Handwritten Text Recognition Best Practice in the Beta maṣāḥǝft workflow

Hizkiel Mitiku Alemayehu

Electronic version
URL: https://journals.openedition.org/tei/4109
DOI: 10.4000/tei.4109
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Hizkiel Mitiku Alemayehu, "Handwritten Text Recognition Best Practice in the Beta maṣāḥǝft workflow", Journal of the Text Encoding Initiative [Online], Rolling Issue, Online since 02 February 2022, connection on 26 May 2022. URL: http://journals.openedition.org/tei/4109 ; DOI: https://doi.org/10.4000/tei.4109

For this publication a Creative Commons Attribution 4.0 International license has been granted by the author(s) who retain full copyright.
Handwritten Text Recognition Best Practice in the Beta maṣāḥǝft workflow

Hizkiel Mitiku Alemayehu

ABSTRACT

This contribution describes the workflow used to transcribe Manuscripts from the Ethiopian and Eritrean Tradition. The goal of the workflow is to obtain a TEI file with an initial text transcription that profits from a wealth of machine-generated information collected through community-based contributions. The author sets the framework of interest of this effort to discuss available state-of-the-art options and the actual workflow implemented. It is argued that a workflow that prefers expert post-processing in the TEI instead of refinement of the preprocessing techniques is
preferable for this specific use case. The publication of large quantities of text although, not 100% correct, when done in a collaboratively edited and open environment, can still be used and provide a user with information reusable for research.

INDEX

**Keywords:** Ethiopic literature, Manuscripts, Transkribus

ACKNOWLEDGEMENTS

This article was produced thanks to the support of the project “Beta maṣāḥǝft: Manuscripts of Ethiopia and Eritrea (Schriftkultur des christlichen Äthiopiens und Eritreas: eine multimediale Forschungsumgebung),” a long-term project funded within the framework of the Academies’ Programme (coordinated by the Union of the German Academies of Sciences and Humanities) under the survey of the Akademie der Wissenschaften in Hamburg. The project is hosted by the Hiob Ludolf Centre for Ethiopian Studies at the Universität Hamburg and is lead by Alessandro Bausi.

1. Introduction

The basis of any philological work, and to some extent also codicological work, is the transcription of the contents of the sources. In most cases for Ethiopian Studies these primary sources are manuscripts in the form of the codex. For each material witness of a given product of human intellect in writing, we can produce a TEI encoded transcription in the same way in which we can encode in TEI an edition of that text which is based upon those witnesses. If we start from the edition, we can virtually reconstruct the witnesses taken into consideration starting from a TEI encoded apparatus. If we do not yet have an edition and are working out all the steps towards a digital edition with TEI, then we will start from the transcriptions of the manuscripts, which we will eventually collate with dedicated software, then analyze in order to have the ground knowledge to encode a critical edition. Transcriptions of the witnesses require specific knowledge and expertise and are often a long and time-consuming task.
The transcription stage is not an avoidable part of the work required to get sufficient acquaintance with a text to be able to carry out the editorial and philological work, so it is not in order to replace this step in the research process that technology comes into play, but only in the time-consuming aspects of it. Sometimes available witnesses abound and the textual tradition to be taken into consideration is simply too vast to assume a human can guarantee precision and consistency. The latter are areas where current technologies can help.

In this contribution I will describe how we are integrating transcription technologies in the Beta maṣāḥǝft research environment for the manuscripts from Ethiopia and Eritrea. In order to obtain a TEI file with an initial text transcription from manuscripts, to be published alongside the catalogue description of the manuscript itself, we have investigated a series of options, among which we have chosen to use the Transkribus software by READ Coop.

The transcription matches and complements the cataloguing efforts of the project, which encodes in the `<teiHeader>` the description of the manuscript. Within `<msDesc>` the project is providing both new catalogue descriptions created directly in TEI, refined reworking of existing catalogue descriptions in historical catalogues, and TEI versions of catalogue descriptions originally done digitally but not in TEI. This involves a lot of restructuring of the source information, where, traditionally, a lot of text is copied in the catalogue descriptions, for example, to facilitate identification of intellectual contents which are not otherwise identified in the witness, e.g., by a clear title. TEI has all the tagging baggage to connect the actual text transcription with the description of features of the object described in `<msDesc>`, but, thinking of the work for an historical catalogue that involves copying from the former cataloguer transcription. Having a new transcription, based on autopsy or at least on the images of the manuscript would be preferable and technology as Transkribus allows one to obtain this transcription in an almost entirely automated way. Additionally, most of the internal referencing within a manuscript is done with the indication of the ranges of folios, and in TEI with `<locus>`. While normally a transcription either in an historical catalogue, or done by a researcher by hand, seldom records more than the folio ranges, automated techniques can identify text areas, like columns, for example, and each line, and have the necessary information to encode the `<pb>`, `<cb>` and `<lb>` which would be needed to encode the structure of the manuscript and would be a tedious task, when manually done. Having the structured text completed by the information about the layout, linked to the images, already in the
source and linking to it with the <locus> elements allows to effectively point to an exact section of the transcribed text. The syntax of the textual content of the attributes @from, @to and @target is defined by the project’s Guidelines and so is the use of the above elements, which make the references machine operable.\(^6\)

The output of any automated process will never be perfect. However, we have experienced, and learned, as it can be seen in the following section, that one can get very close to a perfect result, in facts. Researchers can benefit from the wealth of information which is generated by the software even if it is not entirely perfect, and a quicker and less perfect output is preferable in the current state of research and workflows development compared to a high expenditure of time in preprocessing to reach perfection in the automated process, where editions are badly needed but very time consuming. Besides the above-mentioned benefits, the fact that this is not a replaceable step in the research process, but one that needs to be improved, remains of the utmost importance. The editor of a text will always want to check its transcriptions with care and we assume no editor should ever trust a machine-run transcription, even if it claims to be absolutely perfect. This step, which, from a programmer’s point of view, is post-processing of software output, is from the editor’s point of view, a vital step that cannot be delegated to a machine in its entirety and provides a way to isolate and make mechanically reproducible a distinguishing task of the work. The editor will want and need to do this post-processing, in any case so, the difference is only in how quickly the editor can get going with it. Moreover, when the imperfect-but-usable transcription is made available in a way that makes it editable in a controlled but open way, like in the Beta maṣāḥǝft research environment, it becomes a further place of collaboration by different types of users who can, for example, fix errors and imperfections, add semantic markup, etc.\(^6\)

Once a partially perfect transcription of a source is published, it also becomes useful, indexed, and searchable, for the process of encoding new items descriptions, by providing ways to match text, which is still much less available than the description of it and its support in this context. Especially for the Christian Oriental tradition, this is a vital support for text identification.\(^7\)

The following steps have been taken to carry out an investigation of the possibilities for the automated production of text transcriptions based on images of manuscripts, before we opted for Transkribus and its integration in the workflow to make texts available in the Beta maṣāḥǝft research environment.
2. Related works

2.1 Research on Text Recognition

A few studies have been conducted about handwriting character recognition for Ethiopic, the more comprehensive work being the one done by Assabie and Bigun 2008. In their work the researchers present writer-independent Amharic word recognition for offline handwritten text. The underlying units of the recognition system are a set of primitive strokes whose combinations form hand-written Ethiopic characters. For each character, possibly occurring sequences of primitive strokes and their spatial relationships, collectively termed as “primitive structural features,” are stored as a list. Hidden Markov models (HMM) for Amharic words are trained with such sequences of structural features of characters constituting words. The recognition phase does not require the segmentation of characters but only requires text line detection and extraction of structural features in each text line. Text lines and primitive structural features are extracted by making use of the “direction field tensor.” The performance of the recognition system is tested by a database of unconstrained handwritten documents collected from various sources.

More recent work done by Fitehalew Ashagrie Demile build Ethiopic character recognition system using “deep convolutions neural network” (Sekeroglu and Ashagrie 2019). The proposed system obtained an accuracy of 99.39% with a model loss of 0.044 which demonstrates its efficiency. Another Research performed by Gondere et al. 2019 designs for the first time a model for Amharic hand-written character recognition using a convolutional neural network. The dataset was organized from collected sample handwritten documents and data augmentation was applied for machine learning. The model was further enhanced using multi-task learning from the relationships of the characters.

These models are designed for Amharic, not for Classical Ethiopic, but could theoretically be reused, given the Fidal syllabary is the same for both languages. However, they are not available for this purpose or are too complex to replicate to any degree, thus becoming for our practical purposes unusable. We have thus investigated other options.
2.2 HTR Tools and Computer Vision Libraries

2.2.1 OpenCV

OpenCV is an open source computer vision and machine learning software library. It is primarily used to detect an object. Specific to OCR technologies, OpenCV helps to perform image segmentation. The output of the image segmentation is an input for the model training stage.

Saritha et al. (2020) used OpenCV for preprocessing and segmenting an image. An image is read and then stored in multiple copies for performing different operations. They use the library to read images, detect the letters, crop and resize.

Then they train a Convolutional Neural Network (CNN) model using the preprocessed and segmented image. This model gives them the training accuracy of 91.84% and the test accuracy of 96.26%.

It is a known fact that in relation to image quality most historical documents suffer some common problems which makes the recognition process difficult. Therefore, in order to eliminate undesired noise in an image and prepare the image for the next stage, preprocessing is crucial. For this stage, one can use OpenCV for cleaning a given image. Thus the application of OpenCV for HTR technologies is limited to preprocessing stage.

2.2.2 Google AI Vision

Cloud Vision API’s text recognition feature is able to detect a wide variety of languages and can detect multiple languages within a single image. It includes training data for Amharic language, which however it is not complete.

Google technology, has been recently revised for Syriac by Ephrem A. Ishac, in two blog posts (2020a, 2020b) with encouraging results. Our tests for Ethiopic have not yield similarly encouraging results, especially when it comes to large quantities of images. Although these solutions are good for small tasks they do not seem to be scalable enough or to be able to become part of a stable workflow for a large collaborative project.

2.2.3 Tesseract

Tesseract was initially released as open source in 2005 and is still under development. The newest version 4 of Tesseract adds support for deep network architectures (Wick et al. 2020).
This OCR engine has Unicode (UTF-8) support and can recognize more than 100 languages out of the box. Tesseract supports various output formats: plain text, OCR (HTML), PDF, invisible-text-only PDF, TSV. The master branch also has experimental support for ALTO (XML) output.

Other OCR-based tasks have been carried out for Classical Ethiopic, and satisfactory results have been informally reported, but the investigation run in this context lead to a non-reproducible workflow, which would require many additional steps to get to our desired output. OCR systems remain good individual support especially for printed books and for sources in more than one script.

2.2.4 Transkribus

This software is freely accessible and has a subscription model based on credits. The platform was created within the framework of the EU projects tranScriptorium and READ (Recognition and Enrichment of Archival Documents - 2016–2019). It was developed by the University of Innsbruck and the Digitization and Electronic Archiving Group. Since July 1, 2019, the platform has been operated and further developed by the European cooperative READ-COOP (Guenther et al. 2019). Tools from various research groups from all over Europe are integrated into the platform. The Pattern Recognition and Human Language Technology (PRHLT) group of the Universitat Politècnica de València and the CITlab group of the University of Rostock should be mentioned in particular.

Transkribus comes as an “expert tool” in its downloadable version and its online version, and it allows to upload images privately and perform the transcription task without neural networks knowledge, but just following concise and easily retrievable documentation.

The expert tool offers in one place both the layout recognition, HTR model training and the application of the HTR model of choice, thus constituting a real one-stop-shop for the task of transcription of images. On top of these specialized tools, the export functionality includes a TEI export which produces not only the necessary text structuring elements, but also the elements identified and anchored to them. Not only, then, the text is transcribed, it is also correctly structured and aligned to the exact portions of the images on which the transcription is based. This is an additional benefit that the researcher obtains, and which can be leveraged by any other application reusing that TEI, to retrieve for a given portion of text referred to in not only
the corresponding correct piece of text transcription but also the correct images or portions of
image on which the transcription is based. This is not entirely unproblematic of course, as we shall
note in the following section.

3. Experiment

For testing any of the above tools involving the training of a model, shortage of initial training data
was the main roadblock. Most of the techniques and tools mentioned use deep learning for training
models. This means that they require correct training data. However, there is no organized and
freely available dataset for Ethiopic handwriting character recognition.

Thus, the first stage for developing a model was gathering the data and preparing an initial dataset.
Also for this aspect, Transkribus proved superior to all other options offering support also for this
step. Colleagues which we called to contribute could be added to a collection, share their images
without publishing them and add their transcriptions in the tool with a very mild learning curve.

Within Transkribus we have trained a model called "Manuscripts from Ethiopia and Eritrea in
Classical Ethiopic (Gǝʿǝz)."10 Checked transcriptions for the training set have been kindly provided by

- Alessandro Bausi for Baḥerāwi Kallalāwi Mangašti Tǝgrāy, ‘Urā Qirqos, UM-039,10 ff. 16vb-29va;
- Antonella Brita for Tǝgrāy, Dagwʿa Tamben, Dur Ambā Šollāsse, MS 002, ff. 101va-110ra;
- Dorothea Reule for Gāntā ‘Afāšum, Qandāʾro Qaddus Qirqos, QDQ-004,11 ff. 97ra-101vb, 104ra-109rb;
- Nafisa Valieva for London, British Library, BL Oriental 718,12 ff. 1ra-7vb, images British
  Library;
- Several parts of manuscripts transcribed by Jeremy Brown and pertaining to the Miracle
  of the Cannibal of Qemer (CAe 3615);13
- Other important but more limited contributions.14
Because the style of manuscripts writing is changing over time, we have opted for training a generic model, and thus fed to the training a mixture of different manuscripts and styles. The machine learning is sensitive to the quality of the images, and we have paid some attention to the diversity of types of images, avoiding however to feed very good images to the system, knowing that is often not what users will have. A machine learning algorithm that is trained using the data set of a specific period will not be able to work for other manuscripts, but this mixed one, maybe a good basis for more specific models to be trained and made available in the future. Ethiopic manuscripts are still produced today and the writing style is evolving, especially in conjunction with phenomena such as the production of manuscripts on parchment based on printed editions. A machine learning algorithm cannot take palaeography of manuscripts into consideration, but from a core model, a user could transcribe a small portion of its manuscript, train a specific model based on the generic one and perform the rest of the transcription with higher precision and correctness, specific to the image set.

4. Training a model in Transkribus

Gathering data to train an HTR model in Transkribus was not easy. Researchers were directly asked to contribute images of which they had already done the correct transcription. Sets of images with the relative transcription was thus obtained thanks to the generosity of contributors listed above.

As stated earlier, we have trained a generic model using various styles and manuscripts. The simple fact of having the images and the transcriptions was not enough of course. These needed to be cleaned up at least for what concerns the file naming before being uploaded to the expert tool. After that, the layout analysis was carried out and hand-fixed. This process took some weeks since the diversity of the datasets brought with it also a series of issues in this step, like, for example, the recognition of the folding in the center of an image of an opening as one text area, or the lack of recognition of rubricated text.

Once the alignment was fixed and satisfactory we entered the transcriptions. These often came as running text in a word file and had to be copy-pasted to each line box in the expert tool with a tedious process which however led also to the discovery of several errors in the hand-
made and “correct” transcription for the validation set, thus bringing a further benefit to the contributors, while demonstrating an area of the workflow where the precision of the computer-assisted transcription could be already visible.

The model was initially trained with a smaller dataset of about 15k words, with the intent to use it as a base to produce transcriptions which the colleagues would have checked. Two tests of this kind have however shown us that it took us less time to enter more of the available transcription by hand as discussed above, than to wait for the available time of the colleagues to fix the work of the machine, since we intended to train the model again. After three months with a full-time dedicated person, we had more than 50k words in the Transkribus expert tool, and we could train a model which could be made public, since this is the unofficial threshold to make a model available to everyone.

The features of the final model can be seen in figure 1.
As can be seen from the table the Character Error Rate (CER) is below 6%, with both the train and the test set. It means that it can be expected on average that using this model less than one in 25 letters will be recognized incorrectly.

Most of the errors that occur during automatic transcription are related to diacritic signs and rubricated texts and are thus also easily identifiable.

We still plan to train the model again as more correct transcriptions become available, hopefully as corrections of the transcriptions produced with this model. Once the model is publicly available eventually anyone will be able to do so.
5. Adding transcriptions to Beta maṣāḥǝft from Transkribus

Even if a user already worked through each page of a manuscript to produce a transcription, doing it again with Transkribus and checking it has many advantages, chiefly the alignment of the text regions and lines on the base image to the transcription.\textsuperscript{16}

With the transcribed images, either by hand with the help of the tool, or using the HTR model, the export functionalities of the Transkribus tool, allow to download a TEI encoded version of this transcription where we encourage users to use Line Breaks (\texttt{<lb>}) instead of \texttt{<l>} and preserve the coordinates of the boxes.

This TEI file contains all the aligned transcription, links between the regions of the image, and the text. It has however to follow the structure of the set of images. If you transcribed images, for example, of openings, logically you will have a page break for each image, not for each page-break in the manuscript. This TEI is thus not ready to be copy pasted into the TEI file for a Manuscript in the Beta maṣāḥǝft Research environment where instead the structure of the manuscript expects \texttt{<pb>} and \texttt{<cb>} elements to mark the page and column breaks of the manuscript and not of the image set. Most of this can be fixed by preparing the image set accurately, but we assume in most real-life use cases this will not be the case.

We have then prepared a bespoke XSLT transformation which can be used to transform the rich TEI from Transkribus, called \texttt{transkribus2Beta maṣāḥǝft.xsl}. This transformation, given a few parameters, restructures the TEI to fit the project requirements. The needed parameters are: the total number of foliated leaves, the number of protection leaves at the beginning if this is part of your image set, and the type of images (if single-side or openings). The assumption is that your set of images will be tidy in this respect, that is to say, internally coherent and not made of some openings and some single leaves.

The result of this transformation is not yet ready to be pasted into the correct TEI Manuscript record, because, at least in our experiments, it is more often the case that people require parts of a manuscript, then not the entire transcription. Some hand fixing will still be necessary, for example for the enumeration, of the structure of text regions that contain additions or other types of contents, like legends of decorations, or extras.
The output of the transformation can be added to the TEI file for a manuscript, eventually as one of many possible transcriptions, and we encourage contributors to document the origin and processing of the transcription within the file, as well as to further encode those features of the text which are related to its transcription (erasures, interlinear additions, rubrication, etc.).

6. Conclusions

Working with Transkribus for the Beta maṣāḥǝft project gives the community of users a way to support the process of transcribing to the text on source manuscripts without typing it down. This is not intended to substitute the work of the editor of a text, but to support it, producing a transcription that still needs a lot of care for its content and encoding, but also comes with a lot of added value, like the precise alignment of the text to the image set and its encoding in the TEI. Files thus obtained are huge and not as easy to maintain in a database or edit directly. However, even if the text of the transcription is still unchecked and thus subject to at least the percentage of the error the model provides, several benefits become immediately available to the users, both encoders and users of the web application hosting the texts. Encoders can point to a part of the transcription using <locus> and avoid keying in the text from the transcription of a cataloguer in the <teiHeader>. Similarly, a user of the application can identify an unknown text using the functionality of the search index, which is capable of performing fuzzy searches which will return results also where a query term is partially different from the matching result, e.g., in case this contains an error originated from the automated transcription process.

With this process started, the model publicly available, and thousands of images of manuscripts, we work for more transcriptions, more text distributed as TEI on the web, and more collaboration to improve each of these aspects.

BIBLIOGRAPHY


NOTES


2 Computer supported collation is a whole field of research, see many contributions in the volume edited by Andrews and Macé (2014).

3 Accessed February 2, 2022, https://betamasahfeit.eu/. See Liuzzo 2019. The project Beta maṣḥǝft: Manuscripts of Ethiopia and Eritrea (Schriftkultur des christlichen Äthiopiens und Eritreas: eine multimediale Forschungsbegebung) is a long-term project funded within the framework of the Academies’ Programme (coordinated by the Union of the German Academies of Sciences and Humanities) under survey of the Akademie der Wissenschaften in Hamburg. The project is hosted by the Hiob Ludolf Centre for Ethiopian Studies at the Universität Hamburg and is led by Prof. Alessandro Bausi.


5 On this process see Reule 2018.

6 This is the case for Beta maṣḥǝft, which implements this experimentally, via the Distributed Text Services API Specifications. See Beta maṣḥǝft Guidelines, accessed February 2, 2022, https://betamasahfeit.eu/Guidelines/?id=references.

7 This means that, after preprocessing, the researchers train a model that takes a pixel matrix of an image to extract image features (such as edges, lines, corners and abstract features). Then reduce the features to the spatial volume. Finally, the images (in our case letters) will be labelled and classified accordingly.


13 Especially generous in sharing images and related transcriptions were, Steve Delamarter and Ralph Lee for the Ethiopic Manuscript Imaging Project and Wendy Belcher of the Princeton Ethiopic Miracles of Mary project.

14 Other transcriptions with relative images where contributed by Antonella Brita, Alessandro Bausi, Dorothea Reule, and Denis Nosnitsin.

15 See Uhlig 1990 for the palaeography of Ethiopic manuscripts.

16 Guidelines are provided for this steps to the users in the project Guidelines, accessed February 2, 2022, https://betamasaheft.eu/Guidelines/?id=transcription-transkribus.

AUTHOR

HIZKIEL MITIKU ALEMAYEHU

Hizkiel Mitiku Alemayehu graduated in Information Systems at Addis Ababa University and in Digital Humanities and Digital Knowledge at University of Bologna. His main interests concern programming and natural language processing. Hizkiel did his traineeship at the Hiob Ludolf Centre for Ethiopian and Eritrean Studies, Universität Hamburg. He dealt with Ethiopic HTR using Transkribus. Currently, he is a Research Assistant at the University of Paderborn.