Publishing Text for a Digital Age
Tufts University, March 2014
Workshop Proposal

Computational Processing of Toponymic Data from Classical Arabic Sources

Cameron Jackson,[1] Maxim Romanov[2]

cameron.jackson@tufts.edu, maxim.romanov@tufts.edu

Corpus: Shamela Digital Corpus @ shamela.ws
Problem: Determining the Geographical Coverage of an Arabic Source
Solution: Computational Identification of Toponymic Data & Automatic Mapping

Particular toponymic data, both toponyms proper and toponymic descriptive names that affiliate individuals with specific places, may imply the geographical extent of a given author's network (NB: in this case, "network" is defined rather broadly, including the actual scholarly network [direct connections], as well as that of influences [indirect connections]). Mapping particular sources we are likely to clearly see both the center and the periphery of given authors' networks. Looking into the geographies of dozens, hundreds, or even thousands of classical Arabic sources, on the other hand, we are most likely to get a valuable perspective on the development of the Islamic learned community and written culture in general. Looking at these geographies in historical perspective should allow us to see how the centers of Islamic learning were changing over time. Keeping in mind that the available digital corpus of classical Arabic sources amounts almost to 6,000 titles (spanning chronologically from the late 7th to the early 20th centuries), the cumulative chrono-geography of the classical Arabic corpus should allow us, in the long run, to identify geographical extents of major learned networks of the Islamic world.

Considering the volume of the corpus, manual tagging of toponymic data is not a feasible solution, thus the main goal is to develop a satisfactory computational solution for identifying toponymic data in raw texts and visualizing them. In general, toponymic data in Arabic texts can be identified using toponymic `ngrams` and be extracted with text-mining scripts (written in `Python`). I (M.R.) used this approach to identify toponyms in raw text for my dissertation research and the overall approach appears to be quite efficient. However, the algorithm and its procedural implementation do require improvement and optimization. One of the major problems is the disambiguation of place names that may refer to either different locations or to commons words.

[1] Tufts University, Junior (Arabic (BA)/Computer Science).
[2] Tufts University, Postdoctoral Associate (Department of Classics/Perseus Project).

We will attempt to solve these issues algorithmically. Computationally identified toponymic data are to be mapped using a gazetteer of the Classical Islamic World, which is being created as a part of this project. Visualizations are to be produced for online viewing (for example, in the `KML`-format for Google Earth) and high-quality printing (high-resolution `TIF`/`PNG` files generated in `R`). The end goal is to develop a tool that will allow generating visualizations of toponymic data from any submitted Arabic text. Overall, this project should be viewed as an essential building block in the development of the method for the computational reading of the written Classical Arabic legacy in general.

Now including a collaborator, this work is a direct continuation of research that became possible through my (M.R.) participation in "Working with text in a digital age," the previous summer institute at Tufts University that took place in 2012.