A Digital Humanities for Premodern Islamic History

Maxim Romanov
Department of History, University of Vienna, Vienna, Austria; e-mail: maxim.romanov@univie.ac.at

doi:10.1017/S0020743817001015

Defining digital humanities is tricky. Our scholarship has been intrinsically digital for quite a few decades already, as we rely more and more on electronic storage to save, word processors to write, bibliography managers to organize, databases to consult, digital libraries to search and read. Living in the digital world, however, does not make us all digital humanists—if these digital entities are taken away, we will have their analog prototypes to fall back on, and beyond a certain level of inconvenience, this will not affect the way most of us do our scholarship. The transition to digital humanities must begin somewhere at the point where our humanistic inquiry starts to rely on the machine as the matter of methodological exigency.[1]

In some ways digital humanities is a "no man's land" that, within every national context, is most successfully claimed by scholars of national histories, literatures, and languages—by virtue of their higher numbers and the accessibility of their subjects to national funding agencies and the wider public. In practical terms, one's primary field of academic inquiry, with its specific research questions and available source base, determines the set of computational approaches and thus defines a specific instance of digital humanities. (Without a primary field of academic inquiry we would be talking about technicians rather than scholars.) For example, although methods for analysis of video and audio recordings will be of little practical value to a scholar of premodern Islamic history, there is a lot to be gained from methodological areas such as computer vision,[2] social network analysis, geographical information systems (GIS), and, most importantly, text analysis.

To build a case for text analysis methods, let's consider the example of *Ta'rikh al-Islam* (History of Islam) by al-Dhahabi (d. 1348). This book of great length and scope, whose fifty volumes contain 3.6 million words—the size of *War and Peace* six times over—traces the first 700 years of Islamic history through description of historical events and some 30,000 biographies.[3] Although a great number of modern scholars use this massive "obituary chronicle" as their major source, we lack a deep understanding of its inner organization. With a significant amount of Arabic historical texts available, we can employ the text-reuse identification method developed by David Smith of Northeastern University to build the equivalent of an x-ray image of this chronicle, providing new information about it and raising a series of important historiographical questions.

For example, we gain a detailed perspective on the sources that al-Dhahabi might have used: he mentions some forty of them, and in our x-ray we find traces of these (provided, of course, that we have a relevant text in our corpus, and in most cases we do) and other sources that he may have failed to mention for some reason. Figure 1 shows how passages common to al-Bayhaqi's (d. 1066) *Dala'il al-Nubuwwa* (Indications of Prophecy) also feature in al-Dhahabi's text. The length of a black line corresponds to the number of words in an identified text reuse instance; the dense black block in the beginning of the book indicates the density of text reuse and tells us that all

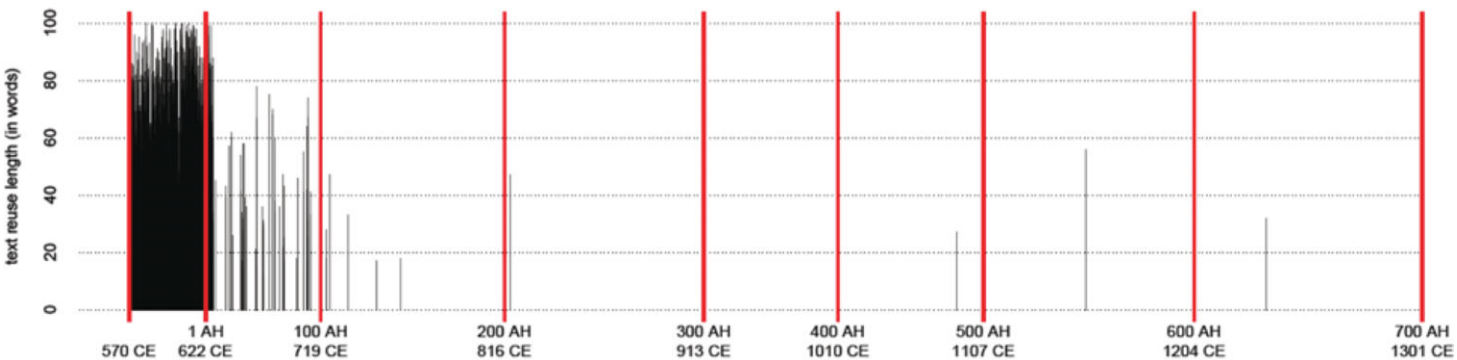

FIGURE 1. (Color online) Common passages in al-Bayhaqi's (d. 1055) *Dala'il al-Nubuwwa* featured in al-Dhahabi's *Ta'rikh al-Islam*: 111,436 words, 371 pages, 50 percent of instances 28–61 words. The graph shows the flow of the text of al-Dhahabi's work in one hundred-word chunks: the beginning of the book is on the left, the end of the book is on the right; the red lines indicate points where al-Dhahabi moves on to cover the next hijri century. Black lines indicate instances of text reuse traceable to al-Bayhaqi's text, and the length of a black line corresponds to the number of words in an identified text reuse instance. The dense black block in the beginning of the book indicates the density of text reuse—with most of it falling on the period up until 640 CE—which means that all these common passages occur exactly where we would expect them to appear: in the part of al-Dhahabi's text that deals with the period of the Prophet's life.

of these common passages occur exactly where we would expect them—in the part of al-Dhahabi's text that deals with the Prophet's life.

With the help of our x-ray, not only do we discover connections with practically all the sources that al-Dhahabi mentions in his introduction, we are also able to gauge the extent to which he engaged with his sources. We find an equivalent of over 800 pages (300 words per page, 246,000 words) of common passages with Ibn ʿAsakir's *Ta'rikh Dimashq* (History of Damascus), some 370 pages with al-Bayhaqi's *Dala'il al-Nubuwwa* (Indications of Prophecy), some 280 pages with Ibn al-Jawzi's *Kitab al-Muntazam* (The Book of Rightly Ordered Things [about Histories of Kings]), some 270 pages with al-Mizzi's *Tahdhib al-Kamal* (Refinement of Perfection), some 250 pages with al-Khatib al-Baghdadi's *Ta'rikh Baghdad* (History of Baghdad), and so on. (Keep in mind, however, that these numbers cannot simply be added up because there is a significant amount of reuse among these texts as well.) In all cases 50 percent of identified shared passages are twenty-five to sixty words long! It is not surprising to see major biographical collections and chronicles on this list, but al-Bayhaqi's *Dala'il al-Nubuwwa* seems to stand out. Our text reuse data suggests that this work is the most heavily reused text—its 370 common "pages" amount to almost 20 percent of its volume (the share of *Ta'rikh Dimashq*, on the other hand, is barely 2.4% of its volume). This does not necessarily mean that al-Dhahabi took all passages common to al-Bayhaqi directly from him—there is always the possibility of a common source or a source between *Dala'il al-Nubuwwa* and *Ta'rikh al-Islam.* (In this particular case, this is quite likely because al-Dhahabi lists *Dala'il al-Nubuwwa* among his main sources.) However, it does mean that our distant reading suggested to us a very interesting connection that deserves further close examination using more traditional methods.

Next, we may attempt to assess the cumulative level of text reuse in al-Dhahabi's *Ta'rikh al-Islam*. Altogether, the currently identifiable amount of text reuse—counting each instance of text reuse only once, even if it is traceable to multiple sources—amounts to *at least* 23 percent of al-Dhahabi's text (750,000 words, 2,500 pages, with 50% of quotations within twenty-five to fifty-nine words). If we look at this text century by century, we discover that for almost every century that he covered, about 20 to 22 percent of his text can be traced to passages from his sources, with the exception of the 1st and the 7th Islamic centuries, where the share of text reuse amounts to 47.8 percent and 8.4 percent, respectively. These numbers confirm, first, that this text is a compilation, and second, that it is the latest material that is least derivative. While there is a tendency to dismiss such "discoveries" as "nothing that scholars don't already know," it is important to stress that they transform "intuitive knowledge" into knowledge backed by a significant amount of textual evidence, which we can then use as a reliable premise to further advance our analysis—something that otherwise would not be possible.

The importance of this seemingly trivial discovery is that it tells us that al-Dhahabi, by quoting his sources so extensively, effectively preserves their *archaic* language. For instance, when he writes about the 1st Islamic century, his narrative is dominated by quotations from texts written in the 3rd Islamic century; when about the 2nd, from the 3rd and the 4th; and when about the 3rd, from the 3rd, 4th, and 5th, and so on. The discovery of these archeological layers of language indicates that al-Dhahabi describes

people and events with a language that is as close to contemporaneous as is feasibly possible in historiographical terms. (It is also likely that his own syntax and word choices were influenced by the language of his sources.)

A rolling stylometry test[4] of *Ta'rikh al-Islam* further shows that al-Dhahabi's writing "style"—defined as a set of the most frequent function words that form a writer's "fingerprint"—changes completely by the end of the book: three samples of 10,000 words were taken from the beginning (red), middle (green), and end (blue) of the book and used to test the extent to which the style of these samples is similar to the rest of the book. Figure 2 shows that the "early style" (red), which dominates the language of the 1st Islamic century, disappears completely by the end of the 3rd century, not reaching even the middle of the book. In stylometric terms this can be interpreted to mean that the beginning and end of the book were written by two different people.

This discovery about the language of al-Dhahabi's *Ta'rikh al-Islam* has far-reaching implications for the data that he collected. For example, although the text comes from the 14th century and inevitably suffers from 14th-century biases when it comes to the representation of the past, at the linguistic level the description of people and events is not as anachronistic as one would think, and, arguably, these properties of the language allow us to use the data from this text for modeling historical processes.[5]

I mentioned earlier that the overall volume of text reuse in al-Dhahabi's *Ta'rikh al-Islam* amounts to at least 23 percent. Our initial text reuse experiment was constrained by the format of our texts—or, to refer to the article in this roundtable, by the lack of a proper scholarly corpus. As a result, we had to compare texts that were mechanically chunked into slices of one hundred words, and with such comparison we could have missed up to 20 percent of reused text. This circumstance also prevented us from performing a more informative distant reading. As our OpenITI corpus develops and texts are supplied with logical markup (i.e., each of their chapters, sections, and subsections are explicitly tagged), we will be able to run more precise and robust experiments. Comparing logical units of texts—for example, a biography with another biography—would open more opportunities for understanding how our texts were composed. For example, we would be able to identify which biographies al-Dhahabi included from any given source and which he omitted. Knowing this would allow us to assess—on the largest scale possible—not only his selection criteria, but also what he suppressed from select biographies and how he modified them.[6] Pushing the point further, this can be accomplished for all historical titles in the OpenITI corpus, which is likely to significantly change our understanding of the Islamic historiographical tradition.

With all this said, the machine will never replace traditional training. No proper distant reading experiment can be designed without a deep understanding of the subject in question, which can only come from close reading. The machine is just another tool in our methodological toolbox that allows us to do something that other methods do not. The machine will never ask novel historical questions, but it will enable us to do so.

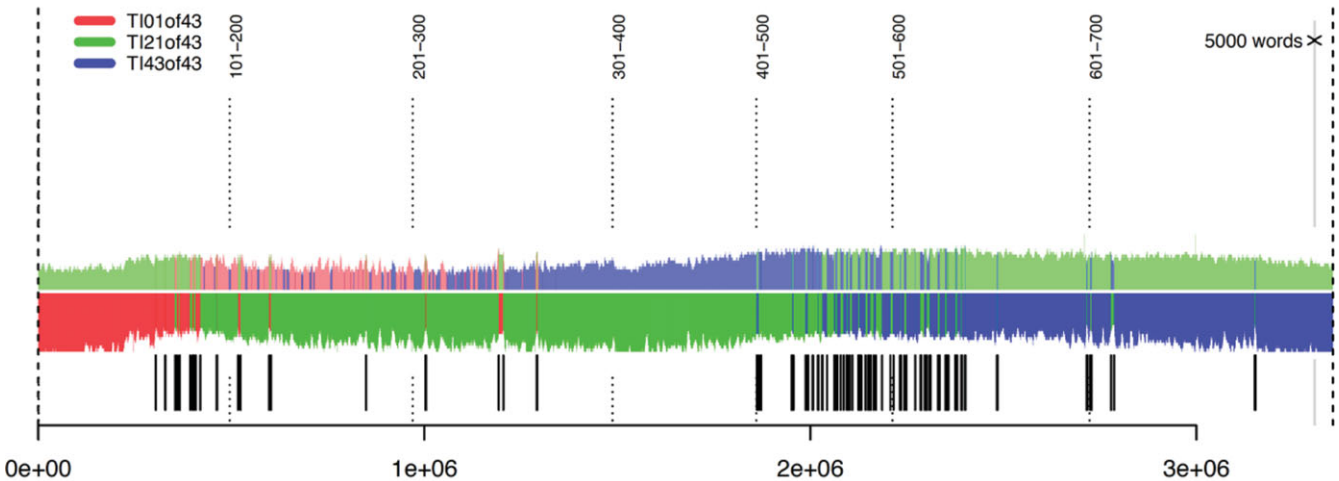

FIGURE 2. (Color online) Results of the rolling stylometry test. Three samples of 10,000 words were taken from the beginning (red), middle (green), and end (blue) of the book and used to test to what extent the "style" of these samples is similar to the rest of the book. The graph shows that the "early style" (red), which dominates the language of the 1st Islamic century, disappears completely by the end of the 3rd Islamic century, not reaching even the middle of the book. The style in the end of the book is completely different from that of the beginning of the book.

NOTES

[1]Our scholarship has been intrinsically digital for quite a few decades already, as we began to rely more and more on electronic storage to save, word processors to write, bibliography managers to organize, databases to consult, digital libraries to search and read. But if these digital entities are lost, we will have their analog prototypes to fall back on, and beyond a certain level of inconvenience, this will not affect the way most of us do our scholarship.

[2]We already have methods to make manuscripts searchable (in a limited way) and soon we'll be able to group manuscripts according to handwriting as well as to identify manuscripts written by the same hand. See, for example, Mike Kestemont and Dominique Stutzmann, "Script Identification in Medieval Latin Manuscripts Using Convolutional Neural Networks," *Digital Humanities 2017: Book of Abstracts* (Montreal, 10 August 2017), 283–85.

[3]Al-Dhahabi, *Ta'rikh al-Islam*, ed. ʿUmar ʿAbd al-Salam Tadmuri, 1st ed., 52 vols. (Beirut: Dar al-Kitab al-ʿArabi, 1990–99).

[4]See the website of the Computational Stylistic Group, accessed 6 October 2017, https://sites.google.com/site/computationalstylistics/projects/testing-rolling-stylometry; and Maciej Eder, Jan Rybicki, and Mike Kestemont, "Stylometry with R: A Package for Computational Text Analysis," *R Journal* 8 (2016): 107–21.

[5]See Maxim Romanov, "Toward Abstract Models for Islamic History," in *The Digital Humanities + Islamic Middle Eastern Studies*, ed. Elias Muhanna (Berlin: De Gruyter, 2016), 117–49; and Romanov, "Algorithmic Analysis of Medieval Arabic Biographical Collections," *Speculum* 92/S1 (2017): 1–21.

[6]See Maxim Romanov, "Observations of a Medieval Quantitative Historian?," *Der Islam* 94 (2017): 462–95.