The history of Arabic books in the digital age

Sarah Bowen Savant reveals how computer algorithms can aid the comparison of medieval Arabic texts

Sarah Bowen Savant is an Associate Professor in the Aga Khan University's Institute for the Study of Muslim Civilisations, which is based in London. In 2015 she received a British Academy Rising Star Engagement Award.

In September 2015, a British Academy-funded three-day workshop in London brought together computer scientists, programmers and scholars to explore the ways that a new digital method can detect the copying of older books into newly created ones. The method is useful for seeing how authors reworked and adapted past texts for new purposes.

In the 7th century, the rise of Islam and the Arab conquests initiated a period of literary creativity at urban centres across the Middle East, as new wealth combined with new faith. By the 10th century, the result was one of the largest textual traditions up to its day, which addressed all aspects of culture in Arabic, the new common language for elites. Works on religion fuelled this creative outpouring, as did many other books, reflecting the subjects of interest to Arabic-reading elites, including philosophy, literature and poetry, as well as medicine, mathematics and geography. Going by accounts from the period and also what survives, individual writers wrote dozens of wordy works. For example, the *History* written by a famous Iranian author named al-Tabari (died Baghdad, 923) covers history from the creation of the world up to his own day and runs to over 1.5 million words, filling 39 volumes in the modern English translation – and this was apparently an abridgment of an earlier version (by comparison Herodotus's Greek *The Histories* has 190,000 words). Furthermore, that is just one of al-Tabari's many books, and he is but one of thousands of authors from the pre-modern period.

What inspired this surge and what kept it going? Why, when Europe and other regions of the world produced books on a much more modest scale, did Arabic writers become so prolific? How did they produce such vast works? The introduction of paper-making technology in the 8th and 9th centuries and the technological revolution it implied can only explain so much. What core practices and ways of thinking about authorship, books, and book production arose in the Middle East to draw new writers into the field and to spur them on? So far, scholars have made only slight headway to explain the driving factors of this new, massively creative tradition.

Detecting book copying and how the technology works

There is now a major clue that can for the first time be probed and this is evidence for verbatim copying gathered electronically. An important point, frequently made but little explored, is that the Arabic textual tradition is not only very large, but very repetitive. About four years ago, I was checking student papers for plagiarism using software purchased by my university, while also writing a book about Islamic history using online collections of Arabic books. For my research, I was particularly interested in the ways that Iranians' memory of their history changed as they converted to Islam between the 9th and 11th centuries. I was trying to work this out by comparing earlier and later accounts of the same matters, and found that many reports about the origins of Islam fine-tuned earlier ones, changing only a name or other choice details so as to give Iranians themselves more of a starring role in events. Frequently, I could find as many as a dozen varying accounts of the same event, as if a game of 'Chinese whispers' had been played out over six or more centuries. Student papers in hand, the idea dawned on me: what if I could use plagiarism detection software to gather evidence for copying in the Arabic tradition, the networks of people who participated in it, and how the tradition developed as a whole over time?

The Arabic textual tradition is not only very large, but very repetitive.

Unfortunately, no software existed that I could apply to my Arabic texts. So after some preparatory work, in 2015 I assembled a team of programming volunteers who decided to try an algorithm developed in the US that detects 'text reuse', much as plagiarism software does. The algorithm's author, David Smith of Northeastern University, had developed it for 19th-century American newspapers, to see how stories went 'viral'. Similar algorithms are employed by search engines to eliminate duplicate results and by security analysts combing the web. KITAB's team[1] – working weekends and nights, including a weekly two-hour Sunday call – applied the algorithm to about 10,000 Arabic books that have been digitally formatted and made available online across the Middle East. We compared every book against every other one using a 'shingling' method, in which the computer searches for similarities between texts broken into units. This comparison generated nearly 10 million files filling 250 gigabytes of data. The algorithm was run on the texts on the supercomputing centre at Tufts University in Boston, thanks to the support of our project partner, the Perseus Digital Library. KITAB's team then designed a database at my university of indexed results and a web application for studying them. We are in the process of loading our files into this database (which, given the size of the data, requires quite a lot of data modelling), but now a sampling of our pilot data can be seen on the KITAB website.[2]

From a research point of view, the complexity of these operations and the enormous quantity of data generated means we need to think carefully about what we would like to find within our data. Furthermore, the present algorithm is designed to discover statistically significant relationships between texts, but there is no one-size-fits-all algorithm to detect text reuse perfectly – it is very different to look for and visualise rare similarities versus rare differences between texts and across our collection. As we detect and graphically display other types of text reuse, we will adapt our algorithm and data models further, and generate still more data for exploration.

The British Academy funding was absolutely critical to our work on the pilot, as it provided a concrete goal: to generate data for analysis by a group of peer historians at our British Academy-funded workshop in September 2015. We met this goal, and KITAB's findings are now beginning to be cited by members of the user group in their own research. The workshop also featured a day and a half initiation into text reuse algorithms themselves, offered by Marco Büchler, of Göttingen University. This was helpful for communicating both the promise and challenges of the method to the field.

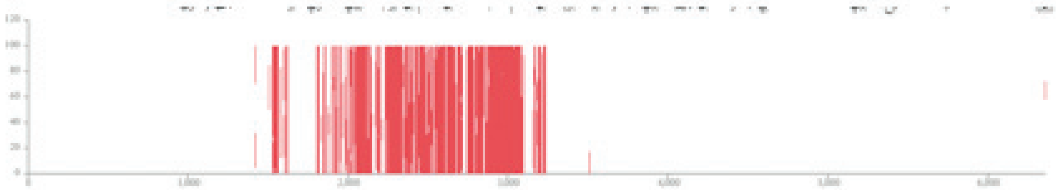
Copying as a creative practice

What we have found is that the scale of copying was absolutely enormous. Just to take one straight-forward example: different accounts about the Prophet Muhammad were written down and then copied repeatedly in the first four centuries of Islamic history. Some accounts eventually won out as the most authoritative, but that was a long time coming. When authors wrote about Muhammad, they went back to previous circulating accounts, copying out the parts most relevant to their own versions. This writing was fuelled and supported by auditory practices, including reading texts aloud. The graph in Figure 1 was generated from our data, with the red lines showing passages common to both the *History* of al-Tabari and an earlier biography of Muhammad. The *x*-axis represents

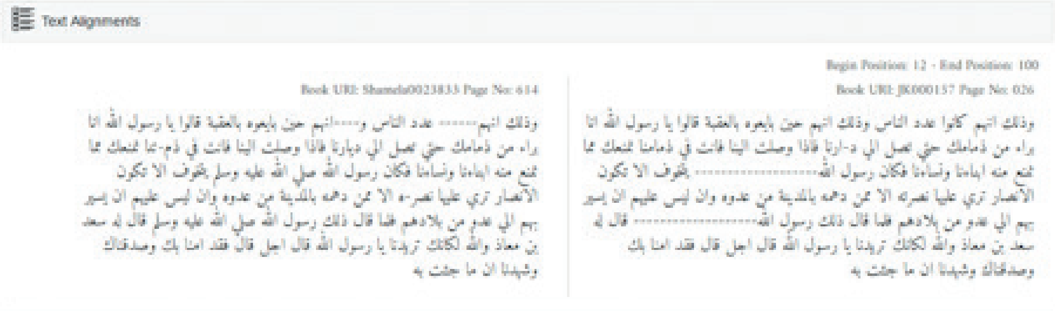
[1] The volunteer team of KITAB (an acronym that means 'book' in Arabic and stands for 'Knowledge, Information Technology, and the Arabic Book') includes: Maxim Romanov (also a historian, of the Perseus Digital Library), Ahmad Sakhi and Malik Merchant (who volunteer their time as part of a programme run on behalf of the Aga Khan), and Sohail Merchant (of the Institute for Ismaili Studies). Romanov has played the leading role in forming the corpus that serves as the basis for our research.

[2] www.kitab-project.org (to search the corpus: ID = kitab-user; password = Kitab321; the K is uppercase).

Figure 1. Comparing works by al-Tabari and Ibn Hisham.



(a) Detecting common passages. This graph may look like a bar code, but in fact shows what of al-Tabari's *History* also turns up in an earlier book, a biography of Muhammad by Ibn Hisham (died 828). The numbers on the *y*-axis represent numbers of words from this second book. Each line represents up to 100 words; hanging lines mean that the length of an identified passage is, for example, 60 words; it starts on the 20th word in the chunk and extends to the 80th word.



Book URI: Shamela0023833 Page No: 614

وذلك انهم------ عدد الناس و-----انهم حين بايعوه بالعقبة قالوا يا رسول الله انا
راء من ذمامك حتى تصل الى ديارنا فاذا وصلت الينا فانت في ذم-تنا نمنعك مما
نمنع منه ابناءنا ونساءنا فكان رسول الله صلى الله عليه وسلم يتخوف الا تكون
الانصار ترى عليها نصرت-ه الا ممن دهمه بالمدينة من عدوه وان ليس عليهم ان يسير
بهم الى عدو من بلادهم فلما قال ذلك رسول الله صلى الله عليه وسلم قال له سعد
بن معاذ والله لكانك تريدنا يا رسول الله قال اجل قال فقد امنا بك وصدقناك
وشهدنا ان ما جئت به

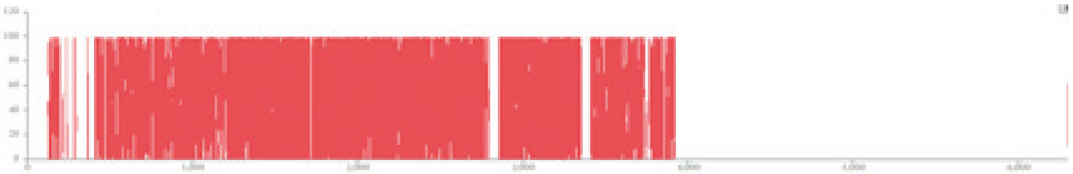
Begin Position: 12 - End Position: 100

Book URI: JK000157 Page No: 026

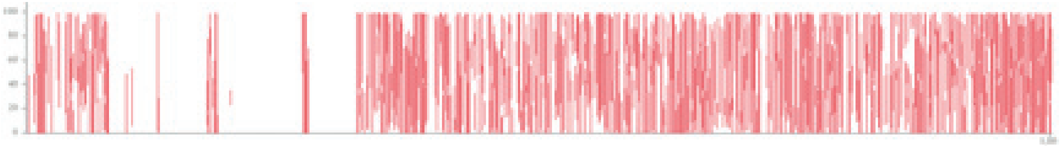
وذلك انهم كانوا عدد الناس وذلك انهم حين بايعوه بالعقبة قالوا يا رسول الله انا
براء من ذمامك حتى تصل الى د-ارنا فاذا وصلت الينا فانت في ذمامنا نمنعك مما
نمنع منه ابناءنا ونساءنا فكان رسول الله-------------------- يتخوف الا تكون
الانصار ترى عليها نصرته الا ممن دهمه بالمدينة من عدوه وان ليس عليهم ان يسير
بهم الى عدو من بلادهم فلما قال ذلك رسول الله-------------------- قال له
سعد بن معاذ والله لكانك تريدنا يا رسول الله قال اجل قال فقد امنا بك
وصدقناك وشهدنا ان ما جئت به

(b) Aligning texts. The passage on the left is from Ibn Hisham's book, whereas that on the right is from al-Tabari's. They treat a moment in the events leading up to the Battle of Badr, when Muhammad's newly established community in Medina had its first significant success against his opponents in his hometown of Mecca. The dotted lines indicate discrepancies in wording between the two texts, which in this case are not significant for the meaning of either text.

Figure 2. Comparing works by Miskawayh and al-Tabari.



(a) Two closely related books. Miskawayh chronology – divided into chunks – runs along the *x*-axis, with the red lines representing 100-word chunks of the text that also appear in al-Tabari's *History*.



(b) Closely related but not the same. This is a closer look at the start of Miskawayh's chronology. The red lines indicate passages common to al-Tabari's *History*, whereas the white indicates that the computer has not detected common passages.

chunked segments of al-Tabari's book, running from the beginning of the book to its end. The red lines running up the *y*-axis represent 100 word segments that are also found in the earlier history. The copied passages start with an account of a Christian king of South Arabia in the middle of the 6th century by the name of Abraha, who Muslim tradition records as leading an expedition against Mecca in the year of Muhammad's birth (*c*.570). They end with the death and burial of Muhammad. As a source for his knowledge, al-Tabari cites in his book an earlier book by Ibn Ishaq (died 767) that was transmitted through intermediaries. Ibn Hisham's book is a revised edition of that of Ibn Ishaq, so what we have in Ibn Ishaq is a common source for both al-Tabari's book and that of Ibn Hisham (Ibn Ishaq's book does not survive independently).

The results of the pilot stage turn up quite extensive and close copying, suggesting a vibrant, if complicated, written tradition. Such wholesale repurposing of an earlier text is in fact extremely common, and it is really just a matter of time – and statistically grounded research – before we can work out how much of the surviving tradition, as a whole, is repetitive.

To say that authors copied out parts of previous works – even very large parts – is not to say that they parroted past knowledge. Rather, they would appear to have viewed earlier books as valuable resources whose contents could be reused, sometimes giving credit, but other times not. We need to understand the assumptions guiding this copying much better. We already know that it occurred in many ways, including excerpts from earlier works, as with al-Tabari's copying of Ibn Ishaq's book, as well as the recycling of textual fragments in anthologies, encyclopaedias, multi-text compilations, commentaries, and abridgements and extensions of earlier books. Did citation practices vary depending on the copied source and/or the book into which it was copied? This seems likely, but in precisely what ways and why? Were Arabic authors more likely than their European and other counterparts to adapt past works than to copy manuscripts in their entirety? Does such copying suggest the evidence of previously unrecognised Arabic literary canons or classics?

For now, we can already see how selective this copying was – sometimes in fascinating ways. Consider how al-Tabari's own book was copied in later centuries, including by a court secretary and librarian named Miskawayh (b. *c*.932, d. 1030), who studied al-Tabari's works under the direction of one of the past master's pupils. A famous Orientalist, David Samuel Margoliouth,[3] completed in 1921 an English translation of the last two volumes of Miskawayh's universal history – at the time, he asserted that Miskawayh copied al-Ṭabari's *History* for his own book up to 908. Since then – that is, for nearly a hundred years – European scholars have largely ignored Miskawayh's book for anything but history from 908 onwards and have viewed it as only original – and valuable – as a source for the Buyid dynastic period in which he worked. Until the turn of the 21st century, in fact, there was not even a completed printed edition to be had.

Both works are chronicles, running from creation up to their own day. In Figure 2, the red lines represent what of al-Tabari's book can be found in Miskawayh's. Read as a whole, the graph shows that Miskawayh relied very heavily upon al-Tabari's *History* – but he did not copy it verbatim. In particular, examining the white parts of the graph now, rather than the red, reveals that when Miskawayh wrote about the history of the Sasanians, the last dynasty of Iran before the Muslim conquests, he copied very little, and likewise, he copied sparingly from al-Tabari's account of Muhammad. By contrast, he relied heavily on al-Tabari when he turned his attention to the history of the caliphate (but even there, he made choices). It would seem that these were parts of his book that he felt needed more editorial work and intervention – these are where we can expect to find his own handiwork. For research purposes and understanding the authorial interventions of Miskawayh, it is very helpful to be able to see the distinction between the white and the red, and then to closely read and consider the parts of Miskawayh's book corresponding to the white.

Future development of KITAB

We have produced a lot of data now, and we need time to examine it carefully and consider its significance. We are applying for further funding to do so, and also for funding to improve the quality of the bibliographic data in our collection of digital books.

As for the technology side, as a result of the pilot, we now have a much better idea of the technology underlying KITAB and the research questions that can realistically be addressed. We can filter our data to show the highest correlation between one book and any other in our corpus. What might we do in the future? We hope to allow users to explore copying and transmission in the tradition in many different ways, for example, to investigate complex texts, built out of earlier texts, to see how diverse elements were combined and amalgamated (as occurred, for example, with anthologies), or to see how single, specific works (such as al-Tabari's *History*) moved across many other later works. We intend to create visual aids that allow users to detect slight differences between two or more texts, as one might wish to do when comparing two editions of the same book, or a book and its abridgements or extensions. We also hope users will also be able to collect common passages scattered across the corpus, and order them using bibliographic data pertaining, for example, to place and date of composition (as would have been useful for researching my first book). Each of these ideas arose through the process of the pilot and is inspiring the next phase of development. This research will undoubtedly show that copying had many different cultural meanings. It is this big picture that drives the work of KITAB. ▬

[3] D.S. Margoliouth was elected a Fellow of the British Academy in 1915.