DHQ: Digital Humanities Quarterly

Preview
2023
Volume 17 Number 3

Automated Transcription of Gə'əz Manuscripts Using Deep Learning

Samuel Grieggs <sgrieggs_at_nd_dot_edu>, University of Notre Dame
Jessica Lockhart <jessica_dot_lockhart_at_utoronto_dot_ca>, University of Toronto
Alexandra Atiya <alexandra_dot_atiya_at_mail_dot_utoronto_dot_ca>, University of Toronto
Gelila Tilahun <gelila_dot_tilahun_at_utoronto_dot_ca>, University of Toronto
Suzanne Akbari <sakbari_at_ias_dot_edu>, Institute for Advanced Study, Princeton, NJ
Eyob Derillo <eyob_dot_derillo_at_bl_dot_uk>, SOAS, University of London
Jarod Jacobs <jarod_dot_jacobs_at_gmail_dot_com>, Warner Pacific College
Christine Kwon <ckwon_at_nd_dot_edu>, University of Notre Dame
Michael Gervers <m_dot_gervers_at_utoronto>, University of Toronto
Steve Delamarter <sdelamar_at_georgefox_dot_edu>, George Fox University
Alexandra Gillespie <alexandra_dot_gillespie_at_utoronto_dot_ca>, University of Toronto
Walter Scheirer <walter_dot_scheirer_at_nd_dot_edu>, University of Notre Dame

Abstract

This paper describes a collaborative project designed to meet the needs of communities interested in Gə'əz language texts – and other under-resourced manuscript traditions – by developing an easy-to-use open-source tool that converts images of manuscript pages into a transcription using optical character recognition (OCR). Our computational tool incorporates a custom data curation process to address the language-specific facets of Gə'əz coupled with a Convolutional Recurrent Neural Network to perform the transcription. An open-source OCR transcription tool for digitized Gə'əz manuscripts can be used by students and scholars of Ethiopian manuscripts to create a substantial and computer-searchable corpus of transcribed and digitized Gə'əz texts, opening access to vital resources for sustaining the history and living culture of Ethiopia and its people. With suitable ground-truth, our open-source OCR transcription tool can also be retrained to read other under-resourced scripts. The tool we developed can be run without a graphics processing unit (GPU), meaning that it requires much less computing power than most other modern AI systems. It can be run offline from a personal computer, or accessed via a web client and potentially in the web browser of a smartphone. The paper describes our team's collaborative development of this first open-source tool for Gə'əz manuscript transcription that is both highly accurate and accessible to communities interested in Gə'əz books and the texts they contain.

Abstract

ጥልቅ እውቀትን ለረቂቅ ጽሁፎች ስለመጠቀም

ሳሙኤል ግሪግስ፡ ኖተርዳም ዩኒቨርሲቲ፤ ጀሲካ ሎክሀርት፡ቶሮንቶ ዩኒቨርሲቲ፤ አሌክሳንደራ አትያ፡ ቶሮንቶ ዩኒቨርሲቲ፤ ገሊላ ጥላሁን፡ ቶሮንቶ ዩኒቨርሲቲ፤ ሱዛን ኮንክሊን አክባሪ፡ አድቫንስድ ጥናት ኢንስቲትዩት፡ ፕሪንስተን ኒው ጀርሲ፤ ኢዮብ ዴሪሎ ሶ.አ.ስ. ለንደን ዩኒቨርሲቲ፤ ጃሮድ ጃኮብስ፡ ዋርነር ፓሲፊክ ኮሌጅ፤ ክሪስቲን ኮን፡ ኖተርዳም ዩኒቨርሲቲ፤ ሚካኤል ጀርቨርስ፡ ቶሮንቶ ዩኒቨርሲቲ፤ ስቲቭ ዴላማርተር፡ ጆርጅ ፎክስ ዩኒቨርሲቲ፤ አሌክሳንድራ ግለስፒ፡ ቶሮንቶ ዩኒቨርሲቲ፤ ዋልተር ሸሪር፡ ኖተርዳም ዩኒቨርሲቲ።

መግለጫ

ይህ ጥናት የሚገልፀው የግዕዝ ቋንቋ ፅሁፍን እና ሌሎች መሰል ትኩረት ያልተሰጣቸውን፣ ባህላዊና እና ጥንታዊ ሥሁፎችን ለመማር ወይም ለጥናት የሚፈልጉ ማህበረሰቦችን ፍላጎት ለማርካት የጥምር የጥናት ቡድናችን ስለቀረፀው ቀላል እና ሁሉም ሊጠቀምበት ስለሚችል መሣሪያ(ዘዴ) ነው።::ይህ መሣሪያ የብራና ፅሁፍን የመሰሉ ረቂቅ ፅሁፎች የተፃፁባቸውን ገፆች ምሥል በማንሳት እና ፊደላትን ለይቶ በሚገነዘብ ጨረር (optical character recognition (OCR)) በመጠቀም ምሥሉን ወደ መደበኛ ወይም ሁለተኛ ፅሁፍነት የመቀየር ችሎታ ያለው ነው። ይህ ኮምፒዩተር ላይ የተመሰረተ ዘዴ ወይም መሣሪያ የግዕዝ ቋንቋን ልዩ ባህርዮች ለይቶ እንዲያውቅ ሲባል ስለቋንቋው ያገኘውን መረጃ ወይም ዳታ የመንከባከብ እና የማከም ሂደቶችን አልፎ አንዴ አንጎል ነርቮች መረብ አሽከርክሪት የሚመስል ኮንቮሉሽናል ሪከረንት ነውራል ኔትዎርክ (Convolutional Recurrent Neural Network) በመያዙ ገጽታዎችን እና ምሥሎችን ወደ ፅሁፍ ይቀይራል። ይህ ለሁሉም ተጠቃሚዎች ክፍት የሆነው ጽሁፍ ለተማሪዎች እንዲሁም ለኢትዮጵያ ጽሁፍ ጥናት ተመራማሪዎች የሚጠቅም ብቃት ያለው እና በቀላሉ በኮምፒዩተር ተፈልጎ ሊገኝ የሚችል ከመሆኑም በተጨማሪ የግዕዝ ጽሁፎቹ የኢትዮጵያን እና የኢትዮጵያን ህዝብ ታሪክና ባህል ግዕዝን በዲጂታል/በኮምፑተር ቀርፆ በማስቀመጥ በቀጣይነት እንዲኖር ያስችላል። አመቺ የሆነ ተጨባጭ ሁኔታ ሲኖር ደግሞ ይህ ለሁሉም ክፍት የሆነ የ OCR የግዕዝን ምስልን ወደ ፅሁፍ የሚቀይር መሣሪያ ወይም ዘዴ ሌሎች ትኩረት ያላገኙ ረቂቅ ፅሁፎችንም እንዲያነብ ተደርጎ ሊሰለጥን ወይም ዲዛይን ሊደረግ ይችላል። ይህ የፈጠርነው መሣሪያ/ዘዴ የተለመደውን ግራፊክስ ፕሮሰሲንግ ዩኒት (GPU) የተባለውን በኮምፕዩተር ምሥሎችን የማንበቢያ እና ማሳለጫ ዘዴ መጠቀም አያስፈልገውም። በዚህም ምክንያት ከሌሎች ዘመናዊ የአርቲፊሻል ኢንተሊጀንስ (AI systems) ዘዴዎች አነፃር ሲታይ ሃይለኛ የኮምፒዩተር አቅም አይፈልግም። ይህንን መሣሪያ/ዘዴ ያለ ኢንተርኔት ወይም በይነ-መረብ ከግል ኮምፒዩተር፣ በኢንተርኔት እንዲሁም ወደፊት ኢንተርኔት ባለው የእጅ ሥልክን በመጠቀም ማስኬድ ይቻላል። ይህ ጥናት የሚገልጸው በአይነቱ የመጀመሪያ የሆነው እና ለሁሉም ክፍት የሆነ እንዲሁም በተገቢ ሁኔታ ጥራቱን ጠብቆ በጥምር ተመራማሪዎቻችን የበለፀገው መሣሪያ/ዘዴ ለማናቸውም በግዕዝ መጽሀፍቶች እና ውስጣቸው በያዙት ፅሁፎች ላይ ጥናት ለማድረግ ለሚፈልጉ ግለሰቦችም ሆኑ ማህበረሰቦች ሁሉ ጠቃሚ መሆኑን ለማስገንዘብ ነው።

Introduction

This paper summarizes an interdisciplinary collaborative project to create an easy-to-use open-source tool that converts an image of a manuscript page written in the historical Ethiopic script of Gə'əz into a transcription using Optical Character Recognition (OCR). 1

OCR tools can transform the life of a text in a digitized manuscript by rendering its images readable and searchable. This ability allows readers to engage directly with the contents of manuscripts and significantly eases the tasks of those describing or cataloging endangered collections or obsolescing files. OCR thus has a pivotal role in the preservation of historical texts into the future. However, recent years and new users have brought into focus a range of ethical challenges in the evolving 2

field of digital preservation of cultural heritage, emphasizing the need for any digitization to serve first and foremost the needs and wishes of the communities who created and safeguard the cultural heritage being preserved [Manžuch 2017] [Liuzzo 2019, 239] [Sutherland and Purcell 2021].

With this in mind we have designed an open-source tool that is accessible outside of a university setting, that can transcribe batches of images with no requirement that the text generated through transcription leave the control or the home environment of the person running the program. Our tool has low operational requirements in terms of computing power, and it does not require an internet connection to run, although we have also created a web interface to demonstrate it on a line-by-line basis. It can be used broadly by students and scholars of Ethiopian manuscripts to create a substantial and computer-searchable corpus of digitized and transcribed Gə'əz texts. 3

Our tool advances AI-driven methods of OCR for manuscripts, through adaptable strategies that can be used to enhance research on other under-resourced textual traditions. The software itself is agnostic to language, and thus, although we have developed it for the specific context of cultural heritage preservation of Gə'əz language texts, we hope it will be of interest to other groups, as it can be retrained easily for other languages and scripts given a new JSON file. In this paper we thus describe our own process for preparing ground-truth, and we include some discussion of how to adapt it to a new context in the final stages of the paper. 4

Background on the Gə'əz Language

Gə'əz is largely considered to fall within the North Ethiopic subgroup of Ethiopian Semitic languages [Weninger 2005, 732]. Other languages in the Ethiopian Semitic language family include Amharic, Tigre, and Tigrinya [Appleyard 2005, 51] [Weninger 2005, 732]. Gə'əz has a distinctive script, which — unlike some other Semitic languages, such as Hebrew and Arabic — is read from left to right. The script's characters are phonetic, and arranged in a table called by members of the Ethiopian community a fidäl, rather than in a sequential alphabet. We reproduce a sample table in Appendix A of a fidäl consisting of 245 characters laid out on thirty-five rows across seven columns with the inclusion of characters specific to Amharic script used in some personal and place names, and additional characters representing labiovelars or semi-vowel glides. Words in Gə'əz script are formed by concatenating characters, and words are separated by a character that looks very similar to the colon punctuation mark, '፡'. Sentences are separated by a 'double-colon' looking character, '።'. Sometimes, the end of a sentence is indicated by '፧'. Semicolons are indicated by the character '፤'. This punctuation system developed over time, and thus older documents often lack the usage found in later documents. The Gə'əz numerical system has different characters for digits from one to nine, separate characters for numbers that are multiples of ten, and a character for the value 'hundred'. 5

Gə'əz served as the dominant written language in the Christian kingdom of Ethiopia from the 3rd century until the mid-16th [Kelly 2020, 25]. Today, Gə'əz continues to be used as a liturgical language and tongue of poetic expression both in the Horn of Africa and in the Ethiopian diaspora. Gə'əz is also associated with a living, though endangered, manuscript tradition; manuscript production in Ethiopia continued robustly into the mid- to late-20th century [Winslow 2015, 7–12]. An estimated 200,000 Gə'əz manuscripts, primarily from the 14th century onwards, still survive in Ethiopia [see an overview in [Nosnitsin 2020, 289]], and are kept by the churches and monasteries of the Ethiopian Orthodox Täwahədo Church, as well as public institutions such as the Addis Ababa University's Institute of Ethiopian Studies (IES) and the National Archives and Library of Ethiopia [Bausi 2015, 47]. 6

The continuity of this living manuscript tradition, however, has become critically endangered in Ethiopia.[1] Recognizing the need to document it, a range of international collaborations since the 1960s have endeavored to preserve records and create facsimile surrogates of Gə'əz manuscripts *in situ* through cataloging, photography, microfilming, and digital images [Stewart 2017]. Our project consulted digital facsimiles of 17 manuscripts imaged in Ethiopia by Donald Davies (1968); the Ethiopian Manuscript Microfilm Library (EMML) (1973–1994); the Endangered Archives Programme EAP432, led by Mersha Alehegne Mengistie (2011); the Ethiopian Manuscript Imaging Project (EMIP, 2005) directed by Steve Delamarter; Michael Gervers and Ewa Balicka-Witakowska's digitizations of books at Gunda Gundē Monastery (2006) in association with HMML; and the project Ethio-SPaRe: Cultural Heritage of Christian Ethiopia – Salvation, Preservation, and Research, led by (2009–2015, funded by the European Research Council), led by Denis Nosnitsin (Universität Hamburg, Hiob Ludolf Centre for Ethiopian and Eritrean Studies). The digitized images we accessed are now under the care of the Hill Museum and Manuscript Library (6 manuscripts); EAP (1 manuscript); EMIP (4 manuscripts), and Ethio-SPaRe (6 manuscripts). See Appendix B for further details on these manuscripts and the availability of their digital surrogates. 7

The digital age has ushered in a new wave of international collaborations and research trips focussing on manuscript documentation and image preservation in Ethiopia. Meanwhile, the digitizing of historical microfilms has come to be a separate and essential source of manuscript images. Some of the most important and early collections, including the UNESCO and Ernst Hammerschmidt–Tanasee projects of the 1960s and some of the EMML projects, are in microfilm form. The pace of microfilm digitization has taken decades longer than expected, and the most important and largest of these collections (the EMML collection) still has a significant percentage of image sets unavailable in any digital form. In certain cases, decades after these projects, the whereabouts of the manuscripts the microfilms represent cannot be confirmed, rendering the microfilm effectively the book's only extant witness. Other important microfilm collections represent Gə'əz manuscripts historically based outside Ethiopia, such as the Library of Congress's microfilmed collection of manuscripts held at Sinai, or the University of Utah's microfilms of the manuscripts of the Ethiopian Orthodox Church in Jerusalem. Digitization of microfilm collections greatly increases their accessibility; it has been a priority for institutions such as HMML and the Bibliothèque nationale de France (BnF), and continues to be a pressing need in the field — bringing attendant needs for metadata and appropriate cataloging. 8

Gə'əz OCR and Our Contribution

A range of resources and digital tools are under development to build upon the above imaging efforts and to facilitate broad access to the precious cultural heritage these manuscripts represent — for a recent overview, see [Liuzzo 2019, xxv–xxxii]. To support these and other projects our work contributes a tool: an open-source OCR transcription algorithm that can be widely used by students and scholars of Ethiopian manuscripts to create a computer-searchable corpus of transcribed and digitized Gə'əz texts, and which can be retrained to serve other historical scripts and languages. 9

Our open-source tool is the first to facilitate end-to-end AI-driven transcription of Gə'əz, but our team has learned from previous projects in developing its unique affordances. Daniel Yacob's work on Unicode standards for Ethiopic languages defined the text characters used for automatic transcriptions of Gə'əz manuscripts, making this entire endeavor feasible [Yacob 2005]. Siranesh Getu Endalamaw's 2016 thesis, which introduced a deep learning-based artificial neural network approach to recognize text in Gə'əz manuscripts [Endalamaw 2016] and recent work by Fitehalew Ashagrie Demilew and Boran Sekeroglu [2019] on Gə'əz character recognition were constrained to the classification of pre-segmented characters. While this is an important step in a text recognition tool, it does not lead to readable text output, and the results are not directly comparable to what we present in this paper. Daniel Mahetot Kassa and Hani Hagras [2018] discuss the issue of segmentation of digitized images into individual Gə'əz characters. Segmentation plays an important role in transcription accuracy, but our tool does not require character-level segmentation. With respect to languages close to Gə'əz, OCR for modern Amharic has also been explored by Fitsum Demissie [2011]. 10

Looking beyond Gə'əz, a number of projects have focused on the transcription of handwritten manuscripts and documents of medieval European and colonial origin. 11

For example, Alrasheed et al. use object detection networks to find and classify individual characters in handwritten seventeenth-century Spanish American notary records [Alrasheed et al. 2019]. The medieval Latin script known as Caroline Minuscule has been often explored due to the high availability of creative commons scans from collections such as e-codices - Virtual Manuscript Library of Switzerland. Long short-term memory recurrent models look at raw image data to transcribe these texts [Hawk et al. 2019]. A psychometric loss that measures the samplewise reading performance of paleographers and incorporates it into the training process improves accuracy on transcription of Latin text in Caroline Minuscule script across a variety of models [Grieggs et al. 2021]. Calamari is a high performance OCR tool that transcribes historical printed texts in both English and German with a high degree of accuracy using a CRNN [Wick et al. 2020].

12 Transkribus is a more complete computer-aided transcription platform that does have some Gə'əz support [Kahle et al. 2017]. While it offers a similar feature set to our tool, there are several key differences. In a broad sense, our software can be run locally, and is free and open-source, allowing users to add functionality as they see fit. Transkribus offers more sophisticated layout analysis tools, but requires the user to upload documents to an external server, raising questions about intellectual property. Transkribus's performance is comparable to the algorithm proposed in this paper under certain conditions. We have included a more detailed comparison of the two tools and performance on our data in the results section below.

13 In building our tool we were aware of some of the ethical challenges that Gə'əz digitization projects face more broadly — particularly with respect to issues of selection, copyright and access, short- and long-term stewardship, and appropriate uses of those images once created.[2] As Lara Putnam argues, "the digitized revolution is not inherently egalitarian, open, or cost-free" and digital corpora can threaten the "place-specific learning that historical research in a pre-digital world required" [Putnam 2016, 389, 377]. A nuanced discussion of these issues is not possible here and we do not attempt it, although navigating them is the central work of several of our team members, who have published on this elsewhere. [3] In this project we have endeavored to operate under principles of community-driven development and open collaboration, focussed on issues of access. This work too is still ongoing — we describe some next steps, including accessibility mechanisms in the technology for further community engagement, at the conclusion to this paper.

Data Collection and Collaborative Workflow

14 Collecting ground-truth data, or labeled examples from which the algorithm can learn, is always one of the biggest challenges when implementing a machine learning-based tool, but this rings especially true for specialized handwritten text transcription tasks. Historical documents often require significant domain expertise in ancient languages and scripts to read, meaning that the data collection process cannot simply be outsourced to human intelligence crowdsourcing services like Amazon Mechanical Turk. Furthermore, documents such as these, when transcribed by humanists, are typically not in a format that is suitable for use as ground-truth. When consolidating into an edition, edits are made that make the documents more accessible, but also make the result unsuitable for use as ground-truth for an algorithm, since it no longer represents a one-to-one (diplomatic) transcription of the original text.

15 One of the most important aspects of our collaboration was facilitating the collection of ground-truth data with which to train the algorithm. We experimented with two different workflows for this collection. In one method, co-author Eyob Derillo (SOAS, University of London; Asian & African Collections Reference Specialist at the British Library) transcribed samples of whole pages of text from two manuscripts from Gunda Gundē Monastery, viewable on vHMML's Reading Room interface as GG 00004 and GG 00016 (see Appendix B). These manuscripts were chosen essentially at random with respect to content, but with pages screened for clear and readable appearance.

16 Our second method, developed and largely executed by co-authors Gelila Tilahun (University of Toronto) and Sam Grieggs (Notre Dame), emerged from meetings of the University of Toronto and Notre Dame teams with the *Cannibal of Qəmər* project team led by co-author Steve Delamarter in collaboration with Wendy Belcher (Princeton University). The "Cannibal of Qəmər" project is a research collaboration between EMIP and Princeton University, based on the foundation of decades of imaging and cultural heritage preservation work performed by Delamarter and others, as detailed above. Workers in the Qəmər team including Delamarter, Jeremy Brown, Jonah Sandford, and Ashlee Benson had made diplomatic transcriptions of 95 manuscript witnesses of one of the best-known Tä²ammərä Maryam (the Miracles of Mary) stories of the Ethiopian Orthodox Church — the tale of "The Cannibal of Qəmər." The Qəmər team shared these transcriptions with the University of Toronto and Notre Dame collaborators as they did with Pietro Liuzzo of Hamburg University. With this contribution, our teams were able to make significantly faster progress towards an OCR reader, as Tilahun and Grieggs coordinated with Sindayo Robel (University of Toronto) and Enkenyelesh Bekele (volunteer) to lead the work to reformat the Qəmər team's transcription data into a format suitable for training.

17 Thus, the experimental data upon which we ultimately drew consist of images and transcriptions of lines of handwritten text from seventeen manuscripts written by scribes in Ethiopian church communities and monasteries between the sixteenth and twentieth centuries. As stated above, these manuscripts were originally imaged in Ethiopia with the consent and collaboration of their home institutions in a series of cultural heritage preservation projects between the 1960s and 2000s — see Appendix B for further details on these manuscripts and the availability of their digital surrogates.

18 The manuscripts are written on parchment folios, with text blocks divided into two or three columns depending on the manuscripts' size. The majority contain compilations of Tä²ammərä Maryam (the Miracles of Mary) stories of the Ethiopian Orthodox Church. We drew our data from copies of a single text from this collection, specifically, the tale of "The Cannibal of Qəmər." The full tale of "The Cannibal of Qəmər" is described in approximately twelve to eighteen columns of writing. The sequential reading of the story starts from the top of the leftmost column and continues down to the end of the column. The story then continues from the top of the closest right column. Below is an image of two pages/folio sides from our manuscript image repository.

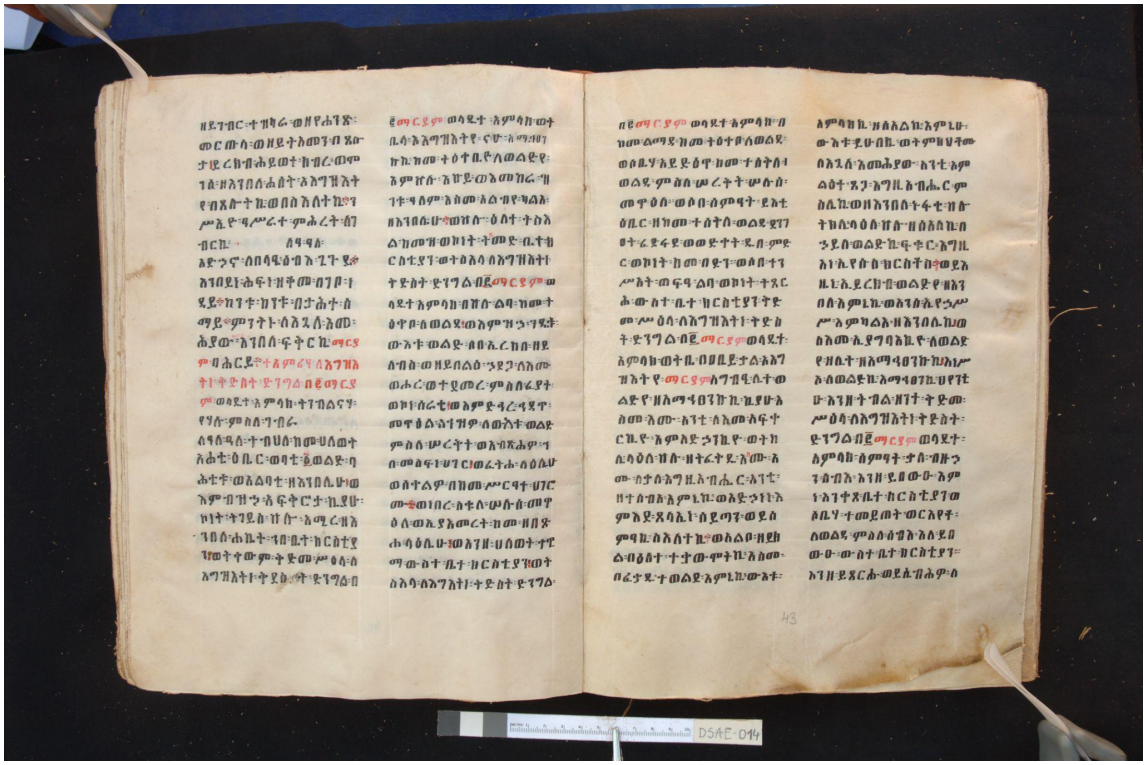

Figure 1. Ethiopia, Tegrāy Province, Dabra Śāhel Agwazā Monastery, HMML Pr. No. DSAE 00014, fols. 40v–41r, from the Cannibal of Qəmər. Image courtesy of the Dabra Śāhel Agwazā Monastery and the Hill Museum & Manuscript Library. Published with permission of the owners. All rights reserved. An example of Gəʿəz text used in our experiments.

Data Pre-Processing

For all of the manuscript samples, we have digital images of the handwritten texts and corresponding transcriptions in standard unicode. These transcriptions have been prepared by human experts and the accuracy of the transcriptions has been thoroughly verified. For each manuscript, we prepared two files: an image file and a text file. For each manuscript, starting from the column and line at which the Qəmər text begins, we sequentially worked our way down the column by manually cropping an image of each of the lines separately and copying them onto the image file. These cropped line images were sequentially numbered. In the text file, we copied the transcribed text that corresponded exactly to the cropped-and-copied line images. This way, the line number of a cropped line image corresponds exactly to the line number of a transcribed text. Since the goal of the project is to train a handwriting transcription algorithm, we note that if the scribe of a manuscript inadvertently misspelled or omitted a word, or left out a punctuation mark, the training transcription does not make a correction. 19

When pre-processing the data, there were a few things that were left out from consideration. For example, when we encounter rubricated words, such as a mention of “Mary” (**ማርያም**) or “Son of Christ” (**መልዱ ክርስቶስ**), the textual information is recorded in a homogenous color, and we thus do not capture the rubrication. Similarly, the rubricated parts of punctuations are also recorded in a homogenous color. In addition, pictorial illustrations, as well as their associated text bubbles, are omitted in the transcribed texts. Therefore, scribal information conveyed through the emphasis of rubricated words or pictorial illustration will not be available. Below is an excerpt of a pictorial reproduction of the “The Cannibal of Qəmər” tale. 20

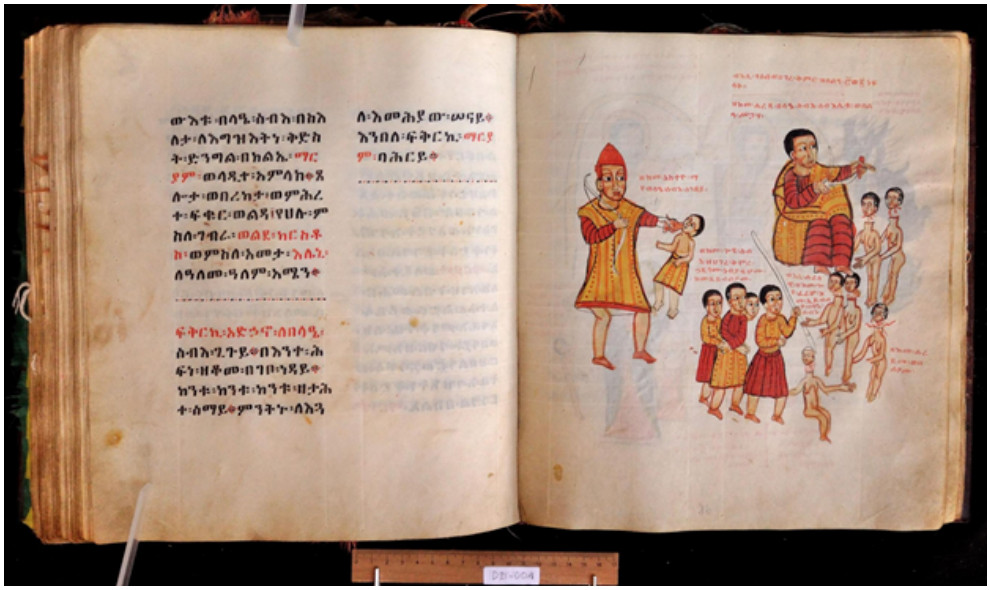

Figure 2. Ethiopia, Təgray Region, Däbrä Dammo ʔAbunä ʔArägawi, MS C3-IV-229 (digitized through the Ethio-SPaRe Project as EthioSPaRe DD-001), fols. 85v–86r. 1632–1664. Cataloged by Susanne Hummel. Accessed 2 Feb 2022 at https://mycms-vs03.rrz.uni-hamburg.de/domlib/receive/domlib_document_00001852. An example from the Cannibal of Qəmər showing a page with an illustration containing text.

Synthetic Ground-Truth Generation

One of the reasons that automatic document transcription is such a worthwhile task is that it is a very expensive process to transcribe documents manually — in both time and money. Unfortunately, this also means that the process of collecting ground-truth data is similarly resource-intensive. This is especially true for texts like Ethiopic manuscripts, which require specialized knowledge to read. Therefore we also make use of synthetic ground-truth data — i.e., images generated using transcriptions for which we don't possess a corresponding original image, for training purposes. Such data helps us account for variance in the documents that will ultimately be transcribed by the trained neural network model in the transcription tool. 21

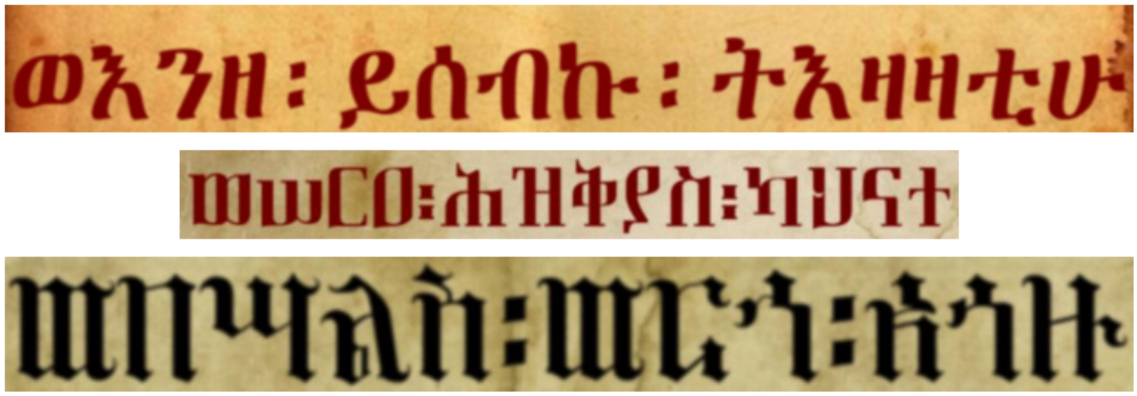

Figure 3. Examples of synthetic ground-truth. The images are created by taking transcribed text, splitting the lines up to match the length of the text lines in the target dataset, and then generating corresponding images for that text using different fonts, backgrounds, and colors to resemble possible configurations of the real documents.

21,873 lines transcribed from the legacy of the hagiographic Christian tradition in Ethiopia were used as the basis for generating the synthetic data. To better match the formatting of our real data, the text lines were assembled into 106,616 images, since the real data we were working with were in a 2-column format, with significantly fewer words per image than using the lines associated with the formatting in this separate transcribed text. Therefore, we added additional line breaks, such that the resulting images contained a similar distribution of the number of words per line as the dataset of real documents. 22

This data was provided by the Textual History of the Ethiopic Old Testament Project, and represents a gold standard for direct transcription data [Assefa et al. 2020]. As their goals were textual analysis, they went to great lengths to ensure the transcription recorded was directly what was written on the page, which increases its value as a representation of handwritten Gəʿəz. The synthetic ground-truth text was generated using the Text Recognition Data Generator, with special fonts added to account for the unicode characters used in Gəʿəz script. Examples are shown in Figure 3. 23

A Deep Learning-Based Transcription Algorithm for Gəʿəz

Convolutional Recurrent Neural Networks (CRNNs) are commonly used for the task of handwritten text recognition [Shi et al. 2016]. A CRNN is an artificial neural network architecture that is specifically designed for sequential text processing and meant to be trained over many different image samples of writing. The resulting trained network is commonly referred to as a model, which represents the core functionality of any transcription tool. CRNNs have recently been used in transcription tools that have achieved state-of-the-art results on the IAM and RIMES datasets of handwritten text [Xiao et al. 2020]. The English IAM and French RIMES datasets are commonly used by the computer vision and document processing communities to benchmark handwritten text recognition tasks. Despite the significantly expanded character set, the CRNN technique can be applied to Gəʿəz without a significant processing penalty. 24

CRNNs offer good performance, while not requiring excessive resources. In this case, we have found that it is reasonable to run a trained CRNN model at inference time without GPU acceleration, with CPU speeds of up to 4 lines of text per second on modern hardware from the past several years. However, even with the older hardware commonly found at under-resourced institutions, performance is still very good. For instance, on a 2013 Thinkpad T440p running Microsoft Windows we averaged 2.2 lines per second. While this is significantly slower than when processing on a GPU (usually more than 20 lines per second), it makes it possible to run this software to transcribe documents on just an old laptop. A GPU is still recommended to train models. 25

CRNNs work by passing the learned feature representations of a convolutional neural network into a series of recurrent layers that can analyze the text as a sequence. In essence, this means that the first convolutional layers of the network learn the parts of the image that are most important for identifying individual characters, encoding them into a sequence. This sequence is then analyzed by a series of recurrent layers that can understand context. This allows the model to understand, to some extent, which characters are likely to appear together, and to utilize that information when making a prediction about what an individual character should be. 26

While the specifics of the network can vary, we have drawn inspiration from the work of Joan Puigcerver [2017] and Shanyu Xiao et al. [2020], who use a base CRNN model with 5 convolutional layers which pass into 5 Bi-directional Long Short Term Memory (BLSTM) recurrent layers. We do not use any of the image rectification of Xiao et al.'s project. Additionally, we did not use dropout, a regularization technique that zeros out random parameters within the network to improve generalization [Srivastava et al. 2014], as extensively as they did. Only a pooling layer was dropped. Removing that layer appeared to marginally improve performance in our implementation. The exact architecture is shown in Figure 4. The code was written in Python using the PyTorch framework and is available on github here. 27

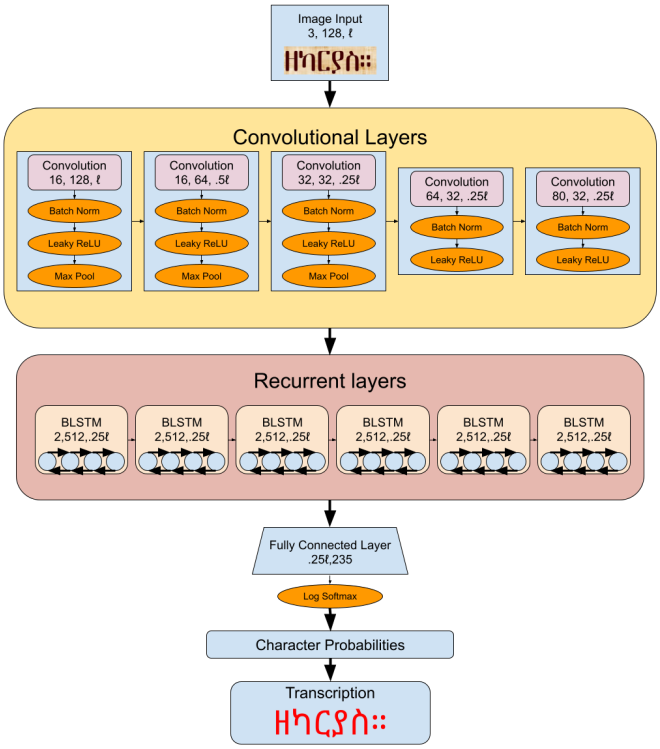

Figure 4. A diagram of the neural network architecture used for transcription. The image passes through 5 convolutional layers into 5 Bidirectional Long Short Term Memory (BLSTM) recurrent layers. The output size is specified at each layer, and as images of text lines are inherently of variable length, ℓ refers to the length of the input image.

The output of the network is not a string of text, but a matrix containing character probabilities for timesteps that roughly (but not exactly) correlate with a length of 4 pixels in the original image. For example, when the ground-truth is ስተ፡ፖርዓሁ፡ዘእንበለ, and we just take the highest probability character for each time step, the output looks something like this (the '~' character represents a blank space): 28

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ስስስ~~~~~~~~~~~~ተተተ~~~~~~~~~~~~~::
~~~~~~~~~~~~~~~ፖፖፖፖ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ርር~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ዓዓዓ~~~~~~~~~~~~~~~~~~ሁሁሁሁሁ~~~~~~~~~~~~~::
~~~~~~~~~~~~~~~~~~~~~~ዘዘ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~አአ~~~~~~~~~~~~~~~~~~~~~ንንን~~~~~~~~~~~~~~~~~~~በበበበ~~~~~~~~~~~~~~~~~~~~~~~~~ለለ~~~~~~~~~~
```

Example 1.

This turns out to be a perfect transcription, because if we merge all the repeating characters not separated by blank characters (which to the network represent the end of a character) we get a matching transcription candidate of ስተ፡ፖርዓሁ፡ዘእንበለ. This same example is shown graphically in Figure 5. 29

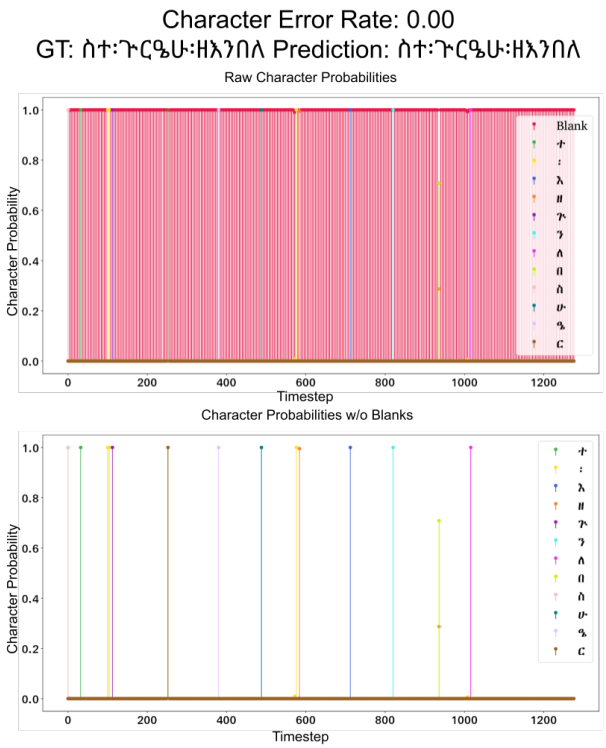

Figure 5. An example of the model output for one line of text. The first plot is a chart showing the probability of a specific character occurring at each time step, and the second plot is the same, but without the 'blank' character that specifies that a particular timestep doesn't correspond with a character.

A loss function is the measure of performance that a machine learning model is optimized against. It is a differentiable formula that can be used to calculate the model's accuracy. With this, we can measure the gradient of this function when we change the model's parameters, ideally changing them to move towards a more optimal result. 30

For this work, we use a Connectionist Temporal Classification (CTC) loss, the most commonly used loss function in handwritten text recognition [Graves 2018]. The CTC loss attempts to evaluate the network's prediction by taking all possible alignments of the ground-truth and the output matrix and finding the sum of the probabilities of the correct alignments. For a simple synthetic example, assume the network output shown in the top-most table of Figure 6. If the ground-truth transcription is "H," then the possible alignments are "HH," "~H," and "H~". This gives the probabilities shown in the center-left table, with the transcription receiving a CTC loss value of 0.76. The same process is illustrated for "h" and the blank character in the remaining two tables of the figure. Note that the blank character's only possible alignment is "~~," based on the information in the other two tables. The training objective is to maximize the probability that we have the correct transcription, a probability which is well represented by taking the CTC loss of the network prediction. 31

Character	Probability at Timestep	
Timestep	1	2
ሀ	0.1	0.75
ካ	0	0.15
~	0.9	0.1

Network output

Possible Solutions	Calculation	Result
ሀሀ	0.1*0.75	0.075
~ሀ	0.9*0.75	0.675
ሀ~	0.1*0.1	0.01
Total:		0.76

Probability that the Transcription is “ሀ”

Possible Solutions	Calculation	Result
ካካ	0*0.15	0
~ካ	0.9*0.15	0.135
ካ~	0*.1	0
Total:		0.135

Probability that the Transcription is “ካ”

Possible Solutions	Calculation	Result
~~	0.9*0.1	0.09
Total:		0.09

Probability that the Transcription is “Blank”

Figure 6. Example calculations for the CTC Loss Function. The top table represents the raw network output, i.e., the probability that a character is present at each timestep. The lower tables show the calculation of the CTC loss for each possible character transcription.

Experiments

To evaluate the proposed transcription tool, a series of controlled experiments were performed over relevant Gə'əz texts drawn from the datasets that were prepared for the project. These experiments are described below. 32

Data Augmentation For Training Dataset

When performing character frequency analysis, we found that some of the characters in our dataset appeared very infrequently. Not only does this negatively affect the model’s ability to identify these characters, but it also limits our ability to assess the transcription performance on individual characters. To address this class imbalance, we generated new synthetic lines of text for the training and evaluation (validation and test) datasets. These lines were created by collecting all the individual words from both sets of texts and compiling a list of all the words containing each character. We then randomly selected words from this list such that each character appears a specified number of times in the training and evaluation sets. 33

Since we wanted to make sure that we did not overemphasize synthetic examples during training, we generated synthetic data for the training set until each character appeared 10 times. We did not add any synthetic data to the validation set used for model tuning during training, since the model is trained until performance plateaus on it, and we wanted to emphasize performance on images drawn from real manuscript pages. For the testing dataset, we added data until there were 100 examples of each character. While the model performs better on synthetic data because there is less variance (due to the limited number of fonts), it is trivial to evaluate the model both with and without the added synthetic examples. Thus with this methodology, we could get a more accurate depiction of performance on individual characters, without diluting performance on real manuscripts. Additionally, all images were augmented on-the-fly using the random elastic distortion technique of Wigington et al. [Wigington et al. 2017]. On-the-fly augmentation ensures that each image shown to the model is unique, and greatly reduces the opportunity for the model to simply memorize each image, or “overfit.” 34

CRNN Training

For a rigorous evaluation, we trained three CRNN models in two stages using different random initializations of the parameters that are adjusted as the model learns. First, each model was trained on a dataset made completely from synthetically generated text. These models are used as a starting point for the three models that will be used for the transcription task; each is then fine-tuned on the combination of real and synthetic data mentioned in the paragraph above. Specifically, these models were trained on 106,616 line images of Gə'əz text generated from transcriptions provided by the Textual History of the Ethiopic Old Testament Project [Assefa et al. 2020], with an additional 9,593 images created using disjoint text used as a validation set only during the initial training phase. This selection of text contained 234 different characters. We used an RMSprop optimizer with a learning rate of 3*10^-4, and each model was trained from a random initialization until it did not improve for 80 epochs. 35

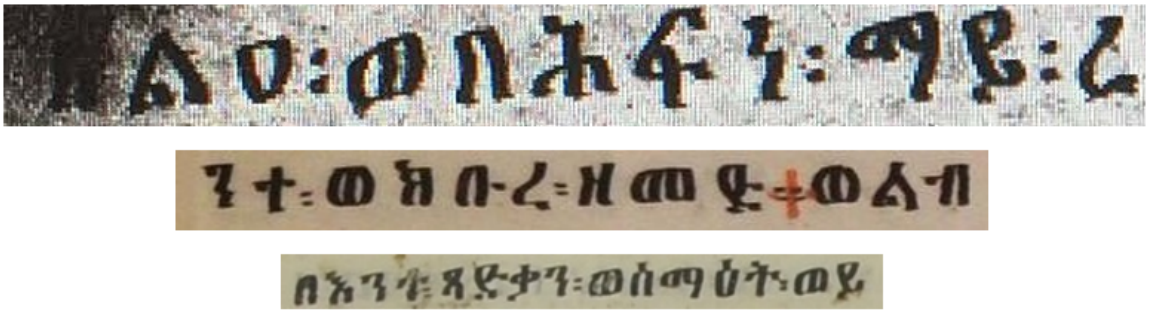

Figure 7. Examples of the various types of manuscript images in the evaluation datasets.

We then proceeded to repeat this training process on our dataset of real Gəʻəz manuscript images augmented with a selection of synthetic examples of rare characters using the three candidate models trained on the synthetic data as the starting parameters. This process is commonly known as fine-tuning, and is helpful to overcome domain mismatch between real and synthetic data. For the dataset consisting of real manuscript images, we have a collection of 1,510 lines from various manuscripts, including both high and low resolution color images, and also lower quality images scanned from microfilm media in black and white. Examples of these images are pictured above in Figure 7. 224 images were withheld as a validation set for this additional training run, which was used to evaluate the model after each epoch. Finally, 394 line images were held out as a test set, which were only used for a final evaluation of a model after all training was completed. This procedure was designed to ensure that the model is not overfit to the data that it has seen in training, and can generalize to unseen data as well. 36

CRNN Evaluation

The metric commonly used to measure transcription performance, Character Error Rate (CER) is calculated by counting the number of edits required to make the predicted string match the ground-truth string, and dividing that number by the length of the string. An example of this in English would be if a model outputs “Helo,” when the ground-truth is “Hello,” which would require a single edit, adding an “l” to correct it, making its “edit distance” 1. This is known as a “Deletion Error.” Similarly, if the model outputs “Helllo,” we reach the ground-truth string by removing a single “l,” which also gives us an edit distance of 1. This is an “Insertion Error”. Finally, if the model outputs “Helfo,” we can achieve the target string by changing the “f” into an “l.” This also has an edit distance of 1, and is considered a “Substitution Error.” 37

To further emphasize what this means, some example mistranscriptions of the English string “The Quick Brown Fox Jumps Over The Lazy Dog” and their respective CERs are shown in Table 1. 38

Proposed Transcription	CER
“The Quick Brown Fox Jumps Over The Lazy Dog”	0%
“The Quick Brown Fox Jumps Over The Lazy”	9.3%
“Th Qick Brwn Fx Jmps vr Th Lzy Dg”	23.2%
“the quick brown fox jumps over the lazy dog”	20.9%
“th aick dlak fao bump the dog”	60.4%

Table 1. Examples of Character Error Rates (CER) of varying degrees. Note that some types of errors are much more legible than others.

The results listed in Table 2 below as “Test” are a reflection of performance on data completely unseen during training time. These data are only used after choosing a model based on the results on the validation set, and help demonstrate model generalizability to unseen data. As a reminder, in order to evaluate performance across all characters, we balanced the test set using synthetic augmentations, so that each character appears at least 100 times. Since typeset characters are easier for our model, we included the results with and without these augmentations. While the results are not perfect, the average CER on the Test set is quite low, meaning the models are able to reliably recognize individual characters. 39

	Validation CER (%)	Test CER (%)	Test w/o Augmentations CER (%)
Average	8.25%	6.35%	10.62%
Error	±0.31%	±0.88%	±0.32%
Best	7.71%	4.60%	10.08%

Table 2. Character Error Rate (CER) of our model on the Validation and Test sets of the data we collected. The performance is aggregated across 3 models. The “Best” row refers to the model we trained that had the best performance, which is the one that would be selected for further use.

Analysis of Transcription Mistakes

While CER is a useful tool to measure performance across transcription models, it is somewhat lacking in assessing the readability of model output. A key question is whether or not the output will be useful to the reader. The human brain does not just sequentially classify each character individually when it reads, so some incorrect transcriptions are more problematic than others, since they make the transcription less legible. We can further analyze a model’s performance by looking at exactly what kinds of mistakes it is making. For this we consider the output of the best performing model. While there can be multiple shortest paths to edit a line of text into a specified target, we systematically generated sets of the minimum edits required to map the predicted strings onto the ground-truth text, and found the characters most commonly missed. 40

Considering the most commonly missed characters we find an interesting trend. The cutoff for the 90th percentile of most missed characters is a CER of 14.8% in aggregate. Table 3 shows the characters that make up the 90th percentile of most frequently misidentified characters. With the exception of two punctuation characters, we see that the vast majority of errors are substitution errors, and for many, greater than 30% of the total errors comes from one specific character 41

substitution.

Ground-Truth Character	MCS	Freq. of MCS	MCS % of Total Errors	Other Common Subs.	Total Sub. Errors	Total % of Sub. Errors	Total Errors
ስ	ሰ	39	29.32%	ዕ,ሱ,ለ,ሶ,እ,ፀ,ል,ዌ,አ	123	92.48%	133
የ	ያ	21	17.80%	፪,ም,ዖ,ዋ,ፆ,ዩ,ዬ,ዮ,ት	113	95.76%	118
ዘ	ወ	39	17.89%	አ,ዝ,በ,ለ,ዞ,እ,ዚ,ኢ,፱	209	95.87%	218
ሰ	ስ	55	55.56%	ዕ,ሱ,ለ,ሶ,ጎ,ኀ,ይ,ገ,ሕ	93	93.94%	99
ቤ	ቢ	5	31.25%	ሊ,ጤ,ር,ጌ,ኬ,ሲ	13	81.25%	16
ዓ	ፀ	8	19.05%	ዲ,ት,ዔ,ዒ,ኖ,ባ,ጓ,ጎ,ን	41	97.62%	42
ገ	ን	9	16.98%	ጉ,ግ,ጓ,ፒ,ፐ,ፓ,ጌ,ጔ,ፔ	48	90.57%	53
ዳ	ዷ	4	13.79%	ዱ,ድ,ዴ,ጻ,ጼ,ት,ላ,ኖ,ፉ	28	96.55%	29
ጸ	ጻ	18	23.38%	ጾ,ጽ,ጼ,ጳ,ጹ,፳,፲,ጺ,፲	74	96.10%	77
ሀ	ሠ	4	14.81%	ህ,ሁ,ዌ,ቢ,፱,ኔ,ይ,ዕ,ዐ	23	85.19%	27
፪	ቋ	4	14.29%	፸,፶,፮,፫,፱,ዌ,ዩ,፪,ዼ	23	82.14%	28
፮	፳	21	61.76%	፯,፶,፴,፲,፺,፸,፹,፫	31	91.18%	34
ጠ	ጡ	7	36.84%	ጣ,ጤ,፲,ጢ	17	89.47%	19
ቊ	ቍ	34	85.00%	ቒ,ኈ	38	95.00%	40
፯	፺	11	44.00%	፲,፪,፫,፮,፻,፫	22	88.00%	25
:	፡	1	4.76%	n/a	1	4.76%	21
፶	፵	10	41.67%	፫,፱,፲,፫	21	87.50%	24
፬	፴	9	25.71%	፹,፮,፵,፱,፲,፯,፪,፸	32	91.43%	35
ኍ	ኍ	19	100%	n/a	19	100%	19
.	:	1	1.85%	n/a	1	1.85%	54
፵	፶	5	50.00%	፴	6	60.00%	10
፫	፲	7	18.42%	፲,፲,፶,፳,፯,፫,፺,፵	35	92.11%	38
ኍ	ኍ	28	84.85%	ኁ,ኈ	31	93.94%	33
ፒ	ፕ	27	65.85%	ፔ,ጊ	37	90.24%	41

Table 3. The most commonly missed characters, and the characters most commonly substituted (MCS) with them using our highest performing model.

Looking at the situations where characters are missed, we see some encouraging signs. We found a large portion of the model's errors were substitution errors amongst characters that are visually similar. This tends to be one of the more legible errors, and means in aggregate that it should be easier to understand the resulting transcription. Furthermore, of the commonly missed characters that are not frequently substituted with one visually similar character, most are punctuation. Another trend we see in the data is that the model performs poorly on numbers. This is almost certainly due to the fact that our real training data was a collection of biblical stories, so these characters did not occur frequently. Digits tend to be an illegible error, since they would be difficult to correct without looking at the source. But looking at the commonly missed characters, we see that when the model misses digits it is mostly confusing them with other digits (see Appendix A for a listing of the numerical characters). 42

This means that in the worst case, our model will produce transcriptions that will be usable right away, with any errors being easily correctable by the human readers. The accuracy of our tool's reading of Gəʿəz might, however, be improved in future. In state-of-the-art transcription tools for modern handwritten text, decoders based on language models are often used to improve transcription quality over the naïve "highest probability character at each time step"' approach. Implementing this, however, creates a significant challenge when historical documents are the source texts: it is notoriously difficult to model their language, because few diplomatic transcriptions of such documents exist, and producing them specifically for the purpose of OCR is laborious. While the language models that are used for decoding are typically very simple n-gram probability-based models, which do not require data on the scale of state-of-the-art transformer-based language models, finding good representations of the target data can be difficult, since most of this type of data is not standardized across time and space. We could address some of these challenges by working collaboratively with our team on more specific language models. In addition, some of the characteristics of the Gəʿəz language could be exploited in the structure of the CRNN. The character family information could potentially be incorporated into the loss function to allow for transcriptions that are more readable than the ones that are currently being produced, if not perfect. 43

Comparison to Transkribus

Transkribus [Kahle et al. 2017] is another tool that can be used not only to automatically transcribe handwritten text, but also to collect ground-truth data for it. The Transkribus project has an Ethiopic text model available; thus we have decided to include a brief comparison to the results from that tool. 44

It can be somewhat difficult to compare transcription performance across algorithms and datasets, for a number of reasons, including the difficulty of the evaluation dataset, the differences in training data, and differences in string tokenization. String tokenization refers to the "rules" on how to handle possible edge cases when making ground-truth data. One example of this is that our data do not have spaces between words, only the Gəʿəz word separator, "፡," while the Transkribus model's data includes a space "፡", which would be counted as an error with a direct comparison to our tool. Because of this, we manually retokenized the model output to match our data. For the purposes of this comparison we share Transkribus's reported error rate. However, since the dataset the Transkribus model is trained and validated on is different from the dataset we put together, and that data is not publicly available, we also provide results on the test set for our data. This may not be a fair comparison, but it allows us to give better context for the results. 45

In order to give Transkribus the best chance on our data, we uploaded each line image of the unaugmented test set to the project's server and manually drew text region boxes around the whole text, lining boxes tightly around the text for each image. This is not quite the intended use case for Transkribus, since the platform supports whole pages, and has a number of automated segmentation tools, but it ensured a more direct comparison, with both tools seeing the same images. 46

Additionally, some of our images contain cropping artifacts, that is, some black space around the edge of the images. These tended to break Transkribus's line segmentation, which is unsurprising, because it would be out of scope for what that tool has been designed to work with. The manual bounding boxes were intended to help give it the best chance at dealing with these as well.

	Character Error Rate	
Transkribus on that project's Validation Set	5.16%	
Transkribus on this paper's Test Set	27.6%	
Proposed tool on this paper's Test Set	10.6%	±0.32%

Table 4. Comparison in performance on our Test Set using only real manuscript images.

From the results in the above table, we can see that Transkribus performs better on that project's own data, but doesn't generalize to our data with a similar level of performance. Note that our result is the average performance over three trained neural networks, but Transkribus only has one model for Gə'əz. By diving further into these results we see a pattern where the Transkribus model has trouble with cropping artifacts in some of our images where the lines were cropped using a Lasso selection tool to prevent overlapping text from being included in the image. The image itself is still a rectangle, so the areas outside of the selection are filled with a solid black color. This accounts for a great deal of the Transkribus model's underperformance on our data because our model sees this at training time and knows how to handle it, while Transkribus does not. 47

We also found that the pattern of character-level errors was different between our model and Transkribus. On the unaugmented test set, we found that Transkribus had 166 replacement errors, 1034 deletion errors, and 18 insertion errors. Our model had 305 replacement errors, 22 deletion errors, and 152 insertion errors. A large portion of the deletion errors from Transkribus were due to an inability to process cropped microfilm images, where the black borders interfered with the transcription process. 48

The most important distinction between the two approaches is price. Our software is open-source and freely available to anyone who wants to use it, while Transkribus is a service that requires tokens. While they give a generous free allocation of tokens, because the transcription is run on the project's own servers, there are maintenance costs associated with the upkeep. Thus additional tokens must be purchased for large tasks. 49

While we would recommend a GPU to train a model from scratch, our software can be run on just a laptop to transcribe the text, so it offers additional portability. Keeping the images local is also of some value. Situations often occur where uploading the images to an external server may cause intellectual property concerns. This is an issue we have encountered frequently when working with digitized manuscripts. 50

The Transkribus platform offers more sophisticated layout analysis software than our tool provides, and contains a variety of pretrained models for multiple languages available out of the box. Thus there are definitely use cases where it could be a more appealing option. For instance, the narrow question of OCR accuracy is not the only point of comparison that represents factors crucially important to the needs of scholars. The functionality of joint access to image and transcription provided by Transkribus, as well as its interface that connects (highlights) the location in the images with the location in the transcription, improves ease of use and the accuracy of the workflow. However, as our tool is completely free and open-source, it might be a more attractive option for people with internal tooling, since all of the source code is available to be integrated into a custom pipeline. 51

Web Client and User Interface

In order to showcase the functionality of our transcription tool, we designed and implemented a publicly accessible web demo which anyone can use to transcribe images of Gə'əz text. We prioritized making this website simple and user-friendly in order for scholars without extensive experience in computer science to use it. Software that utilizes neural networks is often seen as intimidating and prohibitively resource-intensive: we wanted to make sure that ours was accessible to end-users with more limited resources, and we kept in mind that many of these manuscripts are located in remote areas of Ethiopia, where it may be difficult to acquire and run large workstations with expensive GPUs. 52

The neural network that directly transcribes text was originally coded in Python, so we decided to program the website through Flask, a web framework that Python provides, in combination with HyperText Markup Language (HTML). Through Flask's web framework, we were able to code input and output pages that allow for the presentation of clear results for the transcription of an input text, as well as efficient methods of inputting an image into the website. Specifically, we coded the functions of the opening and resulting pages in combination with HTML templates that arrange and outline each page. Using Flask's Dropzone package, it was possible to support dragging an image into one of the site's pages as a method of inputting it into the tool. Though we programmed a Dropzone area for a user to drag in an image of a piece of text, we also coded in an option for the user to upload an image to cater towards different circumstances. Once an image is successfully uploaded, the resulting page will display the uploaded image and the resulting transcription in a formatted table. The original image and the resulting transcription will be adjacent to one another in the output. 53

The website tool currently only accepts images of lines of text. We are currently working on a graphical user interface (GUI) feature that permits a user to input an entire page of text, similar to what our standalone tool supports (described below). The user would input the text in the same fashion. Utilizing contour analysis and thresholding functions within Python, the GUI feature permits the user to clearly distinguish text and eliminate noise within the image by adjusting different parameters on trackbars. Python's OpenCV library includes trackbar functions that can be coded in the GUI feature. With these additional features, transcription accuracy and accessibility of the demo will be greatly enhanced. 54

This web interface is written in such a way that it can interface with a PyTorch backend. The server running this tool has no GPU, which demonstrates our software's flexibility. The tool can be accessed at http://docturk.crc.nd.edu/ and the code used to run it can be found at https://github.com/grieggs/Ge-ez-HWR. 55

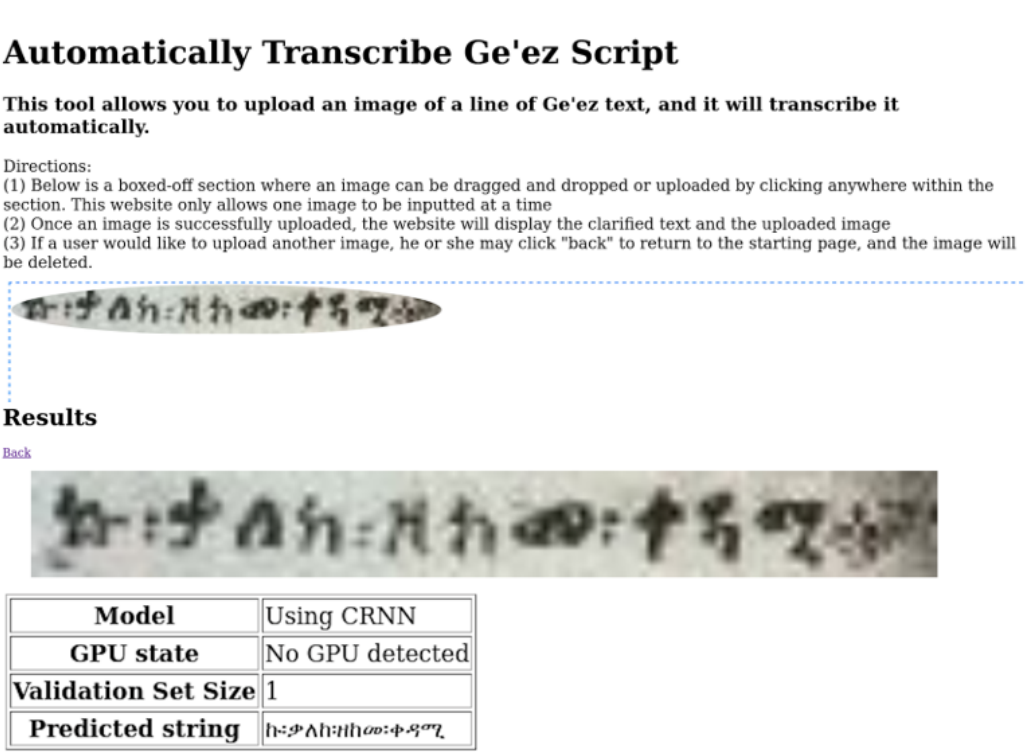

Figure 8. Screenshots of the transcription site. A user can just drag a line of text into the upload box (top), and the tool will transcribe it automatically (bottom).

Additionally, we have a fully offline version that can be run locally on a user's computer. It can process segmented line images in batches of any size, transcribing entire directories of images all at once. This tool can be found at https://github.com/grieggs/Ge-ez-HWR, and instructions on how to run it are included in the repository. The offline tool also includes some simple software for segmenting lines semi-automatically. The offline version offers users the ability to train a new model from scratch, or to fine-tune an existing model on a specific dataset of interest. This feature works generically on handwritten text regardless of language, but we recommend using a CUDA capable GPU for training. All of this software is fully open-source and available under an MIT License. Thus anyone interested in using this software is welcome to branch these repositories and modify the code as needed without any restrictions.

Conclusion

The past decade has witnessed significant growth in the study and conservation of Ethiopian manuscripts and cultural history [Nosnitsin 2020, 282], including the large scale project Beta maṣāḥəft: Manuscripts of Ethiopia and Eritrea at the Hiob Ludolf Centre for Ethiopian Studies at the University of Hamburg (2016–2040, PI Alessandro Bausi). Yet compared with Western heritage materials, such as European manuscripts, Gə'əz manuscripts need more and better tools to aid their study and preservation. Gezae Haile argues there is "a greater need than ever for systematic recording/cataloging, conservation and digitization of Ethiopian manuscripts in Ethiopia" [Haile 2018, 41], and Liuzzo meanwhile notes a "major shift in focus" in the field of digital Ethiopian studies, "to the aim of creating resources available online" [Liuzzo 2019, xxiv].

We hope that our OCR tool will make a significant contribution to the broader project of resourcing digital Ethiopian studies. We also hope it will have significance for research into other historical manuscript cultures. Our tool fits an important niche in the automated document transcription space: a lightweight, offline, and open-source algorithm, with a high degree of accuracy, that both simplifies the process of building a transcription tool from scratch, and allows for keeping the data in-house. It can run at reasonable speeds without a GPU, meaning it will work on most modern laptops, even in contexts where users have poor internet access.

While our focus has been Gə'əz manuscripts, the tool as designed could also be used for other languages and scripts. Reworking the tool itself for a new language and script simply requires new ground-truth — the more the better, but a minimum of 1,000 lines — plus a GPU and computer scientist for the training process. We've included a case study for retraining on modern English text in our github repo, using the same steps that can be generalized to other languages. Our team recently met with a group of Sanskrit scholars led by Ajay Rao at the University of Toronto Mississauga, to discuss OCR — and the potential reworking of our tool — for transcription of Sanskrit manuscripts. Now that the core of this tool exists and is open-source, it is available for a new phase of development along these lines: scholars working in other fields on digitized manuscripts written in under-resourced scripts can tailor and frame the technology to suit their particular needs.

Acknowledgments

Toronto members of the project team received funding through The Andrew W. Mellon Foundation (2019–2021); and the University of Toronto Mississauga Research and Scholarly Activity Fund (RSAF) 2019.

HMML images are courtesy of the Dabra Śāhel Agwazā Monastery, Tegrāy Province; Qundi Giyorgis Church, Šawā Province; Ḥayq Esṭifānos Monastery, Wallo Province; Darafo Māryām Church, Šawā Province; Gunda Gundē Monastery, Tegrāy Province, Ethiopia, and the Hill Museum & Manuscript Library. Published with permission of the owners. All rights reserved.

EAP432 images were accessed courtesy of the Endangered Archives Programme (EAP), Project Lead Dr. Mersha Alehegne Mengistie.

All EMIP images were accessed courtesy of Ethiopic Manuscript Imaging Project (EMIP), Steve Delamarter, Director.

Ethio-SPaRe images were accessed courtesy of the project Ethio-SPaRe: Cultural Heritage of Christian Ethiopia – Salvation, Preservation, and Research, EU 7th Research Framework Programme, ERC Starting Grant 240720, PI Denis Nosnitsin, 2009–2015, https://www.aai.uni-hamburg.de/en/ethiostudies/research/ethiospare.html (the manuscripts now can be accessed also at https://betamasaheft.eu/). 64

For searchability, place names and other manuscript details have been listed with the transcription systems and regional identifiers used on these respective digital platforms. However, readers should note that Ethiopia's regional administrative system has changed since many of these manuscripts were imaged. "Province" refers to a regional unit in the administrative divisional structure prior to 1995, and "region" to the current one. 65

The transliteration tables in Appendix A were reproduced with kind permission of Alessandro Bausi, and originally made for Kelly, S. (ed.), *A companion to medieval Ethiopia and Eritrea*, Leiden: Brill. Freely accessible at https://brill.com/view/book/9789004419582/front-11.xml 66

The authors wish to thank the editors and blind reviewers of *Digital Humanities Quarterly*, and Denis Nosnitsin, for their incisive and generous feedback on drafts of this article. 67

Appendix A: Transliteration Tables for Gə'əz Script

1st order	2nd order	3rd order	4th order	5th order	6th order	7th order
ሀ ha	ሁ hu	ሂ hi	ሃ ha	ሄ he	ህ hə	ሆ ho
ለ lä	ሉ lu	ሊ li	ላ la	ሌ le	ል lə	ሎ lo
ሐ ḥa	ሑ ḥu	ሒ ḥi	ሓ ḥa	ሔ ḥe	ሕ ḥə	ሖ ḥo
መ mä	ሙ mu	ሚ mi	ማ ma	ሜ me	ም mə	ሞ mo
ሠ śä	ሡ śu	ሢ śi	ሣ śa	ሤ śe	ሥ śə	ሦ śo
ረ rä	ሩ ru	ሪ ri	ራ ra	ሬ re	ር rə	ሮ ro
ሰ sä	ሱ su	ሲ si	ሳ sa	ሴ se	ስ sə	ሶ so
ሸ šä	ሹ šu	ሺ ši	ሻ ša	ሼ še	ሽ šə	ሾ šo
ቀ qä	ቁ qu	ቂ qi	ቃ qa	ቄ qe	ቅ qə	ቆ qo
ቐ ǧä	ቑ ǧu	ቒ ǧi	ቓ ǧa	ቔ ǧe	ቕ ǧə	ቖ ǧo
በ bä	ቡ bu	ቢ bi	ባ ba	ቤ be	ብ bə	ቦ bo
ተ tä	ቱ tu	ቲ ti	ታ ta	ቴ te	ት tə	ቶ to
ቸ čä	ቹču	ቺ či	ቻ ča	ቼ če	ች čə	ቾ čo
ኀ ḫa	ኁ ḫu	ኂ ḫi	ኃ ḫa	ኄ ḫe	ኅ ḫə	ኆ ḫo
ነ nä	ኑ nu	ኒ ni	ና na	ኔ ne	ን nə	ኖ no
ኘ ñä	ኙ ñu	ኚ ñi	ኛ ña	ኜ ñe	ኝ ñə	ኞ ño
አ ʾa	ኡ ʾu	ኢ ʾi	ኣ ʾa	ኤ ʾe	እ ʾə	ኦ ʾo
ከ kä	ኩ ku	ኪ ki	ካ ka	ኬ ke	ክ kə	ኮ ko
ኸ ḳä	ኹ ḳu	ኺ ḳi	ኻ ḳa	ኼ ḳe	ኽ ḳə	ኾ ḳo
ወ wä	ዉ wu	ዊ wi	ዋ wa	ዌ we	ው wə	ዎ wo
ዐ ʿa	ዑ ʿu	ዒ ʿi	ዓ ʿa	ዔ ʿe	ዕ ʿə	ዖ ʿo
ዘ zä	ዙ zu	ዚ zi	ዛ za	ዜ ze	ዝ zə	ዞ zo
ዠ žä	ዡ žu	ዢ ži	ዣ ža	ዤ že	ዥ žə	ዦ žo
የ yä	ዩ yu	ዪ yi	ያ ya	ዬ ye	ይ yə	ዮ yo
ደ dä	ዱ du	ዲ di	ዳ da	ዴ de	ድ də	ዶ do
ጀ ǧä	ጁ ǧu	ጂ ǧi	ጃ ǧa	ጄ ǧe	ጅ ǧə	ጆ ǧo
ገ gä	ጉ gu	ጊ gi	ጋ ga	ጌ ge	ግ gə	ጎ go
ጠ ṭä	ጡ ṭu	ጢ ṭi	ጣ ṭa	ጤ ṭe	ጥ ṭə	ጦ ṭo
ጨ č̣ä	ጩ č̣u	ጪ č̣i	ጫ č̣a	ጬ č̣e	ጭ č̣ə	ጮ č̣o
ጰ p̣ä	ጱ p̣u	ጲ p̣i	ጳ p̣a	ጴ p̣e	ጵ p̣ə	ጶ p̣o
ጸ ṣä	ጹ ṣu	ጺ ṣi	ጻ ṣa	ጼ ṣe	ጽ ṣə	ጾ ṣo
ፀ ṣ́ä	ፁ ṣ́u	ፂ ṣ́i	ፃ ṣ́a	ፄ ṣ́e	ፅ ṣ́ə	ፆ ṣ́o
ፈ fä	ፉ fu	ፊ fi	ፋ fa	ፌ fe	ፍ fə	ፎ fo
ፐ pä	ፑ pu	ፒ pi	ፓ pa	ፔ pe	ፕ pə	ፖ po
ቨ vä	ቩ vu	ቪ vi	ቫ va	ቬ ve	ቭ və	ቮ vo

Table 5. *Fidäl Table/Ethiopian Syllabary* (transcription based on the system adopted by the *Encyclopaedia Aethiopica*). Reproduced with permission of Alessandro Bausi.

ቈ qʷä		ቍ qʷi	ቈ qʷa	ቈ qʷe	ቍ qʷə	
ኍ ẖʷä		ኍ ẖʷi	ኋ ẖʷa	ኋ ẖʷe	ኍ ẖʷə	
ኰ kʷä		ኰ kʷi	ኳ kʷa	ኳ kʷe	ኰ kʷə	
ጐ gʷä		ጐ gʷi	ጓ gʷa	ጓ gʷe	ጐ gʷə	
ኸ x̱ʷä		ኸ x̱ʷi	ኻ x̱ʷa	ኻ x̱ʷe	ኸ x̱ʷə	
ቌ qʷä		ቌ qʷi	ቋ qʷa	ቋ qʷe	ቌ qʷə	

Table 6. *Labiovelars* (transcription based on the system adopted by the *Encyclopaedia Aethiopica*). Reproduced with permission of Alessandro Bausi.

፩ 1	፰ 8	፸ 60
፪ 2	፱ 9	፹ 70
፫ 3	፲ 10	፺ 80
፬ 4	፳ 20	፻ 90
፭ 5	፴ 30	፼ 100
፮ 6	፵ 40	፲፼ 1,000
፯ 7	፶ 50	፼፼ 10,000

Table 7. *Numerals* (transcription based on the system adopted by the *Encyclopaedia Aethiopica*). Reproduced with permission of Alessandro Bausi.

Appendix B: Manuscript Images Consulted for Ground-Truth Data

Images accessed through the Endangered Archives Programme (EAP):

EAP432/1/4 a Ethiopia, East Gojjam, Debre Koreb We Qeraneyo Medhanealem Monastery, MS digitized by Hamburg University as EAP432/1/4, fols. 181v-183r. *Täʾamrä Maryam "Miracles of Mary"*, 17th century.

EAP432/1/4 b "Te'ammere Maryam. [17th century]", British Library, EAP432/1/4, https://eap.bl.uk/archive-file/EAP432-1-4, Images 181-182. Digitized through the Endangered Archives Programme supported by Arcadia. Accessed 21 June 2023.

Images accessed through the Ethiopic Manuscript Imaging Project (EMIP):

EMIP 0761 Ethiopia, Tegrāy Province, Selassie Chelekot Church, MS 20 digitized as EMIP0761, fols. 227v–229r. *Täʾamrä Maryam "Miracles of Mary"*, 17th century.

EMIP 0832 Ethiopia, Tegrāy Province, Selassie Chelekot Church, MS 93 digitized as EMIP0832, fols. 98r–98v. "A major collection of 318 *Täʾamrä Maryam 'Miracles of Mary'*", 1750–1849.

EMIP 01154 Ethiopia, Addis Alem, Adbarat Debretsion Mariam Church, MS Addis Alem 112 digitized as EMIP 01154, fols. 30v–31v. *Täʾamrä Maryam "Miracles of Mary"*, 1868–1913.

MS fols. 73r–76v Ethiopia, East Gojjam (?), Dabra Wark (?), Debre Werk Saint Mary Church (?), MS fols. 73r–76v. Contents include *Täʾamrä Maryam "Miracles of Mary"*, 17th century. Microfilmed images held at the Institute of Ethiopian Studies (Addis Ababa), digitized in 2010 at the request of the director of the Manuscripts Department. Microfilm among a collection attributed to Donald Davies. Images courtesy of Ethiopic Manuscript Imaging Project (EMIP), Steve Delamarter, Director.

Images accessed through the Ethio-SPaRe Project:

DD-001 Ethiopia, Tegrāy Region, Däbrä Dammo ʾAbunä ʾArägawi, MS C3-IV-229 digitized through the Ethio-SPaRe project as DD-001, fols. 46v-48r. *Täʾamrä Maryam "Miracles of Mary"*, 1632-1664. Cataloged by Susanne Hummel, description accessed 2 Feb 2022.

DZ-003 Ethiopia, Təgray Region, Däbrä Zäyt Qəddəst Maryam, MS digitized through the Ethio-SPaRe project as DZ-003, fols. 46v-48r, 63r, and 63v. *Haymanotä ʾabäw "Faith of the Fathers"* / *Täʾamrä Maryam "Miracles of Mary"*, 1550–1650. Cataloged by Stéphane Ancel, description accessed 2 Feb 2022.

DD-010 Ethiopia, Tegrāy Region, Däbrä Dammo ʾAbunä ʾArägawi, MS C3-IV-258 digitized through the Ethio-SPaRe project as DD-010, fols. 56r-68v. *Täʾamrä Maryam "Miracles of Mary"* / *Täʾamrä Gäbrä Mänfäs Qəddus "Miracles of Gäbrä Mänfäs Qəddus"*, 1772. Cataloged by Susanne Hummel, description accessed 2 Feb 2022.

GBI-011 Ethiopia, Tegrāy Region, Gwaḥgot ʾIyäsus, MS digitized through the Ethio-SPaRe project as GBI-011, fols. 102r–104v. *Täʾamrä Maryam "Miracles of Mary"*, 17th century. Cataloged by Susanne Hummel, description accessed 2 Feb 2022.

NSM-007 Ethiopia, Tegrāy Region, Nəḥbi Qəddus Mikaʾel, MS digitized through the Ethio-SPaRe project as NSM-007, fols. 90v–92r. *Täʾamrä Maryam "Miracles of Mary"*, 1900–1950. Cataloged by Susanne Hummel, description accessed 2 Feb 2022.

QSM-017 Ethiopia, Tegrāy Region, Qärsäbär Qəddus Mikaʾel, MS digitized through the Ethio-SPaRe project as QSM-017, fols. 48r–50v. *Täʾamrä Maryam "Miracles of Mary"*, 1721–1735. Cataloged by Stéphane Ancel, description accessed 2 Feb 2022.

Images accessed through the Hill Museum & Manuscript Library (HMML):

DSAE 00014 Ethiopia, Tegrāy Province, Dabra Śāhel Agwazā Monastery, MS digitized as HMML Pr. No. DSAE 00014, fols. 40v–42v. *Täʾamrä Maryam "Miracles of Mary"* 20th century(?). Digitized by Ewa Balicka-Witakowska and Michael Gervers. Metadata supplied by Ted Erho. Accessed 3 Feb 2022.

EMML 1978 Ethiopia, Šawā Province, Qundi Giyorgis Church, MS digitized as HMML Pr. No. EMML 1978, fols. 30v–31v. *Täʾamrä Maryam "Miracles of Mary"*, 1813. Cataloged by Getatchew Haile and William Macomber; metadata added by Ted Erho. Accessed 3 Feb 2022.

EMML 2058 Ethiopia, Wallo Province, Ḥayq Esṭifānos Monastery, MS digitized as HMML Pr. No. EMML 2058, fol. 97rv. *Homily on the glory of Mary and Täʾamrä Maryam "Miracles of Mary"*, 18th century (?). Cataloged by Getatchew Haile and William Macomber; metadata added by Ted Erho. Accessed 3 Feb 2022.

EMML 2275 Ethiopia, Šawā Province, Darafo Māryām Church, MS digitized as HMML Pr. No. EMML 2275, fols. 157r–161r. Contents including *Täʾamrä Maryam "Miracles of Mary"*, 1508/1535. Not yet fully cataloged. Accessed 3 Feb 2022.

GG 00004 Ethiopia, Tegrāy Province, Gunda Gundē Monastery, MS C3-IV-163 digitized as HMML Pr. No. GG 00004, fols. 5r-7r. *Life of Alexis; Life of Yāsāy, the orthodox king of Rome*, 15th–16th century (?). Digitized by Ewa Balicka-Witakowska and Michael Gervers. Cataloged by Ted Erho. Accessed 3 Feb 2022. Also accessible at Gunda Gunde Manuscripts Digital Scholarship Unit Islandora site hosted by the University of Toronto Scarborough at https://gundagunde.digital.utsc.utoronto.ca/islandora/object/gundagunde%3A3307#page/1/mode/2up. Accessed 7 April 2022.

GG 00016 Ethiopia, Tegrāy Province, Gunda Gundē Monastery, MS C3-IV-154 digitized as HMML Pr. No. GG 00016, fols. 1r, 8r-15. *Homilies and Testaments*, 18th century (?). Digitized by Ewa Balicka-Witakowska and Michael Gervers. Cataloged by Ted Erho. Accessed 3 Feb 2022. Also accessible at Gunda Gunde Manuscripts Digital Scholarship Unit Islandora site hosted by the University of Toronto Scarborough at https://gundagunde.digital.utsc.utoronto.ca/islandora/object/gundagunde%3A12438#page/1/mode/2up. Accessed 7 April 2022.

Notes

[1] The precarity of what Eyob Derillo describes as Ethiopia's "distinctive, extraordinary and irreplaceable traditional school and academic system [...] which is still little known and [...] for which support is rapidly dwindling in the 21st century" [Derillo 2019, 112] is illustrated in descriptions of the past forty years. In 1981, Sergew Hable Selassie reported that "[b]ookmaking is alive and well in Ethiopia," but noted a decline in the demand for scribal work as more church books came into print and the cost of materials rose [Hable Selassie 1981, 3, 33–34]. In 2002, John Mellors and Anne Parsons interviewed 30 of "around one hundred" scribes active in South Gondar, "the only area where [manuscripts] are now produced in any quantity" [Mellors et al. 2002, 4]. In 2015, Sean Winslow found and interviewed just over 30 active scribes across multiple regions including Gondar, and reported that "pockets of scribal activity nevertheless survive throughout the country" [Winslow 2015, 23, 140–141]. In 2018, Gezae Haile identified a number of factors contributing to the "alarming rate" of decline of the tradition, especially in remote rural areas [Haile 2018, 35–37]. How the situation may have changed further as a consequence of Ethiopia's current civil war will become apparent in time.

[2] See e.g., [Stewart 2009, 606–13] [Tomaszewski et al. 2015, 92] [Kominko 2015, liii] [Derillo 2019, 104–105] [Woldeyes 2020] [Loyer 2021]. On challenges encountered during the EMML project, see [Stewart 2017, 453, 458–467]. It is worth noting that collaborations surrounding digitization of Gəʿəz manuscripts are further complicated by a history of bad-faith interactions from the 18th century onwards with western scholars and collectors seeking to appropriate Ethiopian manuscripts and cultural patrimony from their original owners [Hable Selassie 1981, 35] [Stewart 2009] [Kominko 2015] [Derillo 2019] [Woldeyes 2020] and by the disjunction of perspective between western scholars seeing manuscripts as historical artifacts, and the Ethiopian religious communities for whom the manuscripts were created, who continue to venerate and employ the manuscripts for their original cultural and religious significance [Kominko 2015] [Winslow 2015] [Haile 2018]. On some principles guiding our own development, see e.g., Liuzzo's 2019 chapters "Introduction" and "Openness and Collaboration," in particular on the need for any digital tools to be useful in Ethiopia and in the manuscripts' home environments: "It might well be that a laudable software methodology produces an output which is from the perspective of the users useless or wrong. [...] The possibility to work digitally should benefit Eritrean and Ethiopian scholars in the first place." [Liuzzo 2019, xvii, 234].

[3] See [Derillo 2019] [Akbari 2019] [Delamarter 2023].

Works Cited

Abebe 2019 Abebe, A. (2019) "Launch of Ethiopic studies program at the University of Toronto", *Tadias Magazine*, 19 December. Available at: http://www.tadias.com/12/29/2015/launch-of-ethiopic-studies-program-at-university-of-toronto/

Akbari 2019 Akbari, S.C. (2019) "Where is medieval Ethiopia? Mapping Ethiopic studies within medieval studies", in Keene, B. (ed.) *Toward a global middle ages: Encountering the world through illuminated manuscripts*. Los Angeles: The J. Paul Getty Museum, pp. 80–91.

Alrasheed et al. 2019 Alrasheed, N., Rao, P., and Grieco, V. (2021) "Character recognition of seventeenth-century Spanish American notary records using deep learning", *Digital Humanities Quarterly*, 15(4). Available at: http://www.digitalhumanities.org/dhq/vol/15/4/000581/000581.html

Appleyard 2005 Appleyard, D. (2005) "Definite markers in modern Ethiopian Semitic languages", in Khan, G. (ed.) *Semitic studies in honour of Edward Ullendorff*. Leiden: Brill, pp. 51–61. DOI: 10.1163/9789047415756_007.

Assefa et al. 2020 Assefa, D., Delamarter, S., Jost, G., Lee, R., and Niccum, C. (2020) "The textual history of the Ethiopic Old Testament Project (THEOT): Goals and initial findings", *Textus*, 29(2020), pp. 80–110. DOI: https://doi.org/10.1163/2589255X-02901002

Bausi 2005 Bausi, A. (2005) "Ancient features of Ancient Ethiopic", *Aethiopica*, 8(2005), pp. 149–169. DOI: 10.15460/aethiopica.8.1.331.

Bausi 2015 Bausi, A. (ed.) (2015) *Comparative Oriental manuscript studies: An introduction*. Hamburg: Tradition.

Bausi 2020 Bausi, A. (2020) "Ethiopia and the Christian Ecumene: Cultural transmission, translation, and reception", in Kelly, S. (ed.), *A companion to medieval Ethiopia and Eritrea*. Leiden: Brill, pp. 217–251. DOI: https://doi.org/10.1163/9789004419582_010.

Delamarter 2023 Delamarter, S. (2019) "Relationships, technology, money, and luck: The back story of six collections containing Ethiopian Arabic manuscripts and how they were digitized", in Butts, A. (ed.), *The Qurʾān and Ethiopia: Context and reception*, 8 April. Washington, D.C.: Catholic University of America (forthcoming).

Demilew and Sekeroglu 2019 Demilew, F.A., and Sekeroglu, B. (2019) "Ancient Geez script recognition using deep learning", *SN Applied Sciences*, 1(1315). DOI: 10.1007/s42452-019-1340-4.

Demissie 2011 Demissie, F. (2011) *Developing optical character recognition for Ethiopic scripts*. Masters thesis, Dalarna University. Available at:http://du.diva-portal.org/smash/record.jsf?pid=diva2%3A519067&dswid=2398.

Derat 2020 Derat, M.-L. (2020) "Before the Solomonids: Crisis, renaissance and the emergence of the Zagwe dynasty (seventh–thirteenth centuries)", in Kelly, S. (ed.), *A companion to medieval Ethiopia and Eritrea*. Leiden: Brill, pp. 31–56. DOI: https://doi.org/10.1163/9789004419582_003.

Derillo 2019 Eyob, D. (2019) "Exhibiting the Maqdala manuscripts: African scribes: Manuscript culture of Ethiopia", *African Research & Documentation*, 135(2019), pp. 102–116.

Endalamaw 2016 Endalamaw, S.G. (2016) *Ancient Ethiopic manuscript recognition using deep learning artificial neural network*. Doctoral dissertation, Addis Ababa University.

Graves 2018 Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2018) "Connectionist temporal classification: Labeling unsegmented sequence data with recurrent neural networks", in Cohen, W. and Moore, A. (eds.), *Proceedings of the 23rd International Conference on Machine Learning, Association for Computing Machinery*, pp. 369–376. DOI: 10.1145/1143844.1143891.

Grieggs et al. 2021 Grieggs, S., Shen, B., Rauch, G., Li, P., Ma, J., Chiang, D., Price, B., and Scheirer, W. (2021) "Measuring human perception to improve handwritten document transcription", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10), pp. 6594-6601. DOI: 10.1109/TPAMI.2021.3092688.

Hable Selassie 1981 Hable Selassie, S. (1981) *Bookmaking in Ethiopia*. Leiden, Netherlands: Karstens Drukkers B.V.

Haile 2018 Haile, G. (2018) "The limits of traditional methods of preserving Ethiopian Gəʿəz manuscripts", *Libri*, 68(1), pp. 33–42. DOI: 10.1515/libri-2017-0004.

Hawk et al. 2019 Hawk, B., Karaisl, A., and White, N. (2019) "Modelling medieval hands: Practical OCR for Caroline minuscule", *Digital Humanities Quarterly* 13(1). Available at: http://www.digitalhumanities.org/dhq/vol/13/1/000412/000412.html

Kahle et al. 2017 Kahle, P., Colutto, S., Hackl, G., and Mühlberger, G. (2017) "Transkribus: A service platform for transcription, recognition and retrieval of historical documents", in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 4. Kyoto, Japan: IEEE, pp. 19–24. DOI: 10.1109/ICDAR.2017.307

Kassa and Hagras 2018 Kassa, D.M., and Hagras, H. (2018) "An adaptive segmentation technique for the ancient Ethiopian Gəʿəz language digital manuscripts", in *2018 10th Computer Science and Electronic Engineering Conference (CEEC)*. Essex, UK: IEEE, pp. 83–88. DOI: 10.1109/CEEC.2018.8674218.

Kelly 2020 Kelly, S. (2020) “Introduction”, in Kelly, S. (ed.) *A companion to medieval Ethiopia and Eritrea*. Leiden: Brill, pp. 1–30. DOI: https://doi.org/10.1163/9789004419582_002.

Kominko 2015 Kominko, M. (ed.) (2015) *From dust to digital: Ten years of the Endangered Archives Programme*. Cambridge, UK: Open Book Publishers. http://dx.doi.org/10.11647/OBP.0052

Liuzzo 2019 Liuzzo, P.M. (2019) *Digital approaches to Ethiopian and Eritrean studies*. Wiesbaden: Harrassowitz Verlag.

Lor et al. 2005 Lor, P.J., and Britz, J. (2005) “Knowledge production from an African perspective: International information flows and intellectual property”, *International Information & Library Review*, 37(2), pp. 61–76. DOI: 10.1080/10572317.2005.10762667.

Loyer 2021 Loyer, J. (2021) “Collections are our relatives: Disrupting the singular, white man’s joy that shaped collections”, in Browndorf, M., Pappas, E., and Arays, A. (eds.) *The collector and the collected: Decolonizing area studies librarianship*. Sacramento, CA: Library Juice Press. Available at https://mru.arcabc.ca/islandora/object/mru%3A793/datastream/PDF/view

Manžuch 2017 Manžuch, Z. (2017) “Ethical issues in digitization of cultural heritage”, *Journal of Contemporary Archival Studies*, 4(4). Available at: http://elischolar.library.yale.edu/jcas/vol4/iss2/4.

Mellors et al. 2002 Mellors, J., and Parsons, A. (2002) *Scribes of south Gondar: Bookmaking in rural Ethiopia in the twenty-first century*. London, UK: New Cross Books.

Nosnitsin 2020 Nosnitsin, D. (2020) *Christian manuscript culture of the Ethiopian-Eritrean highlands: Some analytical insights*, in Kelly. S. (ed.) *A companion to medieval Ethiopia and Eritrea*. Brill, Leiden, pp. 282–321. DOI: https://doi.org/10.1163/9789004419582_012.

Ondari-Okemwa 2014 Ondari-Okemwa, E. (2014) “Ethical issues and indigenous knowledge production and use in sub-saharan Africa in the 21st century”, *Mediterranean Journal of Social Sciences*, 5(23), pp. 2389–2396. DOI: 10.5901/mjss.2014.v5n23p2389.

Puigcerver 2017 Puigcerver, J. (2017) “Are multidimensional recurrent layers really necessary for handwritten rext recognition?”, in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 1. Kyoto, Japan: IEEE, pp. 67–72. DOI: 10.1109/ICDAR.2017.20.

Putnam 2016 Putnam, L. (2016) “The transnational and the text-searchable: Digitized sources and the shadows they Cast”, *The American Historical Review*, 121(2), pp. 377–402. DOI: 10.1093/ahr/121.2.377.

Shi et al. 2016 Shi, B., Bai, X., and Yao, C. (2016) “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11), pp. 2298–2304. DOI: 10.1109/TPAMI.2016.2646371.

Srivastava et al. 2014 Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014) “Dropout: A simple way to prevent neural networks from overfitting”, *The Journal of Machine Learning Research*, 15(1), pp. 1929–1958. Available at: https://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf.

Stewart 2009 Stewart, C. (2009) *Yours, mine, or theirs? Historical observations on the use, collection and sharing of manuscripts in western Europe and the Christian Orient*. Piscataway, NJ, Gorgias Press. DOI: 10.31826/9781463216801.

Stewart 2017 Stewart, C. (2017) “A brief history of the Ethiopian manuscript microfilm library (EMML)”, in McCollum, A.C. (ed.) *Studies in Ethiopian languages, literature, and history: Festschrift for getatchew haile presented by his friends and colleagues*, pp. 447–472.

Sutherland and Purcell 2021 Sutherland, T., and Purcell, A. (2021) “A weapon and a tool: Decolonizing description and embracing rediscription as liberatory archival praxis”, *The International Journal of Information, Diversity, & Inclusion*, 5(1). DOI: 10.33137/ijidi.v5i1.346669.

Tadias Staff 2020 Tadias Staff (2020) “Ethiopic studies at the University of Toronto becomes permanent,” *Tadias Magazine*, 4 November. Available at: http://www.tadias.com/11/04/2020/ethiopic-studies-at-university-of-toronto-becomes-permanent-update/.

Tomaszewski et al. 2015 Tomaszewski, J., and Gervers, M. (2015) “Technological aspects of the monastic manuscript collection at May Wäyni, Ethiopia”, in Kominko, M. (ed.) *From dust to digital: Ten years of the Endangered Archives Programme*. Cambridge, UK: Open Book Publishers, pp. 89–133. DOI: 10.11647/OBP.0052.

Weninger 2005 Weninger, S. (2005) “Gəˈəz”, in Uhlig S. and Bausi, A. (eds.) *Encyclopaedia aethiopica*, Vol. 2. Wiesbaden: Harrassowitz, pp. 732–735.

Wick et al. 2020 Wick, C., Reul, C., and Puppe, F. (2002) “Calamari: A high-performance tensorflow-based deep learning package for optical character recognition”, *Digital Humanities Quarterly*, 14(1). Available at: http://www.digitalhumanities.org/dhq/vol/14/2/000451/000451.html.

Wigington et al. 2017 Wigington, C., Stewart, S., Davis, B., Barrett, B., Price, B., and Cohen, S. (2017) “Data augmentation for recognition of handwritten words and lines using a CNN-LSTM network”, in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 1. Kyoto, Japan: IEEE, pp. 639–645. DOI: 10.1109/ICDAR.2017.110

Winslow 2015 Winslow, S.M. (2015) *Ethiopian manuscript culture: Practices and contexts*. PhD thesis, University of Toronto.

Woldeyes 2020 Woldeyes, Y.G. (2020) “‘Holding living bodies in graveyards’: The violence of keeping Ethiopian manuscripts in Western institutions”, *M/C Journal*, 23(2). DOI: 10.5204/mcj.1621.

Xiao et al. 2020 Xiao, S., Peng, L., Yan, R., and Wang, S. (2020) “Deep network with pixel-level rectification and robust training for handwriting recognition”, *SN Computer Science*, 1(3), pp. 1-13. DOI: 10.1007/s42979-020-00133-y.

Yacob 2005 Yacob, D. (2005) “Ethiopic at the end of the 20th century”, *International Journal of Ethiopian Studies*, 2(1/2), pp. 121–140.

This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.