HTR4PGP: Bootstrapping Automatic Transcription of Medieval Documents in Hebrew Script from the Cairo Geniza

Marina Rustow*[1], Daniel Stoekl Ben Ezra*[2,3], Bronson Brown-Devost[2,3,4], Jessica Parker[1], Zohar Berman[1], Itay Zandbank[5], and Devorah Witty[5]

[1]Princeton University – United States
[2]EPHE, PSL University – EPHE, PSL University – France
[3]U.M.R.8546-Laboratoire AOROC, 4 Rue Lhomond, 75005 Paris, France – U.M.R.8546-Laboratoire AOROC, 4 Rue Lhomond, 75005 Paris, France – France
[4]Göttingen University – Germany
[5]The Research Software Company. – Israel

Abstract

The Cairo Geniza is a cache of manuscripts discovered in the 19th century in a medieval Egyptian synagogue. Roughly 40,000 geniza fragments are documentary texts such as letters, legal deeds, accounts and lists, most dating from the 11th–13th centuries and written in Hebrew, Aramaic and Judaeo-Arabic (Arabic in Hebrew characters). They are now housed in more than sixty libraries and private collections.
Geniza documents have transformed the history of the medieval Islamicate world and its Jewish communities with fine-grained information about the daily lives of medieval women, children, enslaved people, peasants and the anonymous masses, offering a corrective to sources focused on state officials and religious experts. But the texts have remained the preserve of hyperspecialists with training in history, semitic languages and palaeography. To date, only 10% of geniza documents have been transcribed and rendered as searchable text, work that has occupied the Princeton Geniza Project team since 1986. At the current pace, transcribing the rest of the cache would take centuries.

The Princeton team has therefore partnered with the eScriptorium team from the EPHE,PSL, to accelerate transcription through machine-aided palaeography. Our project, HTR4PGP (Handwritten Text Recognition for the Princeton Geniza Project), uses an open-source annotation platform for machine learning called eScriptorium, built around the kraken HTR engine and designed to handle historical languages and complex page layouts.

The challenges the documents present are multiple: they are highly fragmentary, they code-switch, some layouts are complex, and the scribes were writing over a broad geographic expanse and many centuries in widely variable hands and scribal registers. The legacy PGP transcriptions, moreover, use different sets of conventions for scholarly transcription and editing.

Our paper will present the bootstrapping process we've devised to overcome the challenges

*Speaker

the documents pose to automatic analysis. Our dataset consists of 2,300 fragments (4,600 images) with TEI-XML transcriptions from PGP and IIIF-compliant images from the Cambridge University Digital Library (CUDL). About 40% of the texts are laid out simply in single columns and straight horizontal lines; the remaining 60% contain lines written perpendicularly, diagonally, or upside-down, and in some cases, the writing support has been reused by different hands. For instance, an especially rich, fascinating and common type of geniza text are personal and business letters; these are typically written with diagonal / upside-down lines throughout the margins.

To create ground truth to train general recognition models, we devised an iterative process first to improve existing layout segmentation models, beginning with the simple layouts, then continuing to complex layouts. Our team manually corrected the results of the automatic segmentation, automated the homogenization of PGP edition conventions, then automatically transcribed the texts and manually aligned them with the PGP transcriptions. Our challenge has been to create a new set of HTR models that are few enough to be based on an adequate quantity of ground truth but also specialized enough to handle a heterogenous corpus.

Keywords: HTR, Medieval Documents, Hebrew script, Judeo, Arabic, Aramaic, Hebrew