Large Scale Computational Analysis of Historical Manuscripts: The MiDRASH project and Its Applicability to Other Cultures – First Steps

Daniel Stökl Ben Ezra, Luigi Bambaci, Nachum Dershowitz, Ben Kiessling, Avi Shmidman

MiDRASH is an ERC Synergy project funding six teams in France and in Israel with more than 10 M euros for the period October 2024 to September 2029. It strives to reconstruct Jewish scribal culture from roughly 800 CE to 1600 CE – from England in the West to Iran, Yemen and India in the East – by transcribing, lemmatizing and analyzing intertextually at least 30.000 literary manuscripts with millions of pages and 300.000 fragments written in Hebrew script in three languages, Hebrew, Aramaic and Judeo-Arabic (as well as mixtures of these).

We present the most recent state of our AI pipeline from catalog mining, layout classification, layout segmentation, language and script classification, via script recognition and post-OCR correction, to code-switching detection, morpho-syntactical lemmatization and the detection of intertextual relationships. While the specific hyperparameters of the various neural networks are obviously tailored to our material culture, the general pipeline is or shall be partially or completely applicable to other cultures as well.

Most bricks of the ATR (automatic text recognition) part surrounding our open-source ATR platform eScriptorium cum kraken have already been successfully applied not only to historical Hebrew script (4.9% CER) but also to Arabic, Georgian (8.9%), minuscule Greek (5.5%), Latin (6.2%) and Syriac (2.6%), with some of it also to Japanese (5.2%) and Chinese (1.04%) documents. While we almost always use transfer learning, the hyperparameters, especially learning rate, for very large codecs are not the same.

Our presentation will therefore focus on the techniques and workflows that are the most easily transferable: (1) How to exploit existing catalogs? (2) How to exploit existing etexts via self-supervised training to create large amounts of high quality ATR training data with relatively low human labor. (3) How to profit from the existence of digital images of transcribed and untranscribed historical prints? (4) How to define conventions for layout classification and segmentation as well as recognition to arrive as quickly as possible to high quality training data.

(1) Thanks to the KTIV project of the National Library of Israel, we are in the privileged situation of having a unified catalog for (almost) all historical Hebrew script manuscripts worldwide whether in private or in public collections. It comprises among others legacy data on dates and script types, genre, sometimes composition and author of varying quality on the granularity of the codicological unit. We will crosscheck its dates and provenances with the most recent printed catalogs and with a team of human experts and with a date classifier trained on all available manuscripts with internal dates and places. The catalog for the electronic books (see below) is much more detailed and usually contains the precise publishing year(s), place, mall and/or printer, the title and the author.

(2) Several large etext collections exist among which Sefaria is the most well known. Existing etexts are most often results of more or less successful OCRs of historical prints. We currently have several alternative modules to automatically align existing texts with rough manuscript ATR for semi- or self-supervised ground truth creation: passim has been directly integrated into eScriptorium and allows the automatic alignment of a single text. seriatim, closely related to passim but outside of eScriptorium, can automatically align many texts and choose the best candidate. ACDC which combines seriatim and our ATR engine kraken to create entirely self-supervised xground truth. Our teams from Haifa University and Tel Aviv University have developed the Sofer Stam pipeline based on elastic-search combined with a BERT model to automatically propose improvements for the alignments where the etext and the rough ATR differ due to abbreviations or variant readings. The Paris team developed a Smith-Waterman and Needleman-Wunsch based application that allows also the insertion of line-fillers, the replacement of ligatures and the selection of specific code points from the ATR not extant in the etext to be merged into the ground truth.

(3) We have received digital images of almost 15.000 historical books from before 1870. Many of them represent material that existed in manuscript form before 1600. Fonts and layout are frequently modeled on manuscripts from different regions. As prints are usually more regular than manuscripts, it is easier to enlarge the amount of etexts via an ATRisation of these electronic books and then exploit them in the alignment process described in the previous paragraph. Of specific scientific editions with a critical apparatus, even a full retroversion can be useful (Bambaci et al. 2023)

(4) Layout analysis is perhaps the most crucial step in the whole pipeline as any line or region missed or wrongly joined has a huge impact on all subsequent steps. However, manuscripts can have a huge variability in layouts. Many manuscripts contain multiple texts at once, e.g. a central hypotext and commentaries around, or a base text and its facing translation. While it is extremely difficult if not impossible to train a global layout segmentation model, we have had very good experience in training layout segmentation models adapted to a certain type of layout. Therefore, a preliminary layout classification step has proven immensely useful.

Bibliography

Bambaci, Luigi and Stökl Ben Ezra, Daniel: Enhancing HTR of Historical Texts through Scholarly Editions: A Case Study from an Ancient Collation of the Hebrew Bible. CHR 2023: 554-576.

Kiessling, B., Tissot, R., Stökl Ben Ezra, D., Stokes, P.: eScriptorium: An Open Source Platform for Historical Document Analysis, *OST@ICDAR* 2019 (2019).

Miller, Hadar, Philips, Yoav, Kuflik, Tsvi, Lavee, Moshe, Dershowitz, Nachum, and Londner, Samuel: December 2021, "Towards Automatic Cataloguing of Hebrew Manuscripts" (Abstract), *Second Workshop on Digital Technologies to Study the Past and Present (SfP 2021)*, Kinneret College on the Sea of Galilee.

Smith, David A., Murel, Jacob, Allen, Jonathan Parkes, Miller, Matthew Thomas: Automatic Collation for Diversifying Corpora: Commonly Copied Texts as Distant Supervision for Handwritten Text Recognition. CHR 2023: 206-221.

Stokes, P. A., Kiessling, B., Stökl Ben Ezra, D., Tissot, R., and Gargem, H.: The eScriptorium VRE for Manuscript Cultures. *Ancient Manuscripts and Virtual Research Environments*, ed. Claire Clivaz and Garrick V. Allen. Special issue of *Classics@* 18 (2021).

Stökl Ben Ezra, D.: Computational Document Analysis: New and Open Questions from a Pragmatic Perspective, Keynote Lecture, ICFHR 2020.