Automatic Scribal Analysis of Tibetan Writings

Lior Wolf, Tel Aviv University
Nachum Dershowitz, Tel Aviv University

With the increasing access to old Tibetan manuscripts and xylographs in recent decades, scholars of Tibetan textual studies are faced with new challenges and opportunities. Whereas, until recently, the content of this material has garnered the bulk of researchers' attention, we are seeing increasing interest in codicological, paleographical, and material aspects of these documents.

Following a successful interdisciplinary workshop held in Hamburg, we have been collaborating with Orna Almogi and Dorji Wangchuk (University of Hamburg) in analysing Tibetan manuscripts. We apply the same methods to these Tibetan manuscripts as have been successful in our recent work with the Cairo Genizah. The Genizah is a collection of handwritten documents containing some 350,000 fragments discovered in Cairo in the late 19th century. Most fragments were written between the 10th and the 14th centuries, almost all of them in Hebrew characters, but in a variety of languages (Hebrew, Judeo-Arabic and Aramaic). Today, the fragments are spread out in more than seventy collections worldwide. Using computer-vision and machine-learning algorithms, we have been able to automatically classify Genizah manuscripts by script style and to identify hundreds of new "joins", that is, matches between leaves in the same hand and originally part of the same manuscript, but now catalogued separately.

Initial experiments were conducted with the first 30 volumes of the bKa' gdams gsung 'bum collection. These volumes contain 123 different manuscripts written in a variety of scripts, ranging from dBu can to different kinds of dBu med, and in various hands. Our results show that the same software is able to accurately match manuscripts that were written in the same script and subtypes of scripts. Possibly, pending further verification, scribal matching (identification of the same hand) can also be achieved. In addition, our software provides codicological meta-data about each page including the number of lines, the size of the characters, the density of writing, and other structural information.

We are also experimenting with automatically aligning transcriptions with the words in the images, and are encouraged by the initial results. In the coming months, we will extend our efforts to additional collections, including the Dunhuang Tibetan material, and expect to report on the outcome, as well.