Reading and understanding text reuse data through graphs: a case study on pre-modern Arabic book history and historiography

Mathew Barber[1]

[1]KITAB team, Aga Khan University, Institute for the Study of Muslim Civilisations (AKU-ISMC)

1 Pre-modern Arabic and text reuse

The pre-modern Arabic textual tradition is large and astoundingly complex. Text reuse, moreover, is a prominent feature of the Arabic written tradition. As has long been observed, many Arabic texts are compositions made up of multiple earlier texts (which are sometimes cited). This is particularly true of the historiographical tradition. Al-Tabari's (d. 923) Taʾrikh ('History'), one of the first and most famous works of Arabic history is a compilation of multiple earlier sources (potentially received orally) that are typically meticulously cited. This text became the template for later works of history, like Ibn al-Athir's (d. 1233) famous chronicle the Kamil. The Kamil in turn was used extensively in the historical sections of al-Nuwayri's (d. 1333) encyclopaedia, the Nihaya.

Already with the above historiographical stemma, we are presented with a complex (but crucial) problem: when there is evidence that al-Nuwayri is reusing al-Tabari's history, is he using it through the Kamil or is he accessing the text directly? This kind of question would need to be addressed at both the macro and micro level. That is, firstly where do reuse instances occur across the text, are reuse instances from one text or the other more extensive? Secondly, what kind of language is al-Nuwayri using in individual passages shared with both the Tarikh and Kamil, and how does it differ from each of these texts.[1] A number of important questions can be resolved through a detailed study of text reuse in Arabic texts. It can be used to understand teaching and transmission over time (oral vs. written), compositional techniques (such as direct quotation, excerpting and paraphrase), and citation practices (when and how authors attribute their sources). Most importantly for this paper, it can be used as evidence for the recovery of lost texts and traditions (for a recent application of text reuse to this end, see: Bora (2021)).

To thoroughly answer such questions we need to work at scale. Al-Tabari's Tarikh is nearly 1.5 million tokens long, Ibn al-Athir's text is 1.3 million and al-Nuwayri's is 2.3 million. These three works are quite exceptional, but they are not alone. Of

[1] This specific case study will be explored by Sarah Bowen Savant in chapter 6 of her forthcoming monograph.

OpenITI's digital corpus (Nigst et al., 2021) of over 6000 unique Arabic books, just over 100 exceed 1 million word tokens in length. The OpenITI corpus is, moreover, growing and with advances in OCR and HTR it is likely to expand to include many thousands more works. It is best, therefore, to study text reuse on this corpus using computational methods.

2 The computational study of text reuse: problems of scale

This paper will examine uses for the text reuse cluster data produced by the KITAB team from the OpenITI corpus. The KITAB project utilises an algorithm called passim for this purpose. Passim was initially developed by David Smith and his team for the study of nineteenth-century American newsprint and the phenomenon of reprinting and viral texts (Smith et al. (2015), Xu et al. (2014)). This was then adapted by the KITAB team to suit the larger Arabic texts that make up their corpus. First the corpus is pre-processed - every text is split into 300-word chunks. Then passim compares every 300-word chunk to every other 300-word chunk in the corpus. If the algorithm finds an instance of reuse (according to a set of predefined parameters), it records the location of the reuse in both chunks and uses the Smith-Waterman sequencing algorithm to create an alignment of the text reuse instance.

The main result of running passim on the OpenITI corpus is a set of csv files corresponding to pairs of books, where each row of the csv gives an instance of text reuse and its location in each text. These pairwise alignments are useful for understanding reuse between two Arabic works, but they do not provide a clear understanding of text reuse across the entire corpus. This is better understood through the clustering of the aligned pairs, which passim performs as part of its normal program. This process brings together the different aligned pairs that share text. For example, we might have two aligned pairs for chunks A, B and C, where chunk A and B are aligned and A and C are also aligned. This might suggest that there is also text shared between B and C. If passim finds that A, B and C all more-or-less share the same text, then all three chunks would be brought together into one cluster. The resulting output is a list of passages, classified into clusters. These clusters can range from containing two passages (that is, just one pairwise alignment) to hundreds of passages spread across multiple works.

Smith et al.'s research into nineteenth-century newspapers has been primarily concerned with this cluster output, as it has allowed them to explore the circulation and propagation of viral excerpts, particularly vignettes. For example, it was common for certain maxims or moral stories to circulate almost verbatim across newspapers in a variety of US states and across long periods of time. The cluster data allows one to understand which kinds of excerpts were popular, when they began to circulate and how long they circulated for (for a number explorations, see the team's book in progress Blankenship, Cordell, Fitzgerald, Mullen, and Smith (Blankenship et al.), especially the chapter 'Classifying Vignettes, Modeling Hybridity'). This kind of research primarily involves studying the large clusters. If a cluster contains a large number of passages, then this suggests it was a popular passage. Through the addition of metadata (such as newspaper publication dates and locations) one can easily trace how one cluster of passages circulated across time and space.

Of course, the pre-modern Arabic written tradition varies enormously from nineteenth century newspapers. Certainly, particular stories circulated and were repeated verbatim across time and space (particularly snippets of historical information). How-

ever, the most prominent circulating material in the Arabic tradition (and certainly in the OpenITI corpus) are Hadith. These are sayings attributed to the Prophet Muhammad, often used in the formulation of Islamic law . Much ink was spilled qualifying or disqualifying the Hadith on the basis of their chains of transmission, and they were constantly compiled and recompiled into new collections and commentaries. In the OpenITI text reuse and cluster data, passages corresponding to Hadith are phenomenally common.

In part because of this, the OpenITI corpus text reuse data are enormous. A passim run produces over 1.5 million book pairs. Although many of these books pairs have only a few aligned passages between them, there are over 2 trillion aligned passages in the OpenITI corpus as a whole (for only 6000 texts). Many of these passages form part of large clusters. In the cluster data, there are over 3.5 million clusters of which 346 contain more than 100 passages.

Students of pre-modern Arabic are not just interested in the large clusters. Even small clusters could indicate how a group of authors had access to a shared source tradition or formed a textual community. Many modern researchers are interested in how small pieces of text propagated across multiple texts across time. Network visualisations are an incredibly valuable (especially for those who do not specialise in digital methods) for understanding and exploring text reuse data. The data is, however, much too large to handle in its raw form (contrast this with Ryan Cordell's informative use of Social Network Analysis for the 19th-century text reuse data set, Cordell (2015)). If we were to produce a network graph from the entirety of the cluster data, the result would be a mesh of unreadable nodes and edges. The data instead needs to be filtered and sliced before the network visualisation is produced. As every researcher approaches this data with a distinct question, such filtering choices should be left to the individual scholar.

3 A case study: the lost (or dispersed) Fatimid tradition

The Fatimids were a Shi'i dynasty that ruled from North Africa from 909, before then moving to Egypt in 969 and ruling there until their fall in 1171. Despite this being an important period of North African and Egyptian history, very few Fatimid-era texts survive in the modern day, especially from the Egyptian period of their rule (in part, perhaps, because of anti-Shi'i biases in later periods). Fatimid texts were, however, available to authors in the past and we find excerpts of Fatimid texts quoted by Egyptian authors (particularly historians) up until the end of fifteenth century.

For a long time, modern scholars have used these later Egyptian texts as windows into the Fatimid past and into Fatimid-era sources. There are, however, a number of challenges for accessing Fatimid-era sources through later texts. Firstly, they are very rarely cited by authors (and where there are citations they are often vague and unclear - referring only to a title or an author that might match a number of potential source texts). Secondly, where authors do cite their sources, the boundaries of the quotations are often indistinct. Authors did not uses quotation markers and so we rely on language usage to determine where a quote ends, or where an author interjects before returning to the quotation. Thirdly, authors quoted these texts with varying degrees of precision. Sometimes they paraphrase extensively, other times they quote very exactly (for detailed examples of reuse of Fatimid sources by one author, al-Maqrizi (d. 1442), see Bauden (2010)). Given these problems, text reuse and particularly cluster data present an exciting avenue for exploring how multiple authors shared Fatimid

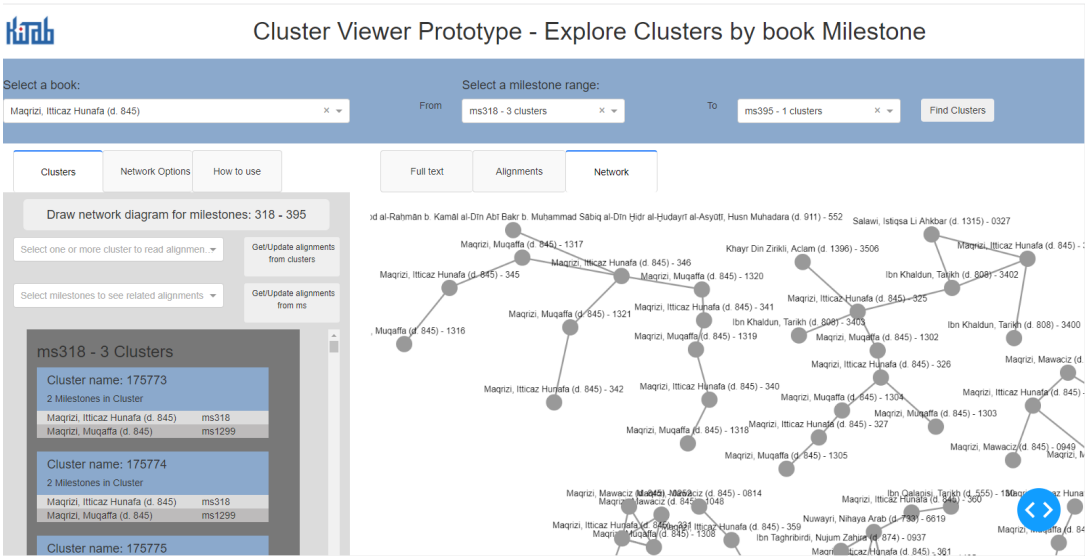

Figure 1: A snapshot from the prototype interactive viewer, showing a graph for clusters corresponding to around 30 years of al-Maqrizi's Itti'az. Note the extensive self-reuse (text shared between several of al-Maqrizi's works), which suggests he is using the same source across his works

source texts. Text reuse evidence cannot be used to reconstruct these earlier sources, but it can provide valuable insights into their content and agendas.

Al-Maqrizi's Itti'az is a chronicle of Fatimid history, which often contains some of the most detailed material on the Fatimids, including some quite unique materials. Even for this unique, relatively short text, passim has identified 941 clusters, of these 298 contain more than 2 texts (that is, they record more than a pairwise relationship) and 93 contain more than 4 texts. For a historian to meaningfully interpret this data, they need to be able to filter it, to see the bigger picture and to read the individual text reuse alignments.

One way in which to interrogate the data, is to view it (or subsets of it) as a graph. To explore the data for Fatimid-era sources, I have developed an application that allows users to create graphs by focusing on sections of a work and to read the alignments associated with each cluster (see figure 1). By filtering the data through this application the data is made more manageable.

For the graph, I have treated the 300-word chunks of each text as nodes, and then drawn an edge between nodes that belong to the same text reuse cluster. As a 300-word chunk might belong to more than one cluster (perhaps the first half of the chunk is clustered differently to the second half), using 300-word chunks as nodes allows one to explore the consecutiveness of reuse clusters. Text clusters that appear across multiple chunks in sequence might suggest, for example, that a cluster of texts is all reusing the same source in the same arrangement.

This approach to drawing graphs allows one to quickly see all the texts that belong to each cluster and how this compares to other sections of the same work. One can then select nodes within the graph and read the text of the corresponding alignments. Thus the graph can act as an exploratory tool, quickly showing connections between works and allowing one to explore further.

To further reduce the complexity and make the data more readable, it is possible to add filters. For example, to see only clusters in texts written before or after a certain

date, or to look at clusters than contain more than a certain number of texts.

As this paper will show with more specific examples, this interactive visualisation can become a key tool for understanding Egyptian history writing and the usage of Fatimid histories. This can of course be applied to a variety of other Arabic textual traditions. It, therefore, is a step towards reading these complex works as part of broader networks of shared text, which provide evidence of shared source usage. Networked text reuse data provides another way of reading, but it must be paired with interactivity to be meaningful, flexible and useful to the historian.

References

Bauden, F. (2010). Maqriziana XII. Evaluating the Sources for the Fatimid Period: Ibn al-Maʾmūn al-Baṭāʾiḥī's History and Its Use by al-Maqrīzī (with a Critical Edition of His Resumé for the Years 501-515 A.H.). In B. Craig (Ed.), Ismaili and Fatimid Studies in Honor of Paul E. Walker, Chicago Studies on the Middle East. Chicago: Middle East Documentation Center (MEDOC). Publisher: Middle East Documentation Center (MEDOC), The University of Chicago.

Blankenship, A., R. Cordell, J. Fitzgerald, A. Mullen, and D. A. Smith. Going the Rounds: Virality in Nineteenth-Century American Newspapers (Online Draft). Manifold.

Bora, F. (2021). Writing history in the medieval Islamic world: the value of chronicles as archives. London: I. B. Tauris. OCLC: 1243259743.

Cordell, R. (2015). Reprinting, Circulation, and the Network Author in Antebellum Newspapers. Blog.

Nigst, L., M. Romanov, S. B. Savant, M. Seydi, and P. Verkinderen (2021, October). OpenITI: a Machine-Readable Corpus of Islamicate Texts (Version 2021.2.5) [data set]. Type: dataset.

Smith, D. A., R. Cordell, and A. Mullen (2015). Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers. Blog.

Xu, S., D. Smith, A. Mullen, and R. Cordell (2014, June). Detecting and Evaluating Local Text Reuse in Social Networks. In Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media, Baltimore, Maryland, pp. 50–57. Association for Computational Linguistics.