Digital Research Library for Multi-hierarchical Interrelated Texts: From "Tikkoun Sofrim" Text Production to Text Modeling

Moshe Lavee, The University of Haifa

Tikkoun Sofrim Team: Daniel Stoekl Ben-Ezra, Benjamin Kiessling, Pawel Jablonski, Avigayil Ohali, Lily Stevenson (EPHE, PSL, Paris), Tsvi Kuflik, Moshe Lavee, Vered Raziel-Kretzmer, Uri Schor, Dror Elovits, Alan Wekcer, Moshe Schorr (The University of Haifa).

"Tikkoun Sofrim" is a multifaceted project, integrating ML based Handwritten Text Recognition with citizen science correction for full textual availability of Hebrew manuscripts. The project developed a comprehensive pipeline including the following elements:

a. Automatic transcription of Manuscripts (Kraken).
b. Ergonomic crowdsourcing platform for crowd correction (https://tikkoun-sofrim.haifa.ac.il/).
c. Producing reliable texts by supervised algorithmic aggregation of crowd corrections.
d. Modeling data structures for efficient digital critical editions based on the aggregated transcriptions.
e. Creating a demo digital edition (following http://erabbinica.editions.org/).

The tools and models developed in the project are expected to enable the fulfillment of a wider vision – STAM (Systematic Textual Availability of Manuscripts). In the near future, with the advancement of digitization of Hebrew manuscripts through "Ktiv", more than 12M images will be available. Our project lay the foundations for making this cultural heritage treasure available and computable not only as in enriched text formats, and through combined image-texts viewers.

We have chosen to experiment with a specific sub-set of medieval Hebrew manuscripts, the homiletic genre of Tanhuma-Yelamdenu. Based on homilies from Late Antiquity, these texts underwent a complex process of creative transmission through the middle ages. As such they Constitute an extremely rich and complicated corpus, which was never fully treated in a sufficient critical edition, and they posit unique challenges which may be addressed through a dynamic digital critical edition.

After reviewing the wider context of the project, the current presentation will focus on its latest stages [C-E]. We will present the algorithms used for aggregating crowd corrections, and the data model. Inspired by Sharing Ancient Wisdoms (SAWS) project and based on an extension of CapiTainS the data model contains (1) TEI schema for encoding each witness (2) tailored CTS, and (3) a Semantic Web (RDF) to represent relationships between different kinds of parallel passages.