MAXIM ROMANOV

TOWARD THE DIGITAL HISTORY OF THE PRE-MODERN MUSLIM WORLD: DEVELOPING TEXT-MINING TECHNIQUES FOR THE STUDY OF ARABIC BIOGRAPHICAL COLLECTIONS

Introduction

Historians of the pre-modern Muslim world are blessed with hundreds of Arabic historical sources, abundant in well-structured biographical records.[1] Largely multi-volume, each chronicle or biographical dictionary includes biographies in numbers that range from several hundred to tens of thousands. The largest source of this kind – *The History of Islām* (*Ta'rīkh al-Islām*) of al-Dhahabī († 748/1347 CE) – comprises 52 volumes, covers 700 years of Islamic history, and includes approximately 30,000 biographies.[2] The overall number of biographies in these sources reaches hundreds of thousands.[3]

A great number of these biographies are rather short notices – they often list only the name of a person with dates of his (more rarely, her) birth and death, whether precise or approximate. However, even onomastic data alone provides historians with a great deal of valuable information, mainly due to the part of the Muslim name known as the *nisba* ('descriptive name'). Let us take a close look at the following name:

[1] Paul Auchterlonie's reference – now outdated and incomplete, but still very helpful – lists over two hundred biographical sources. See Auchterlonie 1987.

[2] al-Dhahabī 1990; on this source, see Somogyi 1932.

[3] Over ninety years ago Italian scholars Leone Caetani and Guiseppe Gabrieli collected 250,000 biographical references, see Malti-Douglas & Fourcade 1976. My own biographical databank, which is still in the process of preparation, already includes over 86,000 biographies and biographical records (with only 24 biographical dictionaries processed).

10.1484/M.LECTIO-EB.5.102573

> Abū l-Faraj ʿAbd al-Raḥmān, the son of (*ibn*) ʿAlī, the son of (*ibn*) Muḥammad, the son of (*ibn*) [so-and-so] [...], the son of (*ibn*) Muḥammad, the son of (*ibn*) Abī Bakr al-Ṣiddīq, al-Jawzī, al-Qurashī, al-Taymī, al-Bakrī, al-Baghdādī, al-Ḥāfiẓ, al-Mufassir, al-Ḥanbalī, al-Wāʿiẓ, al-Ṣaffār.

This name includes nine meaningful descriptive names,[4] which tell us that this particular person belonged to the clan of Taym (al-Taymī) of the tribe of Quraysh (al-Qurashī) and was a descendent of Abū Bakr al-Ṣiddīq (al-Bakrī), the first of the four Rightly-guided caliphs of the Islamic community; a native of Baghdād (al-Baghdādī) and a jurist of the Ḥanbalī school of law (al-Ḥanbalī), he distinguished himself as a knowledgeable transmitter of Islamic tradition (al-Ḥāfiẓ), an exegete of the Qur'ān (al-Mufassir) as well as a public preacher (al-Wāʿiẓ); the last *nisba* (al-Ṣaffār) also tells us that he came from a family that earned its living selling copper utensils.[5] Thus, the onomastic information alone is tantamount to the social profile of a person. Studied as a whole, such social profiles can serve as unique lenses through which the historian can study different aspects of social history, which would otherwise be indiscernible.[6] Additionally, such profiles – in their entirety – form a unique body of data, which is ideally suited to different forms of sociological and spatial analyses.

People who became the subjects of these biographical records, of course, were not simple commoners. By and large, they were representatives of religious, administrative, military and literary

[4] In strict grammatical terms, a *nisba* is an adjective formed from a noun by means of adding suffix 'ī' and thus denoting a relation to the noun from which it was formed (Baghdād + ī => Baghdādī, meaning something or someone related to the city of Baghdād). In historical terms, however, the *nisba* does not seem to be limited to this particular morphological pattern; the term is rather used for any word that can meaningfully describe a person (including but not limited to such morphological patterns as *fāʿil*, *faʿīl*, *faʿʿāl*[*a*], *dhū shay'*[*ayn*] etc.). This is particularly true in the case of al-Samʿānī who included all of them in his *Kitāb al-ansāb* (*The Book of Descriptive Names*).

[5] In addition, the number of ancestors mentioned in the name (each begins with 'the son of [...]') tends to be proportional to the overall fame of a person.

[6] For a clear conceptualization of the approach, see Bulliet 1970. A somewhat similar approach, although more from a literary studies perspective, was offered in Fähndrich 1973. Unfortunately, the study promised in Fähndrich's work never came to fruition.

elites. Nonetheless, the lives of these notables are often presented with so many details that studying them as a whole will also shed light on the life of rank-and-file believers.

Prior attempts

A number of scholars have realized the potential of Islamic biographical sources, yet very few have ventured to approach them in this manner. In the 70s and 80s, on the wave of popularity of quantitative methods in history,[7] several scholars from different countries conducted methodologically similar studies, largely independently from each other. Analyzing biographies *en masse*, historians looked for answers to often quite different research questions. In Israel, Hayyim Cohen studied economic backgrounds of the early religious elite (Cohen 1970). In the USA, Richard Bulliet studied the social and religious elite of Nīshāpūr (Bulliet 1972), and later the process of conversion to Islām (Bulliet 1979); Carl Petry studied the civilian elites of Mamlūk Cairo (Petry, 1981). In the USSR, a group of Soviet scholars inspired by Piotr A. Griaznevich studied the development of Arabic historical and religious writings in different areas of the Caliphate.[8] The scholars of the Onomasticon Arabicum project produced a series of publications on a number of biographical dictionaries.[9]

Of all these scholars, however, only Bulliet and Petry remained faithful to the quantitative approach and produced more than a single study. The neglect of this kind of approach was mainly due to its extremely laborious and time-consuming nature (even with the help of early computers, which were anything but user-friendly at that time). The abundance of information was – and still remains – unfathomable, so research was extremely time-consuming and even the brave ones soon opted out from this kind of studies. From the middle of the 1980s until the end of the

[7] On the fate of quantitative methods in history, see Reynolds 1998.

[8] All from the Leningrad Branch of the Institute of Oriental Studies of the Academy of Sciences of the USSR. Of four planned books, three were published. Unfortunately, written in Russian, they remained unknown to Western scholars. See Boyko 1977; Prozorov 1980; Boyko 1991. All books have summaries in English.

[9] See Graff & Bichard-Bréaud 1971; Pascual 1971; Bichard-Bréaud 1973; Malti-Douglas & Fourcade 1976; Rowson & Bonebakker 1980.

1990s there were almost no studies that relied on this approach. In the late 1990s, advancements in computer technologies and the availability of personal computers stimulated a few more attempts (most notably the Jerusalem Prosopography Project [JPP] founded and directed by Michael Lecker;[10] and the [Netherlands] Ulama Project [NUP] of John Nawas and Monique Bernards),[11] but the potential of the approach is still far from having been realized.

The main problem with the quantitative approach was posed by its very advantage: the limitless data available for analysis. Anyone attempting to implement this approach had to set very strict limits in order to accomplish one's research project: limits on the number of sources, historical periods, geographical areas, and clearly formulated research goals; all this at a time when computers were not widely available and even when they were their use often posed new kinds of problems (no support for Arabic, encoding issues, etc.).[12] Overall, one had to define clearly the research goals, select a limited number of sources, and carefully consider kinds of data required for the research. After careful planning one had to peruse the selected sources, manually extract the required information, and then record it either on paper media or to encode it for transfer to 'the memory bank of a computer'. These technologically imposed limitations also affected the usability of the extracted information: in many cases, the potential of the created databanks was exhausted by the end of the research project, for which they were created. This must have also contributed to the unattractiveness of such endeavors.

[10] On Michael Lecker's work and his Jerusalem Prosopography Project, see his webpage: ‹http://micro5.mscc.huji.ac.il:81/JPP/homepage/›. Overall, this seems to be a rather conventional prosopographical project, in many aspects similar to those on Ancient Rome, Byzantium and Medieval Europe. It deals with Early Islamic Administration (*c.* 622–800) and includes 1,650 persons. For a study based on this project, see Ebstein 2010.

[11] For the technical description of the project, see Nawas & Bernards 1998; for studies based on the NUP database, see Bernards & Nawas 2003; Nawas 2005 and 2006.

[12] Essentially, each project had to develop its own coding system, but in the case of computerized databases it was a particularly important and complex task, especially if a group of scholars was expected to be involved. To appreciate the scale of such an enterprise, see Malti-Douglas & Fourcade 1976, which is a 100 page manual for the Onomasticon Arabicum project.

Needless to say, these technical limitations inhibited the realization of the promises offered by the quantitative approach. Later, when computers became personal and more user-friendly, several more attempts were made to study large bodies of biographical records. (The change in computer standards also played a nasty trick, since previously digitized information now had to be converted to a format readable by new machines, which did not always work out well and some computerized databanks remained on obsolete media.[13])

It was expected that new computerized databases would 'open up a new range of questions that can be asked that would hitherto have been unthinkable "without 500 monks at hand"'.[14] In real life, however, they offered only marginal advantages over the old-fashioned pen-and-paper systems. It is true that they increased the speed and complexity of data retrieval from databases. Yet, they offered no significant improvements for the most tedious process of entering data into databases. Thus, the newcomers continued to suffer from the same limitations as their analog predecessors. Their creation remained equally time-consuming and for this reason a number of projects were never finished.[15] This seems to be true of the above-mentioned JPP and NUP databases: they were created over a rather long period of time, included a relatively small number of biographic profiles (1,650 people in the JPP and 1,000 in the NUP), and their 'coefficient of efficiency' – in terms of numbers of studies based on them and their overall impact on the field – was insignificant, playing well into the hands of the critics of prosopographical and quantitative studies. Another problem with large-scale database projects

[13] For example, this happened to Carl Petry's databank, which still remains on magnetic tapes and requires special equipment and expertise in order to be transferred to a modern type of media (from personal conversations with Carl Petry); interestingly enough, old-fashioned analog databases remained immune to these advancements in technology – Richard Bulliet's analog database on McBee Keysort cards still serves him almost fifty years after his project began (from personal conversations with Richard Bulliet). For an example of usage of McBee Keysort cards in humanities, see, e.g., Anderson 1953.

[14] The quote is from Mathisen 2007, p. 95. The article is an interesting overview of the use of databases – both analog and digital – in history (prosopography, to be more precise); it is also an excellent representation of the conventional relational database approach with all its advantages and disadvantages.

[15] On this issue, see Mathisen 2007.

is that they often tend to take on a life of their own, gradually transforming from the means into an end in themselves.

I myself had experience in dealing with conventional relational databases when I suggested that the methods used by the above-mentioned Boyko, Griaznevich and Prozorov be computerized. Working with Stanislav Prozorov, my mentor at the St Petersburg Branch of the Institute of Oriental Studies of the Russian Academy of Sciences,[16] I developed a database for the study of Arabic historical sources; however, its earthly manifestation turned out to be quite different from what we both envisioned and hoped for.[17] As always, the bottleneck of the database was the process of entering data, and 'without 500 monks at hand' it was an incredibly inefficient and time-consuming task for a lone graduate student. Eventually, I had to abandon this method and look for alternative ways of analyzing historical data.

New approach

Almost ten years later, however, I am hopeful that there is an efficient way of overcoming these limitations of conventional relational databases. Proposed here is a method that capitalizes on a number of developments in the digital sphere that allow for a very different angle of approach. First, a great number of Arabic historical texts became available in fully searchable format – huge digital libraries are now available on a number of Arabic websites, CDs, DVDs and even dedicated hard drives.[18] Second, the development and wide acceptance of the Unicode standard has made it possible to work efficiently with Arabic texts on personal

[16] This is the same academic institution where Piotr Griaznevich's project on the study of the formative period of Arabic historiographical literature took place. Currently, the name of the institution is the Institute of Oriental Manuscripts of the Russian Academy of Sciences (‹www.orientalstudies.ru›).

[17] The project was described in details in Prozorov & Romanov 2003.

[18] Most of the available texts are very careful reproductions of printed editions, often with pagination and inconsistencies of printed editions faithfully preserved; there are exceptions and one should always be collate such electronic texts with their paper editions. It should be noted here that most of the editions published in the Arab world are not critical (at least not in the rigorous sense of European medievalists and classicists), however, they are widely used by the scholars of medieval Islam if only because there are no other editions available and most of the manuscripts are not easily accessible.

computers (although it must be said that quite a few frustrating quirks, especially when dealing simultaneously with Arabic and non-Arabic characters, still remain unsolved). Third, Unicode also made it possible to use powerful tools such as scripting languages and regular expressions in the analysis of Arabic texts. Python is one of the most popular scripting languages for text-mining tasks, while regular expressions are a powerful tool for manipulating patterned text.

The method is meant to overcome two major issues. The first issue is to reduce time costs of laborious data entry by 'delegating' the task as much as possible to the computer. The second issue is to overcome the structural rigidity of conventional relational databases in order to make the database easily adaptable to new tasks, to make it available for research purposes as soon as possible, and to avoid the trap of the database becoming the goal in itself.

The solution for the first issue rests on the premise that we are dealing with highly structured texts where specific kinds of information (e.g., chronological, onomastic, and toponymic data) conform to distinctive textual patterns, which can be described with regular expressions. The solution for the second rests on the following proposition. Keeping in mind that the development of a database structure is a lengthy process that requires advance knowledge of all relevant types of information and their interrelations (the finished structure does not tolerate well any later alterations), this task should be postponed until the moment when all the required information is extracted and research questions are not only clearly formulated, but also adjusted to the available data. Thus, data are stored in 'source files', which become a databank that can be updated with new kinds of information at any moment; they also serve as the source from which a database could be automatically (re)generated at any time to fit current research agenda.

Given the unique organization and peculiar textual patterns of each individual source, it has to be treated separately and as a whole. First, the source must be tagged in a specific manner, so that a computer (or, more precisely, scripts) may differentiate between its structural elements and split the text of the source into independent data units, such as descriptions of events and texts of biographies/obituaries. These independent units of in-

formation become 'source files', where required information is stored in machine-readable format, which can be fed into a database. Mapping the structure is a rather tiresome task, but it allows preservation of the entire text of the source and it need be done only once. Moreover, there are ways to simplify this task, first, by using short tags for the markup of all structural elements (e.g., '|' – for a chapter of the first level; '||' – for a chapter of the second level, etc; '$' – for a biography of a man; '$$' – for a biography of a woman)[19] and, second, by using highlighting schemes, which helps both to avoid typos and to make structural elements easily visible (if tagged correctly, headings are highlighted according to the user-defined conditions, see fig. 1).[20]

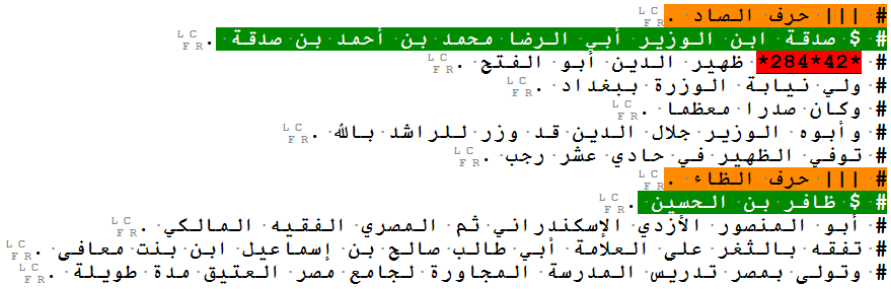

FIGURE 1: The third level headings are highlighted with orange, while the headings of biographies are highlighted with dark green.

When a source is tagged, it can be parsed into separate files (depending on the length of the source, it may take up to a few minutes for the script to generate these individual files). Each of the newly generated files has two parts: the first one is a *cubaron*, [21] the paragraph that contains tagged metadata extracted from the source; and the second one is an e*Naṣṣ*,[22] the actual text of

[19] These tags can later be easily transformed into TEI tags. In the case with Arabic texts, this is actually the best way to follow, since angle brackets as well as other technical symbols behave erratically when combined with Arabic letters; adding corrections and changes to such text is particularly difficult and annoying.

[20] This functionality is available in EditPad Pro (‹http://www.editpadpro.com/›), the only text editor, which handles long text files, supports Unicode, regular expressions and customizable highlighting schemes.

[21] I borrowed this term from the Onomasticon Arabicum project, where it is used for a paragraph, which 'gathers the totality of biographical and bibliographical information concerning one sole person' (Malti-Douglas & Fourcade 1976).

[22] From an Arabic word *naṣṣ* ('text'); the choice of both *cubaron* and *eNaṣṣ* is

a biography or an historical event. Initially, the *cubaron* contains the following information: the name of the source from which it was extracted, the logical path to the e*Naṣṣ* within the source (that is, the names of chapters and subchapters wherein the e*Naṣṣ* was nested), the volume number and page locations of the e*Naṣṣ* text for easy reference, and the type of information that the *eNaṣṣ* contains (e.g., *biography* or *event*). The source files are now ready for text-mining scripts.

Within the Onomasticon Arabicum project, participating scholars extracted and encoded all relevant information from biographies, which included the following rubrics: names with each of their numerous elements coded separately; physical attributes; honorifics; religious affiliations; occupations; scientific and intellectual interests; places of birth, residence, death; affiliated religious institutions; family and tribal relations; age of death; dates of birth and death.[23] Now, significant amounts of this information can be extracted and processed automatically.[24]

In the case of manual data processing each biography is encoded one by one. Computationally, however, it is more efficient to work not with biographies, but with specific kinds of data – i.e. to extract one particular kind of data at a time from all the biographies of a particular source. Each script can be easily adapted to process historical dates, descriptive names (*nisba*s), toponyms, etc.; moreover, dealing with the same type of information makes it easier to discern patterns and thus adjust both text-mining scripts and regular expressions for better performance. Another advantage of such an approach is that it will allow the historian to begin the analysis of data long before the database is complete. Starting the analysis with only a few parameters, the historian will gradually be able to increase the complexity of analysis as new parameters become available.

arbitrary, but necessary, since the usage of more common words makes communication very confusing.

[23] For the complete list of rubrics, see Malti-Douglas & Fourcade 1976, p. 133-134.

[24] This is not to say, however, that a computer will do all the work. The main idea behind such automation is to let the computer perform the most laborious tasks and generate information in the form of suggestions, while a human checks the suggested information and manually corrects it if necessary.

Let us return for a moment to our first example of a traditional Islamic name, which included nine meaningful descriptive names (*nisba*s). The issue is how these descriptive names can be extracted and coded automatically. To resolve this issue, we can create a machine-readable list of descriptive names from *Kitāb al-ansāb* of al-Sam'ānī († 562/1166 CE), a pre-modern dictionary of descriptive names with over 4,400 entries.[25] First, the structure of the source must be tagged in the same manner as described above and then parsed into individual source files. Each entry from *Kitāb al-ansāb* includes three important units of information: 1) a descriptive name itself, 2) its vocalization, i.e. how this *nisba* should be pronounced,[26] and 3) its definition. Definitions are particularly important, since they can be used for automatic sorting of descriptive names into categories (ancestral, geographical, religious, occupational, tribal, etc.).

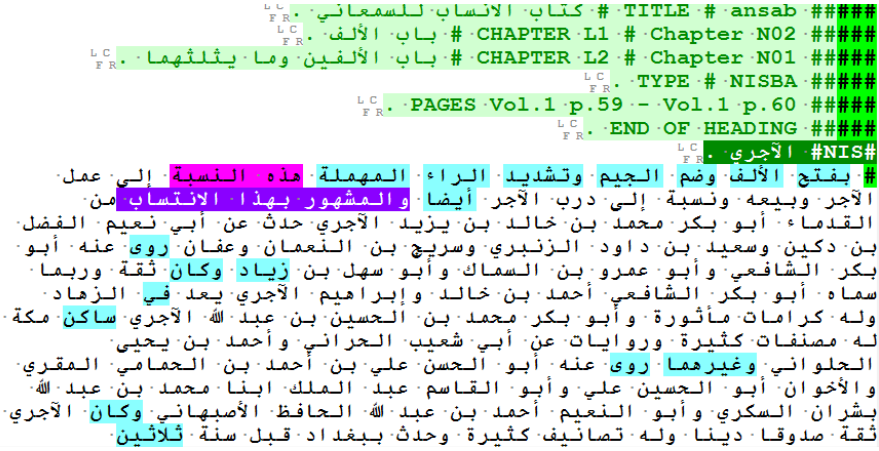

FIGURE 2: A sample entry from the Kitāb al-ansāb of al-Sam'ānī.

[25] For a similar task of extracting toponymic data another pre-modern dictionary can be used – *Mu'jam al-buldān* of Yāqūt al-Ḥamawī († 626/1229 CE), which includes over 14,000 entries.

[26] In medieval Arabic handwriting, short vowels and consonantal diacritical dots were often omitted. This could lead to a lot of confusion, especially in cases of words of non-Arabic origin (e.g., without diacritical dots, letters *b*, *t*, *th*, *n*, and *y* in the beginning and in the middle of a word look exactly the same). For this reason authors 'spelled out' difficult words using different descriptors of vowels and consonants.

Fig. 2 illustrates a sample entry from *Kitāb al-ansāb* with the *cubaron* above the line 'END OF HEADING,' and a beginning of the *eNaṣṣ* right below this line. #NIS# is followed by the actual *nisba* – al-ājurrī, the subject of the entry. Highlighted with light blue, fuchsia and purple are the elements of textual patterns, which are used to extract required information (highlighting is automatic and based on regular expressions).

The line of word highlighted with light blue is the vocalization of the *nisba*, the pattern for which is rather simple – it includes names of Arabic letters, their descriptors, short vowels and a few other technical terms, which can follow each other in any order and make up a phrase of at least three words. Fig. 3 demonstrates how this pattern is rendered into a regular expression. The results of this automatic extraction were quite impressive: I had to correct only about 300 entries (i.e., less than 7%), while over 4,000 vocalizations were extracted correctly. It took about four minutes for the script to parse the entire book.

```
\b(prefixBeg|(إأآا]لف||همز|[إأآا]?ل)?([إأآا]?(و)?|prefixEnd)?(letters||
ب[إأآا] (ؤائ|اوي) |ت[إأآا] (ؤائ|اوي) |ث[إأآا] (ؤائ|اوي) |ج (ياى)م|اح
[إأآا] (ؤائ|اوي) |خ[إأآا] (ؤائ|اوي) |د[إأآا]ل|ذ[إأآا]ل|ر[إأآا] (ؤائ|
واي) |ز[إأآا] ((ياى)ن?|(ؤائ|اوي))|س(ياى)ن|ش(ياى)ن|ص[إأآا]د|ض[إأآا]د|ط
[إأآا] (ؤائ|اوي) |ظ[إأآا] (ؤائ|اوي) |ع (ياى)ن|غ (ياى)ن|ف[إأآا] (ؤائ|اوي) |
ي) |ق[إأآا]ف|ك[إأآا]ف|ل[إأآا]م|م (ياى)م|انون|ه[إأآا] (ؤائ|اوي) |و|إأآا]
موحد|مهمل|مهل|معجم|اخف(ياى)ف||letter descriptions|(و|(ياى)[إأآا]|(ؤائ|اوي)
مخفف|صر(ياى)ح|مقصور|قصر|قطع (ياى)|مثلث|مثن[إأآا]|محف[إأآا]خر [إأآا]
مهموز|مد|ممدو?د|تحر(ياى)ك|مفتوح|مضموم|مكسور||harakat|?(لحروف|[إأآا]خر(ياى
ت?فتح|ت?ضم|ت?كسر|س[إأآا]كن|سكون|ت?شد (ياى)د|مشدد|تخف(ياى)ف|فوقه[إأآا]
ي)||connectors|نقط|من فوق|تحت كل و[إأآا]حد(ه|ة) منهم[إأآا] نقطت[إأآا]ن
ى)لتق(ياى) ف(ياى)اس[إأآا]كن|ووجدته ف(ياى) كت[إأآا]ب نصر|ك[إأآا]ن|اجمع|
بعد|[إأآا]خر|ال(ياى)س ب(ياى)ن|ارو[إأآا]ه بعض|ثم|(ياى)لتق(ياى) مع|(ياى)لتق
(ياى) عند|[إأآا]ول|ث[إأآا]ن(ياى)ث|ث[إأآا]لث|تثن(ياى)(ه|ة) لفظ|تثن(ياى)|
مذكور قبل|[إأآا]لذ(ياى) قبل|از(ياى)[إأآا]د|لفظ|اب(ياى)ن|اب(ياى)نهم[إأآا]||
نحو جمع|وزن|تصغ(ياى)رات[إأآا]ن(ياى)ث|مصغر|اغ(ياى)ر مض[إأآا]ف|ال[إأآا] (ي|
ى)جوز|(ياى)جوز [إأآا]ن (ياى)كون|رو(ياى) |شطره ك[إأآا]لذ(ياى) قبل|اشطر|مثل
شطر [إأآا]لذ(ياى) قبل|مثل [إأآا]لذ(ياى) قبل|[إأآا]لت(ياى) قبل|تشبه|وربم
[إأآا] مدت|قد تقدم ف(ياى)|قد (ياى)كتب|هكذ[إأآا] وجدت|هكذ[إأآا] ج[إأآا]
(ؤائ|اوي) امض[إأآا]ف [إأآا]ل(ياى)|اقر[إأآا]ت بخط|بعضهم (ياى)قول
ه|ة) |ت?[إأآا]ن|ات?(ياى)ن|ات?ه[إأآا]|ت?ه[إأآا]?|ت?هم)|suffixBeg)
|هم[إأآا]?|suffixEnd)?\b
```

FIGURE 3: Regular Expression for vocalizations.

The task of extracting definitions is more complicated. It is based on the premise that each thematic section begins with a rather limited number of *technical expressions*, instances of which

are frequent enough to make automatic extraction not only possible, but also efficient.

In terms of patterns, the structure of definitions can be described in the following manner. The first common pattern begins with one *technical expression* (TE) and ends with another. In the current example, the *initial* TE is 'this *descriptive name* [refers to]' (*hādhihi l-nisba* [*ilá*], highlighted with fuchsia), while the *terminal* TE is 'and known under this description' (*wa-l-mashhūr bi-hādhā l-intisāb*). This terminal TE is in fact an initial TE for another section of information – the list of people known to bear this name. Thus, the definition itself is the text that starts with the initial TE and ends right before the terminal TE. The second pattern has an initial TE, but does not have a terminal TE; definitions of this pattern are the text of the initial TE plus 10-15 words that follow it. The third common pattern has no initial TE, but does have a terminal TE. Definitions of this pattern are the text from the very beginning of an entry until the beginning of a terminal TE, minus *vocalization*, which was extracted by the previous script and is now available for manipulations.

Instances that do not fall into these patterns are not numerous and can be tagged manually. Fig. 4 shows the results of running the definition script.

```
>>> ================================ RESTART ================
>>>
.........1000.........2000.........3000.........4000....
--------------------------------------------------
Processed definitions:         4450
Unprocessed definitions:          9

--------------------------------------------------

Both topoi:                   3610
Fixed pattern:                  16
Initial only:                  702
Terminal only:                  73
Manually tagged:                39
Without definitions:            10

--------------------------------------------------

Execution time: 0:04:28.878000
```

FIGURE 4: The results of running a script that extracts definitions.

After the application of these two scripts, the updated *cubaron* of each source file looks similar to what is shown in fig. 5, where both the vocalization and the definition of the descriptive name are extracted and tagged.

The last step in text-mining this particular source is to assign each *nisba* to a specific category, or categories, by using sets of keywords. The definition of this particular *nisba*, al-ājurrī, says the following: 'This descriptive name refers to the production and selling of 'baked bricks', it may also refer to the Darb al-ājurr ['the Road of Baked Bricks,' a quarter in the western part of Baghdā d]'. Thus, this definition contains such keywords as 'production' (*'amal*) and 'selling' (*bay'*), which can be used to automatically assign the *nisba* to the category of *occupation*, and the keyword 'road' (*darb*), which can be used to categorize it also as a toponymic *nisba*. This particular case has to be finalized manually, since the same descriptive name refers to two different entities. However, most descriptive names do not pose such a complication. After these tasks are accomplished, the data from the source files can be converted into formats suitable for other text-mining tasks.

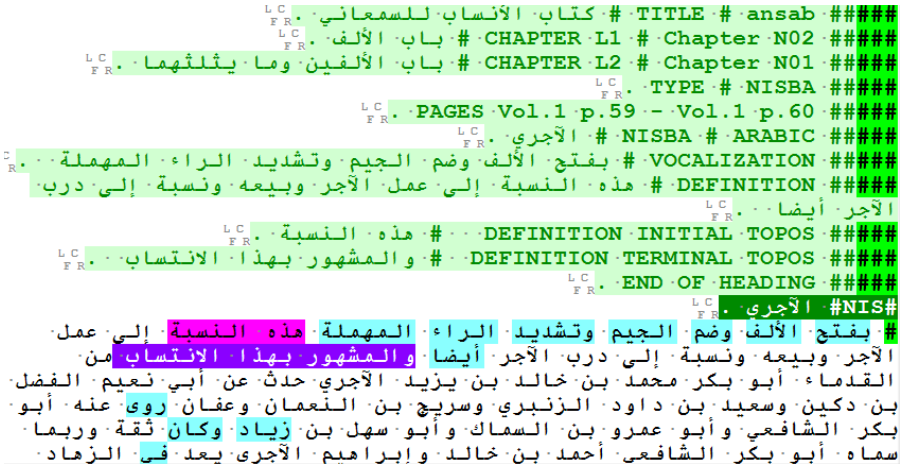

FIGURE 5: The same example from *Kitāb al-ansāb* ('The Book of Descriptive Names') with the updated *cubaron*.

***

The project is still in its rather early stages of development, but I hope that even preliminary results prove that the method is effective and will soon allow us to realize the full potential of the quantitative approach, which was proposed over forty years ago.

Bibliography

M. b. A. al-Dhahabī (1990), *Tārīkh al-Islām wa-wafayāt al-mashāhīr wa-al-aʿlām*, Bayrūt, Lubnān: Dār al-Kitāb al-ʿArabī.

G. L. Anderson (1953), 'The McBee Keysort System for Mechanically Sorting Folklore Data', in *The Journal of American Folklore*, 66, 262, p. 340-343.

P. Auchterlonie (1987), *Arabic biographical dictionaries: a summary guide and bibliography*, Durham: Middle East Libraries Committee.

M. Bernards & J. Nawas (2003), 'The Geographic Distribution of Muslim Jurists during the First Four Centuries AH', in *Islamic Law and Society*, 10, p. 168-181.

P. Bichard-Bréaud (1973), *Traitement automatique des données biographiques; analyse et programmation*, Paris: Éditions du Centre national de la recherche scientifique.

K. A. Boyko (1977), *Arabskaia istoricheskaia literatura v Ispanii: VIII-pervaia tret' XI v*, Moskva: Glavnaia red. vostochnoi literatury.

K. A. Boyko (1991), *Arabskaia istoricheskaia literatura v Egipte, IX-X vv*, Moskva: 'Nauka,' Glav. red. vostochnoi lit-ry.

R. W. Bulliet (1970), 'A Quantitative Approach to Medieval Muslim Biographical Dictionaries', in *Journal of the Economic and Social History of the Orient*, 13, p. 195-211.

R. W. Bulliet (1972), *The patricians of Nishapur: a study in medieval Islamic social history*, Cambridge, Mass.: Harvard Univ. Press.

R. W. Bulliet (1979), *Conversion to Islam in the medieval period: an essay in quantitative history*, Cambridge: Harvard Univ. Press.

H. J. Cohen (1970), 'The Economic Background and the Secular Occupations of Muslim Jurisprudents and Traditionists in the Classical Period of Islam: (Until the Middle of the Eleventh Century)', in *Journal of the Economic and Social History of the Orient*, 13, p. 16-61.

M. Ebstein (2010), 'Shurṭa chiefs in Baṣra in the Umayyad period: a prosopographical study', in *Al-Qantara (Madrid)*, 31, p. 103-147.

H. E. Fähndrich (1973), 'The Wafayāt al-Aʿyān of Ibn Khallikān:

A New Approach', in *Journal of the American Oriental Society*, 93, p. 432-445.

Graff & P. Bichard-Bréaud (1971), *Documents sur la mise en ordinateur des données biographiques*, Paris: Éditions du Centre national de la recherche scientifique.

F. Malti-Douglas & G. Fourcade (1976), *The treatment by computer of medieval Arabic biographical data: an introduction and guide to the Onomasticum [i.e., Onomasticon] Arabicum*, Paris: Editions du Centre national de la recherche scientifique.

R. W. Mathisen (2007), 'Where are all the PDBs?: The Creation of Prosopographical Databases for the Ancient and Medieval Worlds', in *Prosopography Approaches and Applications: A Handbook*, University of Oxford, Linacre College Unit for Prosopographical Research.

J. Nawas (2005), 'A profile of the mawālī 'ulamā'', in M. Bernards & J. Nawas (eds.), *Patronate and patronage in early and classical Islam*, Leiden & Boston: Brill, p. 484-480.

J. Nawas (2006), 'The birth of an elite: mawālī and Arab 'ulamā'', in *Jerusalem Studies in Arabic and Islam*, 31, p. 74-91.

J. Nawas & M. Bernards (1998), 'A preliminary report of the Netherlands Ulama Project (NUP): the evolution of the class of 'ulamā' in Islam with special emphasis on the non-Arab converts (mawālī) from the first through fourth century A.H', in U. Vermeulen & J. M. F. Van Reeth (eds.), *Law, Christianity and modernism in Islamic society. Proceedings of the Eighteenth Congress of the Union Européenne des Arabisants et Islamisants … Leuven … 1996 (Orientalia Lovaniensia Analecta,* 86*)*, Leuven: Peeters, p. 97-109.

J. P. Pascual (1971), *Index schématique du Ta'rih Baġdād*, Paris: Éditions du Centre national de la recherche scientifique.

C. F. Petry (1981), *The civilian elite of Cairo in the later Middle Ages*, Princeton (NJ): Princeton Univ. Press.

S. M. Prozorov (1980), *Arabskaia istoricheskaia literatura v Irake, Irane i Srednei Azii v VII-seredine X v.: shiitskaia istoriografiia*, Moskva: Izd-vo 'Nauka,' Glav. red. vostochnoi lit-ry.

S. M. Prozorov, & M. G. Romanov (2003), 'Principles and procedures of extracting and processing the data from Arabic sources (based on materials of historic-and-biographical literature) / Original title: Metodika izvlecheniya i obrabotki informatsii iz arabskih istochnikov (na materiale istoriko-biograficheskoi literaturi)', in *Oriens/Vostok*, 4, p. 117-127.

J. F. Reynolds (1998), 'Do historians count anymore? The status of quantitative methods in history, 1975-1995', in *Historical methods*, 31, p. 141-148.

E. K. Rowson & S. A. Bonebakker (1980), *A computerized listing of biographical data from the Yatīmat al-dahr by al-Tha'ālib*, Paris & Los Angeles (CA): Centre national de la recherche scientique & University of California (Onomasticon Arabicum, Série 3).

J. d. Somogyi (1932), 'The Ta'rīkh al-islām of adh-Dhahabī', in *Journal of the Royal Asiatic Society of Great Britain and Ireland*, 4, p. 815-855.