It should be no surprise to any reader of this blog that the KITAB project is primarily interested in studying Arabic text reuse. A large number of posts here, including several from myself, are concerned with the text reuse dataset produced by [passim](https://kitab-project.org/methods/text-reuse) and what it can tell us about Arabic book history. But what if I told you that passim produced two datasets and that our work has so far relied on one? If you read our pages on passim, or our blogs, you will see that this work focuses on the study of pairs of texts. That is, how text is reused between two books; where text reuse occurs in those books; how text is rearranged, etc. This post will introduce another dataset, the cluster dataset. It will be part of a series of posts that will follow my ramble[[1]](#footnote-1) (more-or-less in real time) through the cluster data, giving some early insights into how it can be used to understand Arabic book history. This data set could be used to study a range of texts, but here I will take historiography as an example.

# At the trailhead

Even since I began my work with the KITAB project, I have been aware of a large and mysterious entity known as the ‘cluster data set’. In those early days, Ryan Muther gently and generously guided me through passim, its workings, its outputs (and how to read them). He told me that passim produced an output that brought together all of the passages in the corpus that shared text into clusters. Unfortunately, owing to the manner in which these clusters were formed, very common passages, such as the kind found in hadith, created large and unwieldy clusters of largely unrelated material (more on this problem in part 2). The KITAB project chose to primarily focus on the pairwise data set, which was already sufficiently large and (as our blog posts alone have shown) an incredible resource for understanding many facets of Arabic book history.

In contrast, David Smith and his team, who developed passim for the study of American Newsprint, have focussed in their research primarily on the cluster dataset. In their study of ‘viral texts’ they are interested in how stories and anecdotes were reprinted across time. They focus on the larger clusters, which represent popular reprints, and what this can tell us about nineteenth-century newspaper culture.[[2]](#footnote-2)

I was drawn to the cluster data not to study the larger clusters (although that would certainly be a valuable, albeit very complex, study for Arabic). Instead I am interested in the smaller clusters (those that typically contain 3-10 passages). Such clusters can tell us, for example, how one account of the same historical event can be used by multiple authors and see how the different authors make changes or additions in their version. Such clusters could provide evidence of a stemmatic relationship, or hint at how multiple authors have reused the same non-extant source text.

A study of the cluster data is useful in my research for KITAB’s forthcoming, [co-authored book](https://kitab-project.org/research/books#the-erc-team-memories-books-communities-making-and-re-making-the-past-in-the-arabic-textual-tradition). In a chapter of that book, I will be exploring the Arabic accounts of a famine that occurred during the Fatimid *fitna* of 454/1062-466/1073, and what they might teach us about the cultural memory of the famine. I am interested in how accounts of that famine compare to accounts of famine in the Arabic corpus more generally. The chapter will, moreover, attempt to find evidence of quotation of a lost topographical history written by al-Jawwānī (d. 598/1205), which appears to have contained extensive descriptions of the famine. The cluster dataset can be used to explore both of these questions and I will make reference to some early findings of this study in some of the blog posts that follow.

# Some orienteering: pairs vs. clusters

As has been explained [on this website](https://kitab-project.org/methods/text-reuse), we run passim on the whole OpenITI corpus split into 300-word chunks, which we term milestones. Passim compares each milestone to every other milestone and if it finds shared text it outputs a text reuse alignment (a record that contains the shared text in both milestones and some statistics about their length, start and end points, and shared characters). For KITAB’s research, all alignments are grouped into csv files corresponding to book pairs. For example, all alignments between al-Tabari’s *Taʾrikh* and Ibn al-Athir’s *Kamil* are grouped into one file.[[3]](#footnote-3)

As part of its program, passim performs single-link clustering on every pairwise alignment. It takes the text of each alignment and searches for other pairwise alignments that overlap with it by at least 80%. Through this process, passim creates clusters of passages that all contain related text. At this stage, it is perhaps useful to clarify through an example.

Through my study of a pair of histories (Ibn al-Athir’s *al-Kamil fi-l-taʾrikh* and Ibn Taghribirdi’s *Nujum al-Zahira*), I find an alignment of interest, concerning the early-Islamic history of Egypt. The pairwise alignment is given in table 1.

|  |  |
| --- | --- |
| **0630IbnAthirCizzDin.Kamil.JK000911-ara1.mARkdown.ms0810** | **0874IbnTaghribirdi.NujumZahira.JK001330-ara1.completed.ms0017** |
| مريم جاثليق مصر ومعه الاسقف------------------------- بعثه المقوقس---------------- لمنع بلادهم فلما ---نزل بهم عمرو قاتلوه فارسل اليهم لا تعجلونا حتي نعذر اليكم وترون رايكم بعد وليبرز الي ابو مريم وابو مريام فكفوا وخرجا -----------اليه فدعاهما الي -----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------الاسلام او--------------------------------------- الجزية واخبرهما بوصية النبي باهل مصر بسبب هاجر ام اسماعيل------------------------------------------------------------------------------------------------------------------------------- عليه السلام --------------------------------------فقالوا قرابة بعيدة لا يصل مثلها الا الانبياء -------------------------------------------------------------------------------------------------------------------------------------------------------------------امنا حتي نرجع اليك فقال عمرو--- مثلي لا يخدع ولكنني اءجلكما ثلاثا لتنظرا ---------------------------------------------------فقالا زدنا فزادهم يوما فرجعا الي المقوقس فهم فابي ارطبون ان يجيبهما وامر بمناهدتهم فقال لاهل مصر اما نحن فسنجهد ان ندفع عنكم ولا نرجع اليهم------------------------------------------------------ فلم يفجا عمرا الا البيات وهو علي عدة فلقوه فقتل ارطبون وكثير ممن معه وانهزم الباقون وسار عمرو والزبير الي عين الشمس وبها جمعهم وبعث الي فرما ابرهة بن الصباح فنزل عليها وبعث عوف بن مالك الي الاسكندرية فنزل عليها قيل وكان الاسكندر وفرما اخوين ونزل عمرو بعين الشمس فقال اهل مصر لملكهم ما تريد الا قتال يوم قوم هزموا كسري وقيصر وغلبوهم علي بلادهم فلا - | مريم جاثليق مصر ومعه الاسقف ابو مريام في اهل البنيات بعثه المقوقس صاحب الاسكندرية لمنع بلادهم فلما تصافوا قال عمرو بن العاص---------- لا تعجلو-ا حتي نعذر اليكم -----------------ليبرز الي ابو مريم وابو مريام راهبا هذه البلاد فبرزا اليه فقال لهما عمرو انتما راهبا هذه البلاد فاسمعا ان الله بعث محمدا بالحق وامره به وامرنا به محمد وادي الينا كل الذي امر به ثم مضي وتركنا علي الواضحة وكان مما امرنا به الاعذار الي الناس فنحن ندعوكم الي الاسلام فمن اجابنا فمثلنا ومن لم يجبنا عرضنا عليه الجزية و--بذلنا له المنعة وقد اعلمنا اننا مفتتحوكم واوصينا بكم حفظا لرحمنا منكم وان لكم ان اجبتمونا بذلك ذمة الي ذمة ومما عهد الينا اميرنا استوصوا بالقبطيين خيرا فان رسول الله صلي الله عليه وسلم اوصانا بالقبطيين خيرا لان لهم ذمة ورحما فقالوا قرابة بعيدة لا يصل مثلها الا الانبياء معروفة شريفة كانت ابنة ملكنا وكانت من اهل منف والملك منهم فاديل عليهم اهل عين شمس فقتلوهم وسلبوهم ملكهم واغربوا فلذلك صارت الي ابراهيم عليه السلام مرحبا به واهلا وامنا حتي نرجع اليك فقال عمرو ان مثلي لا يخدع ولك-ني اءجلكما ثلاثا لتنظرا ولتناظرا قومكما والا ناجزتكم قالا زدنا فزادهم يوما فقالا زدنا فزادهم يوما فرجعا الي المقوقس---- فابي ارطبون ان يجيبهما وامر بمناهدتهم وقال لاهل مصر اما نحن فنجتهد ان ندفع عنكم ولا نرجع اليهم وقد بقيت اربعة ايام واشار عليهم بان يبيتوا المسلمين ف------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------قال الملا من----هم ما تقاتلون من------- قوم قتلوا كسري وقيصر وغلبوهم علي بلادهم فالح |

*Table 1: A pairwise alignment between Ibn al-Athir’s Kamil and Ibn Taghribirdi’s Nujum. The hyphens indicate characters inserted in the other text (calculated through a Smith-Waterman alignment algorithm).*

From this alignment, one could search passim’s cluster dataset for the two milestones (that is, look for clusters in which either ms0810 of the *Kamil* or ms0017 of the *Nujum* are involved). Table 2 gives the result of such a search.

|  |  |  |
| --- | --- | --- |
| **cluster** | **size** | **id** |
| 68719610468 | 6 | JK000911-ara1.mARkdown.ms0810 |
| **180388738305** | **7** | **JK000911-ara1.mARkdown.ms0810** |
| **249108229347** | **2** | **JK000911-ara1.mARkdown.ms0810** |
| 128849236741 | 2 | JK000911-ara1.mARkdown.ms0810 |
| 17179888967 | 3 | JK000911-ara1.mARkdown.ms0810 |
| 309237764767 | 3 | JK000911-ara1.mARkdown.ms0810 |
| 180388738302 | 5 | JK001330-ara1.mARkdown.ms0017 |
| 128849236740 | 2 | JK001330-ara1.mARkdown.ms0017 |
| **249108229347** | **2** | **JK001330-ara1.mARkdown.ms0017** |
| 17179888964 | 6 | JK001330-ara1.mARkdown.ms0017 |
| **180388738305** | **7** | **JK001330-ara1.mARkdown.ms0017** |

*Table 2: All of the clusters to which ms0810 of the Kamil and ms0017 of the Nujum belong. Bold rows indicate the text reuse that is given in the pairwise alignment between the two texts given in table 1 (this one pairwise alignment is found in two clusters, where part of the alignment is drawn into the larger cluster).*

The milestone in the *Kamil* belongs to six clusters and the one in the *Nujum* belongs to five clusters. Notice how the two texts share two cluster numbers (underlined in the table), cluster 249108229347 and 180388738305, indicating that this long pairwise alignment is split between two clusters. The first of these is only shared between these two books, indicated by the fact that the cluster is of size 2 (that is, part of the pairwise alignment is only shared by the *Kamil* and *Nujum*). However, the second cluster (representing only a part of the same alignment) is of size 7. That means that part of the alignment is shared between the *Kamil*, *Nujum* and 5 other texts in the OpenITI corpus. Table 3 lists the texts in that cluster.

|  |  |
| --- | --- |
| **Book and ms URI** | **Author and Title** |
| 0310Tabari.Tarikh.Shamela0009783BK1-ara1.mARkdown.ms1606 | Al-Tabarī, *Taʾrikh al-rusul wa-l-muluk* |
| 0597IbnJawzi.Muntazam.Shamela0012406-ara1.mARkdown.ms1039 | Ibn al-Jawzi, *al-Muntaẓam fi taʾrikh al-muluk wa-l-umam* |
| 0630IbnAthirCizzDin.Kamil.JK000911-ara1.mARkdown.ms0810 | Ibn al-Athir, *al-Kamil fi-l-taʾrikh* |
| 0634AbuRabicHimyari.Iktifa.Shamela0009770-ara1.completed.ms1000 | Abu al-Rabiʿ al-Ḥimyari, *al-Iktafaʾ bi-ma taḍammanahu min maghazi rasul allah* |
| 0774IbnKathir.Bidaya.Shamela0004445-ara1.completed.ms3276 | Ibn Kathīr, *al-Bidaya wa-l-Nihaya* |
| 0874IbnTaghribirdi.NujumZahira. JK001330-ara1.mARkdown.ms0017 | Ibn Taghribirdi, *Nujum al-zahira fi muluk misr wa-l-qahira* |
| 1411MuhammadTahirBarnaji.SahihWaDacifTarikhTabari. Sham19Y0145435-ara1.ms1957 | Muhammad Tahir Barnaji, *Ṣahih wa ḍaʿif taʾrikh al-tabari* |

*Table 3: A list of texts and milestones belonging to cluster 180388738305*

A simple glance at the list of books in table 3 suggests that this is an al-Tabari historiographical cluster. That is, all seven works have been clustered together because they are all reusing a piece of text that first appeared in al-Tabari’s *Taʾrikh.* In the case of the final work in the list, it is a modern work that appears to reuse the text as part of a discussion of al-Tabari’s work. In the other cases we would need to examine the text of the milestones themselves to ascertain the exact relationship between all of the texts in the cluster – to determine if they have all reused the passage directly from al-Tabari, if they have relied on intermediaries, or even if some share an earlier (lost) source text. For example, Ibn Taghribirdi cites Ibn Kathir as his source for this passage, so it appears that he had not directly reused al-Tabari. The cluster data thus allows us to contextualise pairwise alignments in the broader corpus and figure out the larger stemmatic relationships that can underlie them.

# The next waypoint

In this blog post I have attempted to introduce you to the cluster data set, explain its importance and how it is produced. As the example cluster given above shows, the data set can be valuable for understanding stemmatic relationships that span across the corpus, across time and space. However the neat example that I have presented here obscures some of the complexities of the data set as a whole, of its scale and of the weaknesses of single-link clustering, especially for the pre-modern Arabic tradition. The cluster data set contains over three and a half million individual clusters, many of which are very small and some of which are extremely large. In the next blog post, I will take a closer look at why the data set is so large, and provide some basic statistical analyses. This is necessary groundwork if we are to use the cluster data for historical research.

1. If you will excuse the wordplay, I mean this both in the sense of a ‘pleasant stroll’ through a data set, and to write in a somewhat disorganised fashion – that is, I will present ideas and thoughts that are in progress. [↑](#footnote-ref-1)
2. Ryan Cordell, ‘Reprinting, Circulation and the Network Author in Antebellum Newspapers’, *American Literary History 27.3,* (August, 2015), [*https://doi.org/10.1093/alh/ajv028*](https://doi.org/10.1093/alh/ajv028), 417-445; the (draft) chapter ‘Classifying Vignettes, Modeling Hybridity’ in Ryan Cordell, David A. Smith, Abby Mullen, Jonathan D. Fitzgerald and Avery Blankenship, *Going the Rounds: Virality in Nineteenth-Century American Newspapers*, (Minneapolis: University of Minnesota Press, forthcoming, draft chapters available online), <https://manifold.umn.edu/read/untitled-bd3eb0af-fdad-4dd6-9c94-3fd15d522ab6/section/06899e82-8f06-43d2-9fc9-ea04dffef886>. [↑](#footnote-ref-2)
3. For an example of such a file, see the pairwise file (listed as item 9) in Sarah Bowen Savant and Masoumeh Seydi’s Zenodo release for their recent work on ‘People versus Books’, <https://zenodo.org/record/5074633#.Yl_LZtrMKUk>. [↑](#footnote-ref-3)