Querying Hebrew Texts via Word Spotting

Adiel Ben-Shalom,* Adi Silberpfennig,* Nachum Dershowitz,* Lior Wolf,* Yaacov Choueka

*The Blavatnik School of Computer Science

Tel Aviv University

We report on recent results with word-spotting (WS) in Hebrew historical texts, manuscript and printed. The advantage of such a retrieval system is that it works on images without any need for manual or computer transcription of the texts. The method allows for extremely rapid querying, while still maintaining high accuracy; thus, it should be considered as an important tool in historical textual research.

We describe a successful deployment of our WS system within the Friedberg Cairo Genizah website. The index comprised all Judeo-Arabic documents within the Taylor-Schechter collection (T-S Ar). Querying is performed interactively on the Genizah website by allowing the user to the draw a rectangle around a word anywhere in the entire digitized Genizah and search for visually similar words within T-S Ar.

We have also experimented with spotting for newspapers that have been digitized as part of the Historical Jewish Press project (JPress), but for which poor OCR leads to a high percentage of word identification and segmentation errors. We found that—for the purposes of topic detection and locating related articles—word spotting can achieve reasonably high accuracy, comparable to traditional text-based methods working with noisy OCR.

In addition, we present recent results on using letter-level spotting to compare Qumran texts and to improve transcript alignment in Mishnaic manuscripts.