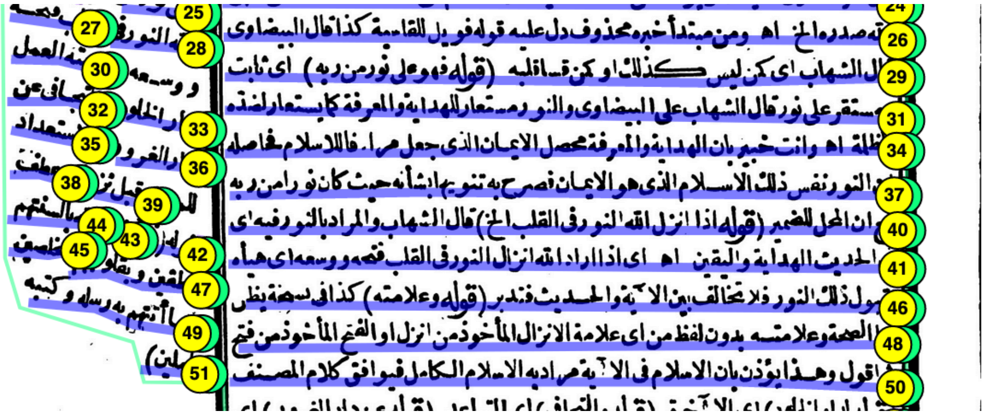


Fig. 16: Jumbled line order caused by parallel texts, prior to correction.

To fix these problems, the user would need to 1) edit the regions by manually deleting, redrawing, and labeling a main text region and a marginal material region, then 2) toggle off regions view and select all the lines on the page, and finally 3) click the ‘Link selected lines to first detected background region' icon, which will associate the lines with their respective regions and ensure that the lines will now be in consecutively numbered order.

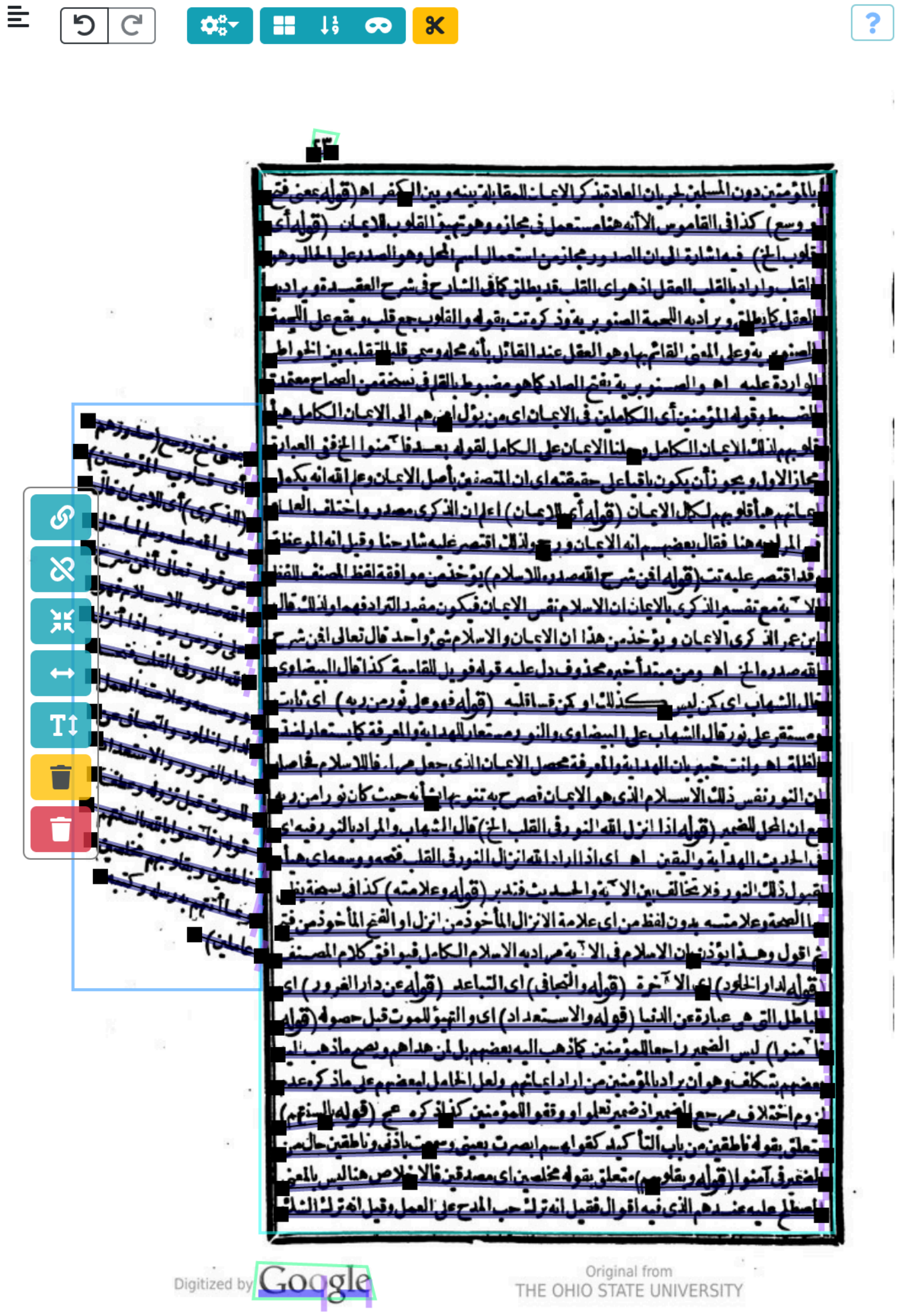
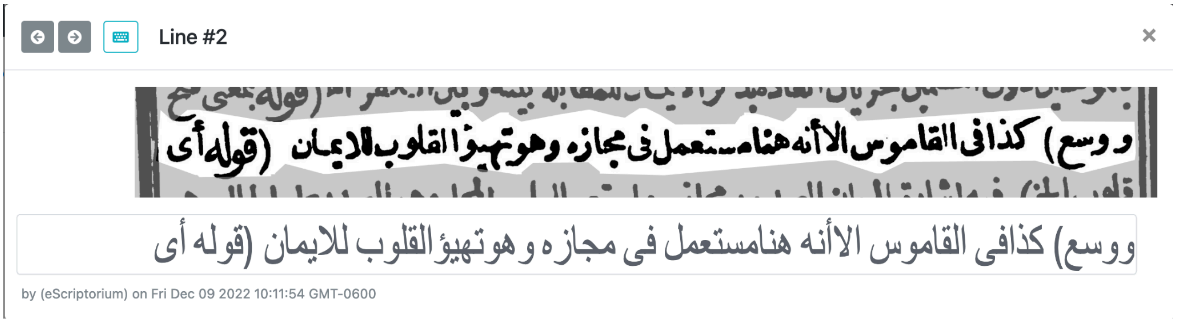


Fig. 17: Highlighted segments ready to be linked to background regions.

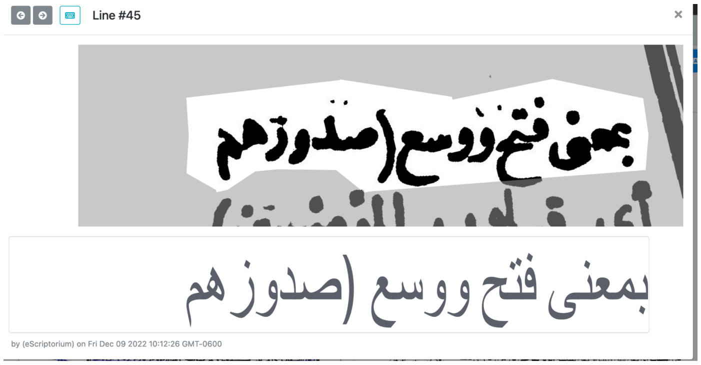
While the preceding exemplifies some of the limitations still existing for digital text creation, we will end on a more positive note, staying with the above 'edge case' text. At first glance, the typography might not inspire confidence in its OCR legibility, but let's have a look at preliminary, uncorrected results when transcription is run on this text:



Line #2

ووسع) كذافى القاموس الاأنه هنامستعمل فى مجازه و هوتهيؤ القلوب للايمان (قوله أى

by (eScriptorium) on Fri Dec 09 2022 10:11:54 GMT-0600



Line #45

بمعنى فتح ووسع (صدوزهم

by (eScriptorium) on Fri Dec 09 2022 10:12:26 GMT-0600

Fig. 18: Uncorrected initial OCR results for a sample line from (above) the main text of our nineteenth-century document and (below) from the slanting marginal material.

The above figures exemplify the degree of accuracy we have obtained, as well as the current limits at the level of segmentation. The initial segmentation, as seen in Fig. 15, of the marginal material was very poor. The solution in this case was to simply bulk delete the original lines and redraw new lines manually. The poor segmentation can be attributed here to the slant of the lines, a relatively infrequent feature in our training data sets due to its chronological limitation to the nineteenth century. However, note that while this particular book is hardly a model of typography clarity, and in fact is at the lower end of print quality a user is likely to encounter, the transcription model has captured the letterforms with very few errors indeed, even in the case of the marginal commentary. Texts of this sort will likely always require some degree of manual postprocess correction–*ṣudūruhum* really does look like *ṣudūzuhum* in the marginal material line above. The human reader will recognize that the apparent dot above the third-to-last letter form is a printing mistake and will correct it accordingly.