MiDRASH – A Project for Computational Analysis of Medieval Hebrew Manuscripts

D. Vasyutinsky Shapira, B. Kurar-Barakat, S. Gogawale, M. Suliman, and N. Dershowitz

Tel Aviv University

Abstract

MiDRASH is an international effort that aims to construct a groundbreaking interdisciplinary methodology for a global approach to the study of the treasure trove of medieval literary manuscripts in Hebrew script. It studies materiality, textuality, transmission and historical contexts of the digitized manuscripts in Hebrew, Aramaic, Judeo-Arabic and other vernacular languages.

CCS Concepts

• Human-centered computing → ***Empirical studies in collaborative and social computing; Collaborative and social computing systems and tools; • Applied computing*** → ***Document searching; Digital libraries and archives; • Computing methodologies*** → *Cluster analysis;*

1. Introduction

MiDRASH – Migrations of Textual and Scribal Traditions via Large-Scale Computational Analysis of Medieval Manuscripts in Hebrew Script, is the first ERC Synergy grant in Jewish studies and the first for computational manuscript studies (2023–2029). The project is led by Daniel Stökl Ben Ezra (École pratique des hautes études [EPHE], Paris Sciences-Lettres University), Judith Olszowy-Schlanger (École pratique des hautes études, Paris Sciences-Lettres University and Oxford University), Nachum Dershowitz (Tel Aviv University [TAU]), and Avi Shmidman (Bar-Ilan University [BIU]), with the participation of the National Library of Israel (NLI) and Haifa University.

The overall ambition of this MiDRASH is to construct a groundbreaking interdisciplinary methodology for a novel global approach to the study of the rich trove of extant medieval manuscripts in Hebrew script. Utilizing computational tools, we analyze – from selected paleographic, codicological, linguistic and literary perspectives – the entire medieval Jewish manuscript culture as reflected in its literary production and as preserved in Hebrew-character manuscripts prior to the printing press. The outcomes will further our understanding of crucial issues of the manuscripts' materiality, textuality, transmission and the historical and intellectual context of their making and readership. Combining traditional philology with machine learning, computer vision, and computational linguistics, we will process huge amount of textual and paleographical data that could not be handled by traditional philology (Figure 1). Our team at Tel Aviv University applies deep machine learning and other advanced computer science techniques to paleographical as well as printed data.

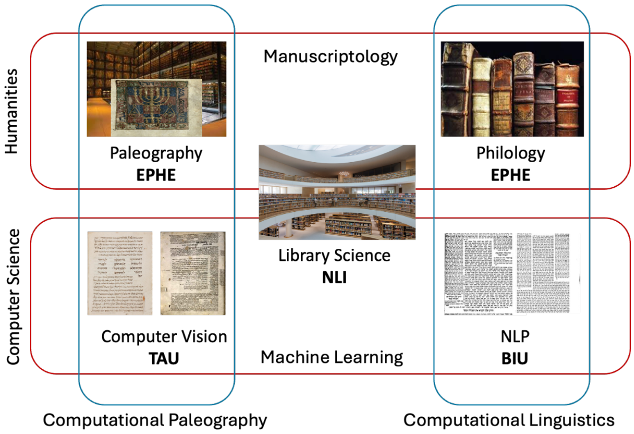

Figure 1: *Synergy in computational manuscriptology.*

2. Aims

The principal aims of the project include:

- Convert the writing appearing in the images of a large fraction of those manuscripts into searchable text.
- Have the computer compare the multitude of texts to find quotations, paraphrases, borrowings, allusions, and other intertextual relations.
- Train computer algorithms to analyze handwriting so as to determine where and when each manuscript was written.
- Train computer algorithms to analyze the extracted texts for lin-

guistic features that can be used in textual searches and that can help place literary works in their historical context.

- Employ both traditional and computational methods for paleographic, philological, and textual analysis, to map out the migration, genesis, and evolution of texts, ideas, readers, scribes, and books within medieval Jewish communities.

3. The Data

The Ktiv project at the National Library of Israel has led an important digitization campaign of all Hebrew-character manuscripts from collections from all over the world and accumulated now more than 80% of the extant manuscripts, accessible via a unified catalog (nearly 100,000 manuscripts). It has been augmented through the addition of the Friedberg Genizah Project with images and metadata of ca. 350,000 fragments from medieval book and document depositories, "genizot". This digital corpus is the source material for our investigation. It is composed of relatively well-preserved scrolls and codices as well as hundreds of thousands of fragments.

4. Methodology

Converting images of manuscripts into searchable text requires significant improvements in the state-of-the-art of computerized page segmentation and handwritten-text recognition. This is the initial stage and a pre-requirement for further computerized research of manuscripts and will serve as the first step to render the massive corpus of images into machine-readable text and make machine-readable annotated text and metadata of this corpus available to all. The different teams of the project are addressing these tasks, joining, extending, and improving our existing cutting-edge open-source tools for historical document analysis.

The accessibility of the textual and non-textual information of the manuscripts can be profitable only if we can approach the texts in their specificity and complexity, in their place and time. Training computer algorithms to determine when and where a manuscript was written is another primary task necessary for future research on digitized manuscripts. Among the whole corpus of medieval Hebrew manuscripts, only some 3500 are dated (they have colophons, i.e., scribe's notes, or owners' notes). The most important methods to establish manuscript provenance are paleography (study of handwriting scripts) and codicology (study of the material aspect of codices). The only currently existing database of dated medieval Hebrew manuscripts, SfarData (https://sfardata.nli.org.il), is first and foremost a codicological project. However, a project that studies digitized images must rely mostly on paleography. The paleography team of MiDRASH aims to propose finer regional and chronological typologies, using extensive but well-defined samples of manuscripts and exploiting the correlations between their local textual features and their script. Their efforts are reflected in HebrewPal (https://www.hebrewpalaeography.com), the comprehensive database of Hebrew paleography. Processing this data involves synergetic feedback between paleographers and modern neural-network machine learning.

The paleographical analysis will be accompanied by linguistic analysis through natural language processing to provenance both manuscripts and compositions geographically, chronologically, and, to a certain extent, sociologically [MDB24]. Morphological annotation and a cross-linguistic intertextuality graph will enable us to identify and position unknown texts or new versions of known texts with respect to the rest of Jewish literature. This analysis will contribute to a better definition of the extant manuscripts, including anthologies and miscellanies, whose texts are often not identified in the existing catalogs.

5. Progress

The MiDRASH project is ongoing; the teams are working in synergy on page segmentation, reading order, automatic dating, and numerous other tasks. The outputs of the project will be publicly available. In some of the tasks, we are trying to advance—by means of computer sciences—beyond the borders of human learning. We describe one example of such a task.

Medieval Hebrew manuscripts are described by paleographers according to their script mode (square, cursive, and the in-between, semi-cursive mode) and their geographical type. The six geographical types are Oriental (Egypt, Palestine, Syria and Lebanon, Iraq, Iran, Uzbekistan and Bukhara, eastern Turkey); Sephardic (the Iberian Peninsula, Provence and Languedoc, North Africa, and Sicily); Italian; Ashkenazi (France and England, the Holy Roman Empire, Central and Eastern Europe), Byzantine (Greece, the Balkans, western Asia Minor, and regions surrounding the Black Sea); and Yemenite. Sometimes, we know that within a certain script type-mode there are distinct subgroups. In some rare cases, these subgroups have been relatively well-studied. However, even the most experienced paleographer will know best those manuscripts s/he mostly works with, and no human memory can keep many thousands of examples of a script. Besides, within some script type-modes, the variations are very subtle. For this reason, we are employing automatic clustering of lesser-studied script types and sub-clustering of the better-studied ones (Figure 2). When this will be successful, it will significantly advance both conventional and computer paleography.

6. Conclusions

The MiDRASH project will further our knowledge of the production and transmission of Hebrew manuscripts and texts, their authors, scribes and readers, and enhance their role as the pivotal aspect of European and Mediterranean intellectual history.

References

[MDB24] Mitelman D. W., Dershowitz N., Bar K.: Code-switching and back-transliteration using a bilingual model. In *Findings of the Association for Computational Linguistics: EACL 2024* (2024), pp. 1501–1511. URL: https://aclanthology.org/2024.findings-eacl.102.pdf. 2

Proceedings published by Eurographics - The European Association for Computer Graphics.

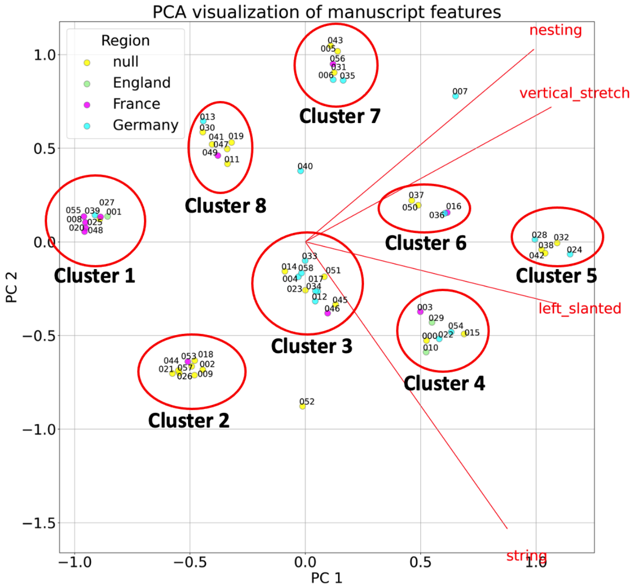

Figure 2: *Preliminary results of automated sub-clustering of Ashkenazi square scripts using paleographical features. Data annotation was facilitated using the Hasty AI assisted annotation tool (https://hasty.cloudfactory.com/).*

Proceedings published by Eurographics - The European Association for Computer Graphics.