Text Mining with the Atthidographers

André Bünte

Abstract

During the Third eAQUA workshop that took place at the University of Leipzig on the 27th – 28th of June 2010, I presented approaches to investigating the Atthidographers using text mining methods. The Atthidographers are a group of ancient authors who had written a history of Athens, whose works, the so-called Atthides, are now lost. The content can be extrapolated only from scarce references in surviving texts. I shall here summarize problems that arise when dealing with fragmentary authors represented in digital text corpora and specifically when text mining methods are applied. My second aim is to present strategies that bypass the methodical disadvantages of the material. At the same time, I am here providing a concise introduction of the tools being used and finally a sample application in order to allow for an evaluation of the results.

Keywords

Atthidography – fragmentary texts – semantic technologies – text re-use – digital text corpora

The Atthidographers though their texts have been lost are nevertheless an important subject of modern research in ancient history. Surviving sources often refer to them when giving important dates or circumstances in relation with Athens that are otherwise not mentioned or contradicted. Today Athens is seen as the birthplace of the concept of modern democracy and so the interest in its history and especially in the cornerstones of the development and changes of its constitutions is great. By using succinct indications, modern scholars try to reconstruct the assemblage of the works. The current *communis opinio* follows the analyses of Felix Jacoby. Jacoby argued that all Atthides dealt with the history of Athens from mythical times down to their present - nested in an annalistic framework and differing only slightly from each other. Because the works themselves have been lost, the evidence of any assumption, even the apparently most self-evident, must be subject to evaluation. Jacoby has himself pointed out the weakness of some of his arguments and one is inclined to think that he stated them not only because the evidence lead him to the respective conclusion but also as he just had no better arguments at hand (Jacoby 1949 and 1954).

Investigating the Atthidographers with text mining methods might throw a new light on the arguments and lead to a new conclusion. However the results can be easily evaluated against the existing literature. Though text mining has its flaws, it draws on data mining, statistics, computational linguistics and information retrieval and tries to derive high-quality information from text. In order to do this automatically, it usually needs a substantial amount of text. In the eAqua project several corpora are incorporated that contain texts deriving from ancient Greece and Rome such as the TLG and the PHI 5 and 7 corpora. The TLG aims to include all merely literary Greek texts that have been produced during a period of more than 2000 years (700 BC to ca. 1600 AD). It is therefore the base corpus for investigating the Atthidography. The way the TLG is composed implies some problems. The main problem is a pragmatic way of selecting editions resulting in a mixed up set of texts. The problem can well be illustrated in the case of the Atthidographers. Despite being lost, the interested student will find them as individual items among the corpus authors' list. Commonly, the following authors are considered Atthidographers: Hellanicus of Lesbus, Clidemus, Androtion, Phanodemos, Demon and Philochorus, all Athenians. Their 'works' are named 'Fragmenta' in the corpus. These fragments are collections of direct references taken out of other texts that have been written in a time when the referenced work was still available, though even the fact of that availability has been disputed by modern scholars (e.g. Costa 2005). Since the

beginning of the 19th century several collections containing the Atthidographers have been published (Lenz & Siebelis 1811; Lenz & Siebelis 1812; Mueller 1849; Jacoby 1954; an online version with emerging new editions: Worthington 2010). The collection of Carl Gotthold Lenz has been published by Carl Gottfried Siebelis and subsequently incorporated into the collection of Greek fragmentary historians by Carl Mueller. Now this Mueller collection has been used in the TLG for the authors Clidemus, Androtion, Phanodemus and Demon, although there is a more recent and more comprehensive edition by Jacoby available. From the latter collection derive the TLG-editions of Philochorus and Hellanicus. The biggest differences can be stated here, as Jacoby distinguishes the references into being a fragment of the actual work of the author or being a testimonium of the actual life of the author. The Mueller collection does not. So the Jacoby editions always have two kinds of 'works', while the Mueller editions only have one. Besides, another more problematic difficulty concerns the integrity of the text. Between the publication of the Mueller collection and Jacoby's 'Fragmente der Griechischen Historiker' there has been quite a boom of publishing new editions of almost all major Greek texts and especially of scholia, were a vast amount of conjectures has been made by ambitious philologists of the late 19th century including also the exchange of proper names like the frequent substitution of Anticleides with Clidemus or of Andron with Androtion. In the TLG the chaotic incorporation of these changes results in a situation in which the original text containing a reference might look different and bear other proper names than the extracted reference that has been drawn from a fragment collection so that the passages don't match. When interpreting the results of tools like the Citationsgraph one has to bear such imponderabilities in mind. Thus, the Citationsgraph can be used to prove the well-known TLG corpus phenomenon that each and every sentence inside the entry of a lost author reoccurs at least one time inside the corpus in the text from which it had been extracted once. In order to do that, first the author has to be selected. An option would be the Atthidographer Clidemus. All the sentences that have been calculated as being similar are displayed in the list. We see that not all fragments are represented among those sentences. An example of a mutilated double quotation due to the use of differing editions has been depicted in figure 10.

Another awkward aspect of analysing the Attidographers using text mining methods shouldn't be omitted. Most of their 'works' are interpolated with other authors and therefore appear more then twice in the TLG: first of all in the quoting text and then under the entry of each of the authors that are mentioned in the reference. Most often this interpolation is circumvented by the fact that in eAqua the semantic window for the analytics concurs with the individual sentence thus connecting semantic content mentioned in conjunction with the authority of its provenance and detaching all other contents that are mentioned in other sentences within the same fragment but derived from other sources. Another advantage of clustering the whole corpus into single sentences is the levelling of the statistical difference between the large text corpus of a surviving work like Herodotus' Histories and the quite small collections of citations that can be attributed to individual Atthidographers. The figures of Herodotus' work with a total number of 189.489 words or the 'Peloponnesian War' by Thucydides with 153.260 words suit well the requirements for existing text mining methods whilst the TLG sub-corpus of all Atthidographers consists of 41.743 words with an average of just 1.500 words for the individual FHG edition author and 17.500 words for the individual FGrHist author. Despite the vast inconsistencies in the sub-corpus of the Atthidographers there is a huge discrepancy between Herodotus and any single Atthidographer. When a statistical text comparison is done based on the semantic window of a single sentence the matter of whether the sentence belongs to a huge or to a small subcorpus is no longer important. The limits of such arbitrary divisions are crucial too. There are of course many cases in which a possible semantic connection between two or more sentences would be dissolved. It is therefore necessary to review the results in their original context.

After this instant introduction of advantages and potential fen fires that arise when the Atthidographers are analysed with text mining methods some samples of the application of two text mining tools, the Citationsgraph and the Co-occurrence analysis, together with concepts of strategies for inquiry that are adapted to the mentioned difficulties of the subcorpus are pointed out.

My first analysis concerns the word ‘Atthis’ with most of its appearances in the corpus.[1] The goal is to examine the context in which the word appears in the corpus in order to find out what Atthis could mean and in how many cases its meaning points to Attic history. An appropriate approach to visualize its context is using the wordnet graph of the co-occurrence analysis tool. This graph displays the semantic relationship of a given word based on data mining results of the whole corpus. The data mining inquires the significance of one word to another. This significance might be varied in accordance with individual goals. There exist several algorithms for calculating the significance of co-occurrences that bear different concepts of significance. Some consider just the observation of occurrences like the SigFreq measure, some put the observation and the expectation of occurrences into relation like the Log Likelihood (Dunning 1993; Rayson & Garside 2000). In addition the concepts differ in the matter of valuating words with high and low frequencies. For the Atthidographers, measures like Dice or Frequency are not useful since they favour high frequency words. More appropriate seem measures like Log Likelihood or Mutual Information that focus on occurrences of words occurring infrequently words and words occurring infrequently with words that occur more frequently. When changing the thresholds of the significance value according to the selected measure the graph re-sort the wordnet and clusters will emerge. Depending on the selected significance the graph clusters individually.

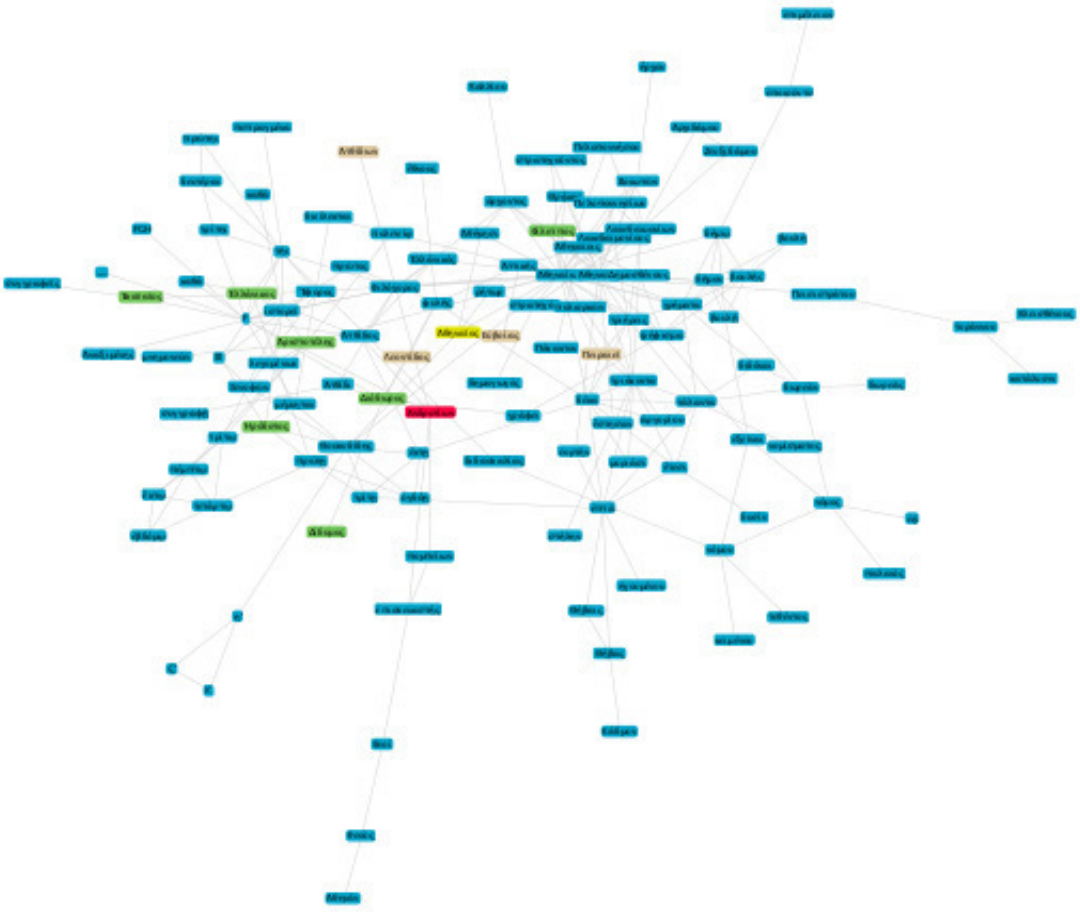

Figure 1

[1] The relevant cases are Ἀτθίς Ἀτθὶς Ἀτθίδος Ἀτθίδι Ἀτθίδα Ἀτθίδων Ἀτθίδας.

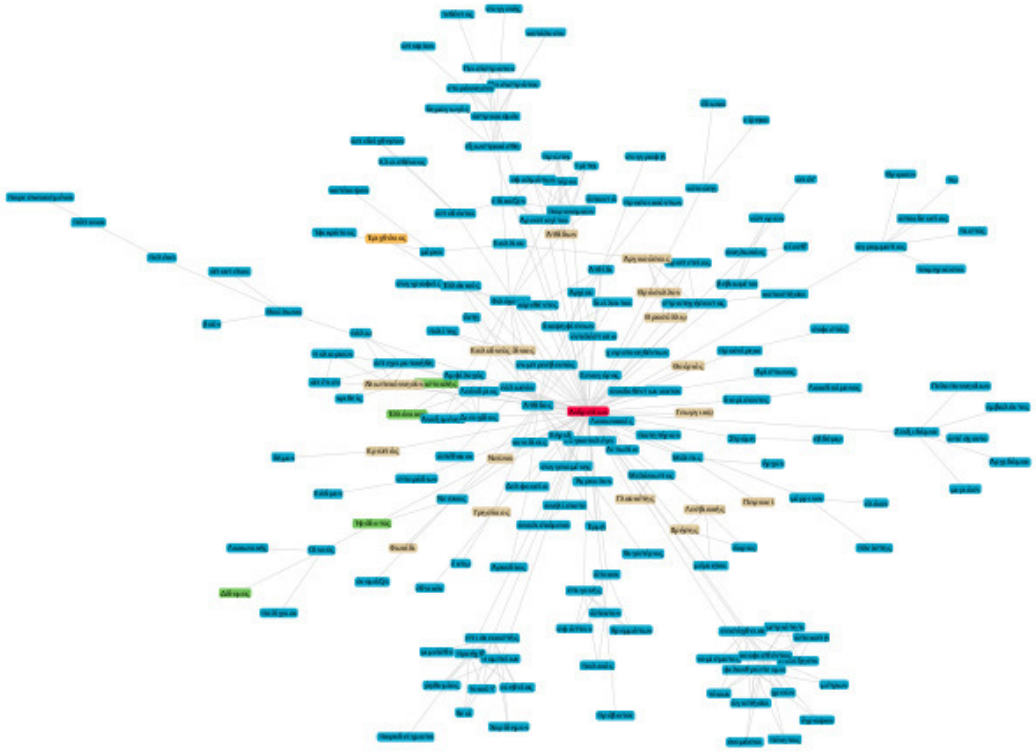

Figure 2

Figure 1 shows the co-occurrence graph of the name of the Atthidographer Androtion (Ἀνδροτίων, the red bar) with a threshold of a minimum significance of 84 based on Log Likelihood. Figure 2 shows the same graph but clustered with a threshold setting of 10.5 based on Mutual Information. The latter graph bears more named entities and proper names indicated by the green, brown and orange bars. Additionally it provides more word clouds, clusters especially at its periphery. For an inquiry of the contexts of the term Ἀνδροτίων within the TLG corpus this graph would be more appropriate than the one in figure 1.

Figure 3

Figure 3 depicts a more detailed snapshot of some of the clusters at the periphery of the graph of figure 2. The reason for the formation of these clouds is a heavy interrelation of the terms that they consist of. As these clouds appear at the periphery the relative independence to other co occurrents of Ἀνδροτίων is clearly indicated. It has been observed that each cloud is formed out of a single sentence that occurs several times in the TLG corpus. Reasons for this have been given earlier.

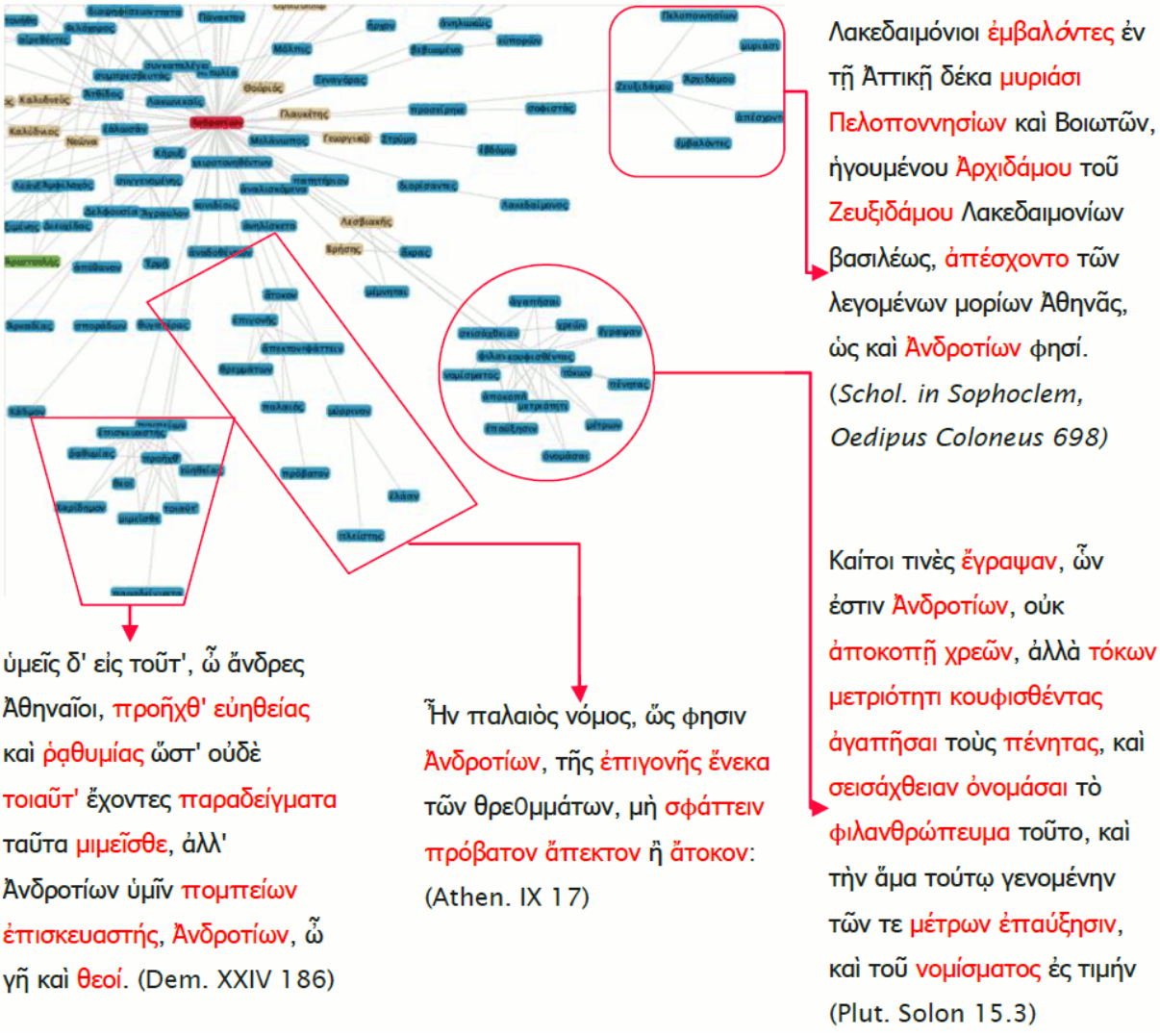

Figure 4

Figure 4 shows that the words in each of the clouds all appear in individual sentences. The corresponding terms have been coloured red. All contexts have Androtion as authority for information derived out of an Atthis except the trapezoid cloud on the lower left which indicates an allusion made by Demosthenes to the politician Androtion.

With clustering as a method of visualising semantic interdependencies between co-occurents, it is now possible to instantly review the context of a single word from inside the whole corpus. In the ancient languages Greek and Latin many words have multiple meanings depending mainly on the context of their appearance. In dictionaries we can find all levels of meaning of the most common words covered. However there is often a lack of that type of information for more infrequent words or proper names. One of these words is Atthis. What information about its use in Greek literary works can we extricate from the clusters?

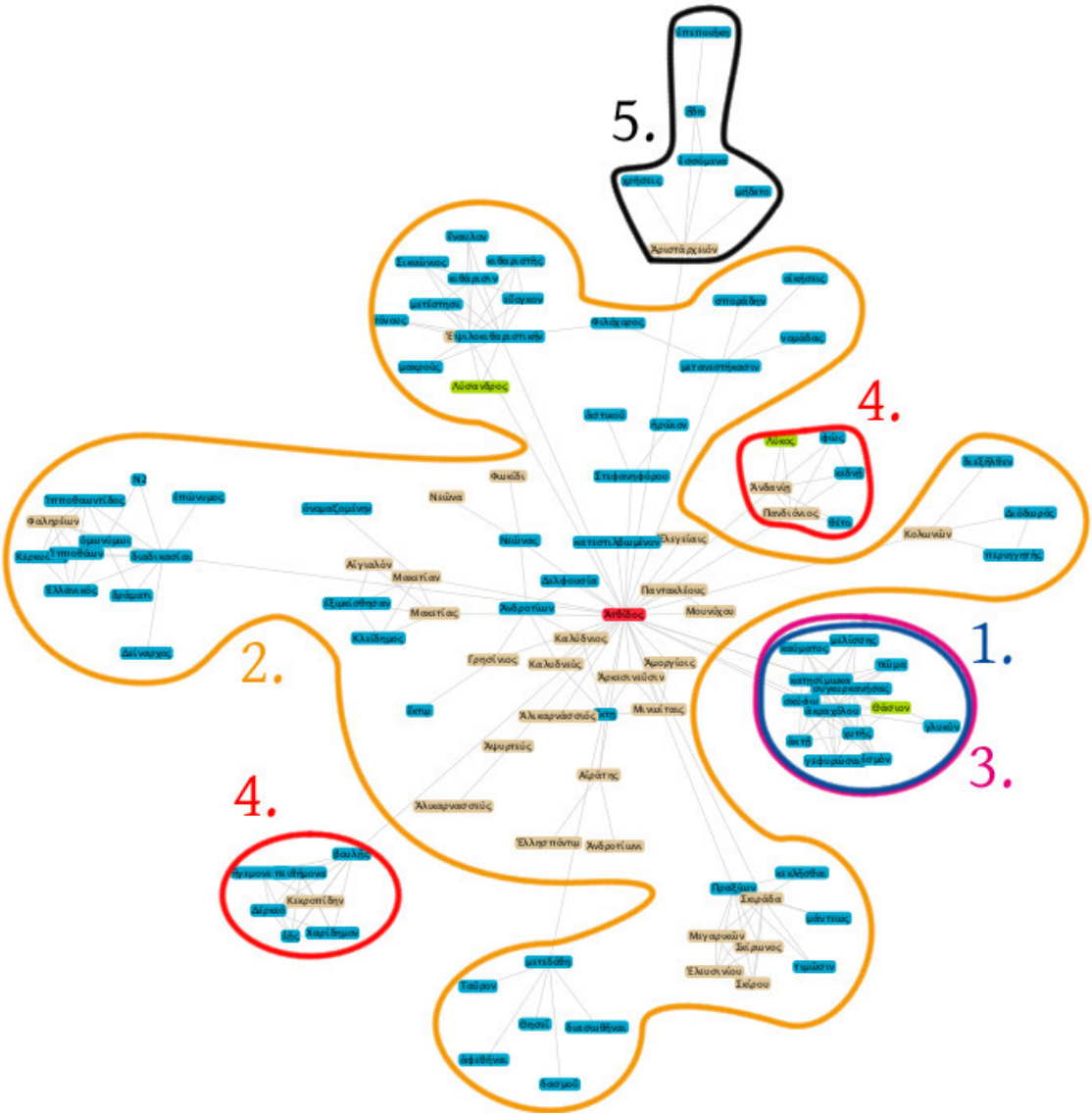

Figure 5

Figure 5 shows the graph of the genitive singular Ἀτθίδος at a threshold of 13.5 based on Mutual Information. Nice clusters enable us to determine its main contexts. In this grammatical form, Atthis appears most commonly in the corpus, with 276 occurrences in total. By inspecting all of them in detail, a list of five types of contexts can be made: The first meaning would be 'Attica' as a poetic or traditional denomination of the peninsular on which Athens is situated. This meaning applies to 23 of all instances in which the word occurs (blue). The second context reers to the book 'Atthis', probably a label for all books that dealt with the history of Attica introduced by the scholar and librarian Callimachus when indexing the content of Alexandria's library during the third century BC.[2] This meaning is covered 187 times and is therefore the major meaning of that word form (orange). The third function of the term is Atthis as referring to the mythical daughter of king Cecrops and the eponym for the peninsular, this could be observed in 16 cases (pink). The fourth use of the term 'Atthis' occurs as the adjective form of 'Attica' and as such it occurs 17 times (red). Because the case of the adjective depends on the case of its related object, one might find this meaning in all grammatical cases of Atthis. In its fifth meaning, occurring 28 times, 'Atthis' signifies the Attic dialect (black). Comparing the given contexts with the clouds shown in figure 5, it can

[2] Jacoby 1949, p. 84; Costa 2005, pp. 5.

be clearly seen that the attribution of the clouds to one of the contexts is possible and that clustering is therefore a good method to quickly line out contexts.

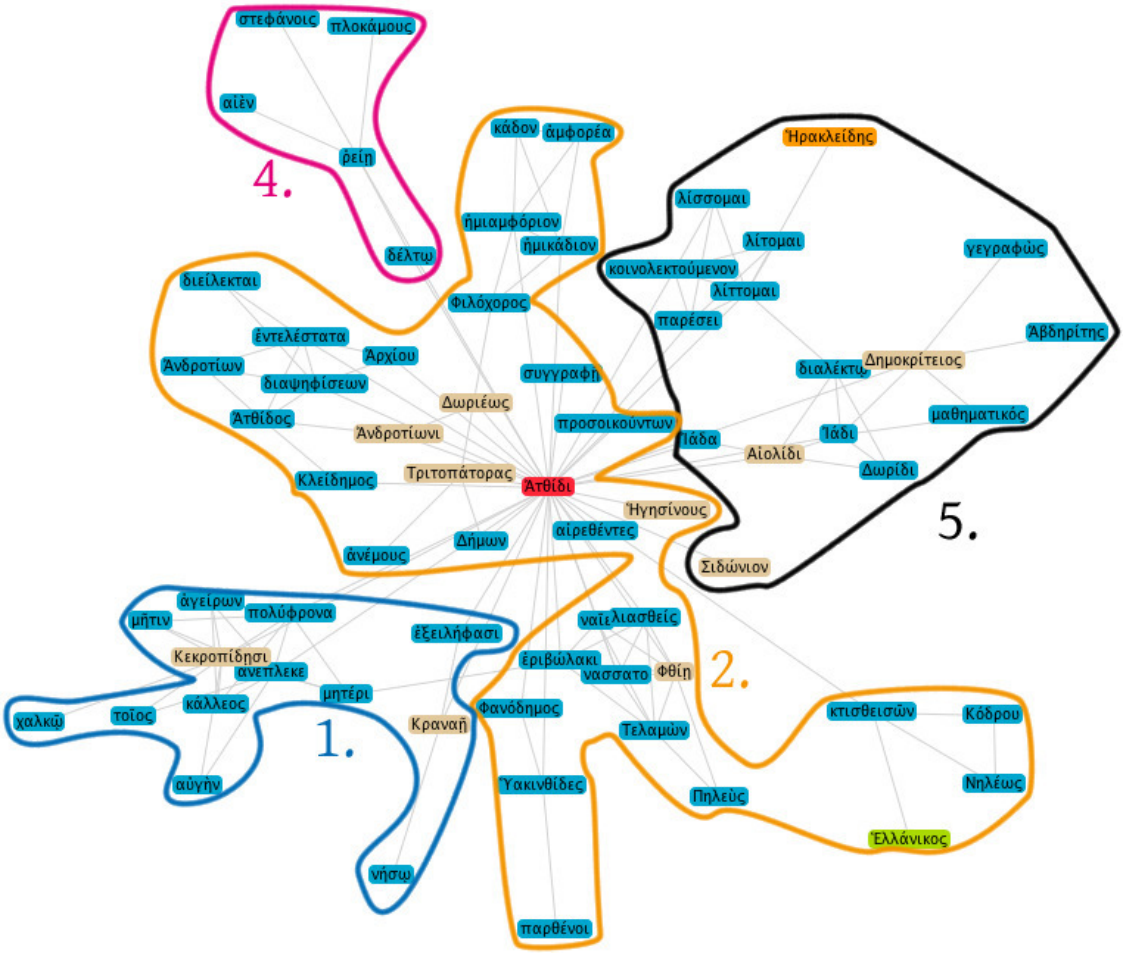

Figure 6

Figure 6 shows the contexts if Atthis appears in the dative singular (taken at 11.399 SigMI). Most of the meanings of the genitive could be found in the dative as well except the meaning of the eponym daughter. Additionally, a new context for the term can be detected, the ethnicon of Attica that refers to its inhabitants, but only two occurrences can be observed of this, too few to be displayed in this graph (blue). The major context again would refer to the book ‘Atthis’ with 42 instances.

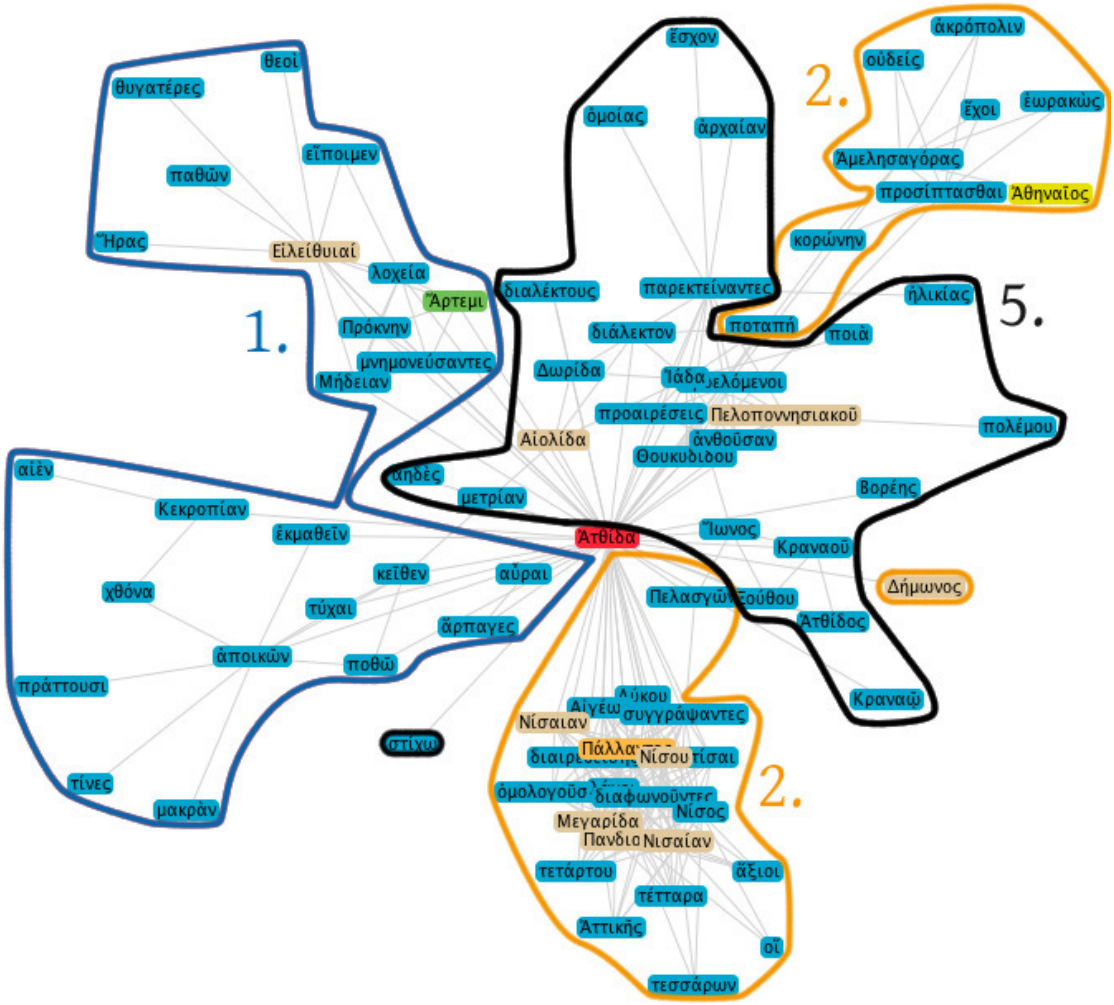

Figure 7

The clusters of the accusative singular show that the use of ‘Atthis’ in this form was most often for referring to Attica (18 times) and the Attic dialect (26 times). The two clouds referring to the book ‘Atthis’ are derived from only eleven instances.

Figure 8

The two contexts of figure 8, which depicts the genitive plural of Atthis, shows clearly that this form is used only as adjective and in the meaning of books about Attica. However one time it is used as dialect but that is not significant enough to be displayed (at SigMI 8.25).

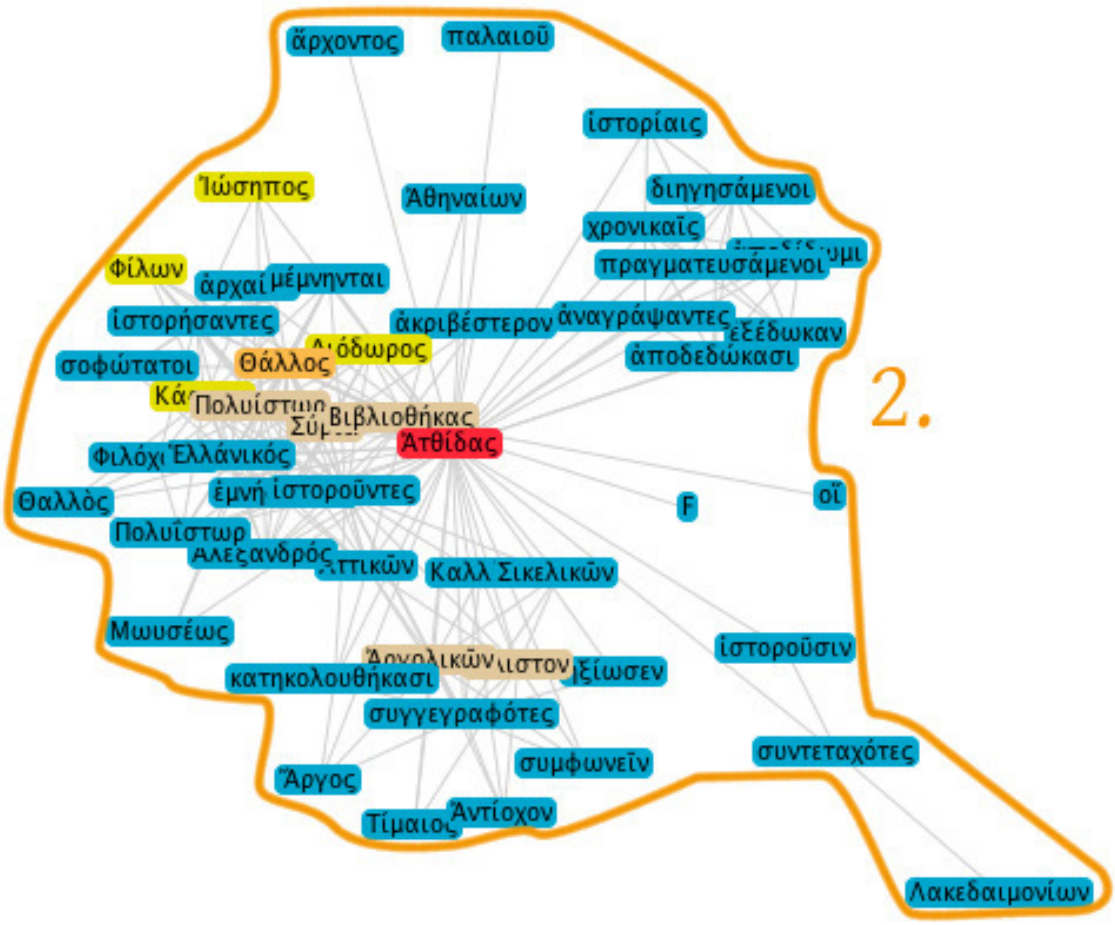

Figure 9

The graph in figure 9 is a clear example for the uniform context of accusative plural of Atthis. In 19 instances (of a total of 20) it refers to the books about Attica, and in one instance it means Attica, yet this last instance has been dropped out of the graph due to its insignificance.

In summarizing the evidence, one might conclude that the word Atthis bears different meanings for each individual grammatical appearance. The co-occurrence graph shows instantly the term's different meanings by grouping the evidence context-wise and is therefore very helpful for every day work as it is very common that one has to figure out the possible meanings of a certain word.

The following section deals with approaches of analysing a small size corpus like that of the Atthidographers using the Citationsgraph tool. If a tool that is designed to compare contents of an individual work with all other works of the corpus in order to look for striking similarities as evidence for text reuse, it is clear that the use of such a tool must be rethought when it is applied to an incomplete text, whose content may have been dispersed. Comparison as such will not work the way it generally does but nevertheless, as I will show, this tool can be very useful.

When dealing with a group of historians like the Atthidographers one is inclined to look which sources they used and who used them in turn as sources. These are important questions that help to classify the kind of work they produced. Especially the quality of their historical research has been often challenged by modern authors but probably not by the ancients (Jacoby 1949, 1954; contra: Schubert 2010). Here are some thoughts about how the Citationsgraph may help in answering these questions.

In order to create a profile of the individual traceable tradition of an Atthidographer one might use the Citationsgraph.

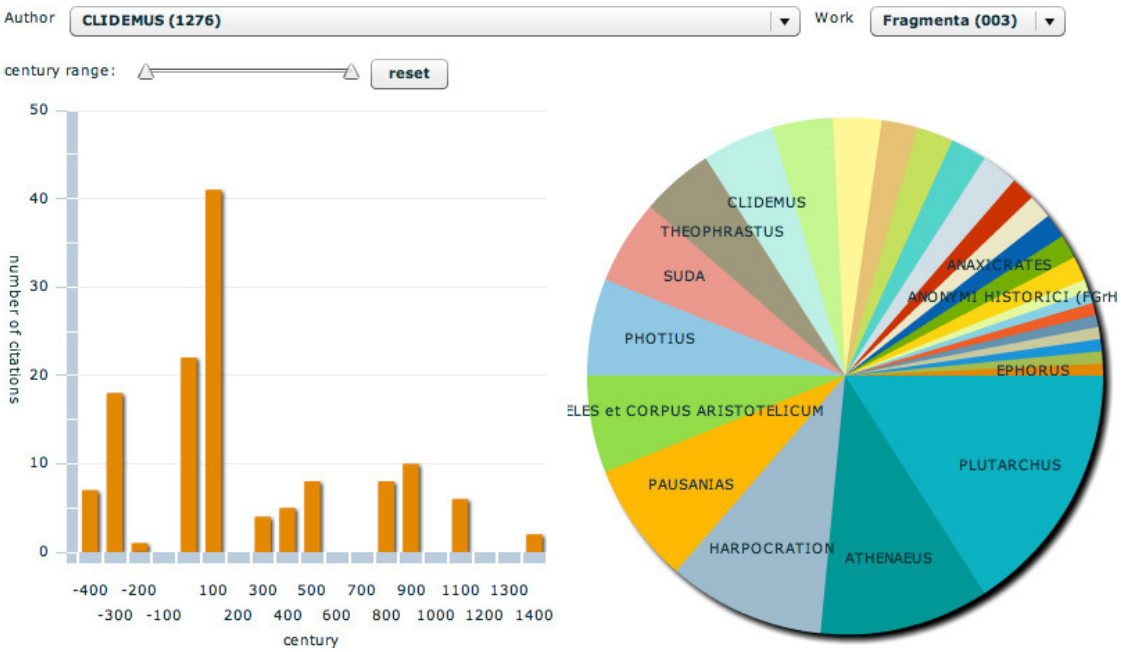

Figure 10

Figure 10 shows the citation profile for Clidemus. The Atthidographer wrote during the third century BC.[3] In the pie chart those authors whose works refer to Clidemus most often are easily visible. We might categorize three major groups of authors from that chart: a) Greek authors writing prose and that flourished during the days of the Roman empire like Pausanias, Athenaeus and Plutarchus; b) Lexicographers Harpocration, Photius and Suda who often depend on each other with Harpocration as the earliest source; c) the Philosophers Aristotle and Theophrastus. The histogram on the left reveals a chronologic overview of the re-users of Clidemus' work with the highest rates concerning the number of citations in the first centuries BC and AD, a time that alludes to group a). The significant decay of re-uses after that period may serve as evidence for the loss of the work and the remains of some epitomes.

[3] Jacoby 1954, pp. 58.

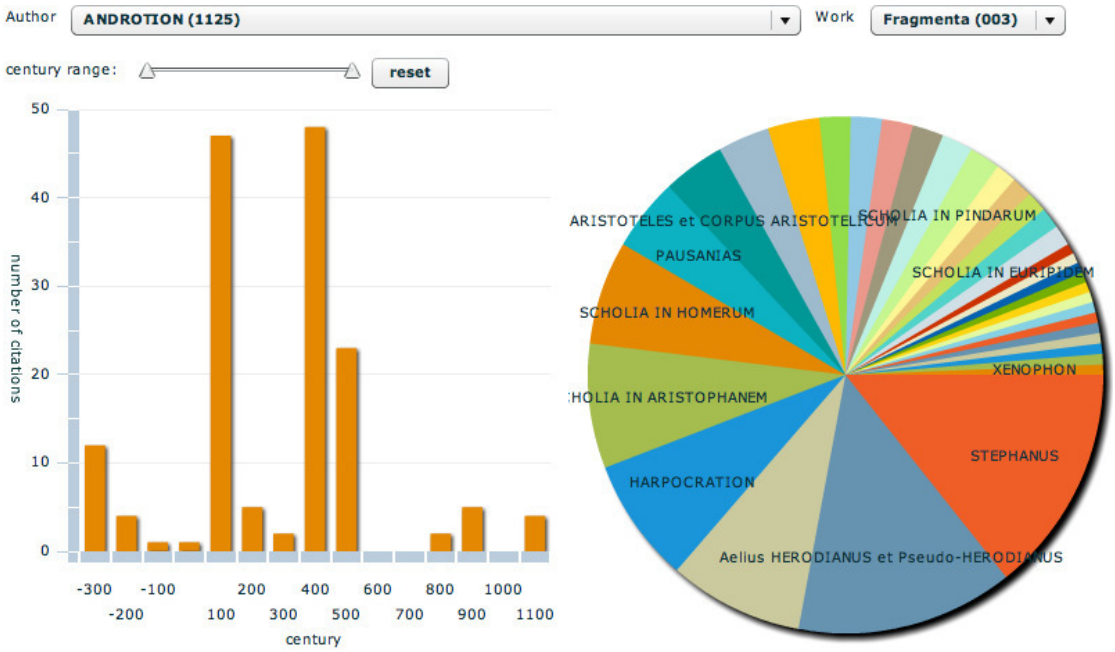

Figure 11

The profile for Androtion is shown in figure 11. Androtion wrote another Atthis a few decades after Clidemus.[4] He was also involved in the political matters of Athens for more then thirty years. His Atthis is known to contain a lot of detailed information concerning important constitutional or institutional changes in the history of the state. The visible reoccurrence of Aristotle in the profile of Androtion seems to contradict the earlier statement that only Clidemus has been used by the philosophers. It is commonly assumed that Aristotle used Androtion for quite significant portions of the Ἀθηναίων πολιτεία ("Constitution of the Athenians") that has been assigned to his corpus. This work is seen as a work of political history rather than a philosophical one. Differences to the profile of Clidemus can be easily seen. Only Pausanias is left for the category a) authors. The lexicographers remain as the only accordance with Clidemus. A new group d) are the grammarians who used Androtion for citing atticisms of ethnica. Category e) finally is also new and comprises Scholia (Aristophanes, Homer, Pindar, Sophokles, Euripides). This indicates that Androtion's Atthis yielded information concerning contemporary events of the plays of Sophokles, Euripides and Aristophanes. Those fifth century play writers were very important in the political propaganda of Athens of the fourth century and it is not surprising that Androtion provided information that has been included by the scholiasts in the manuscripts of the plays, as it was important for the interpretation of the verses.

[4] Jacoby 1954, pp. 90.

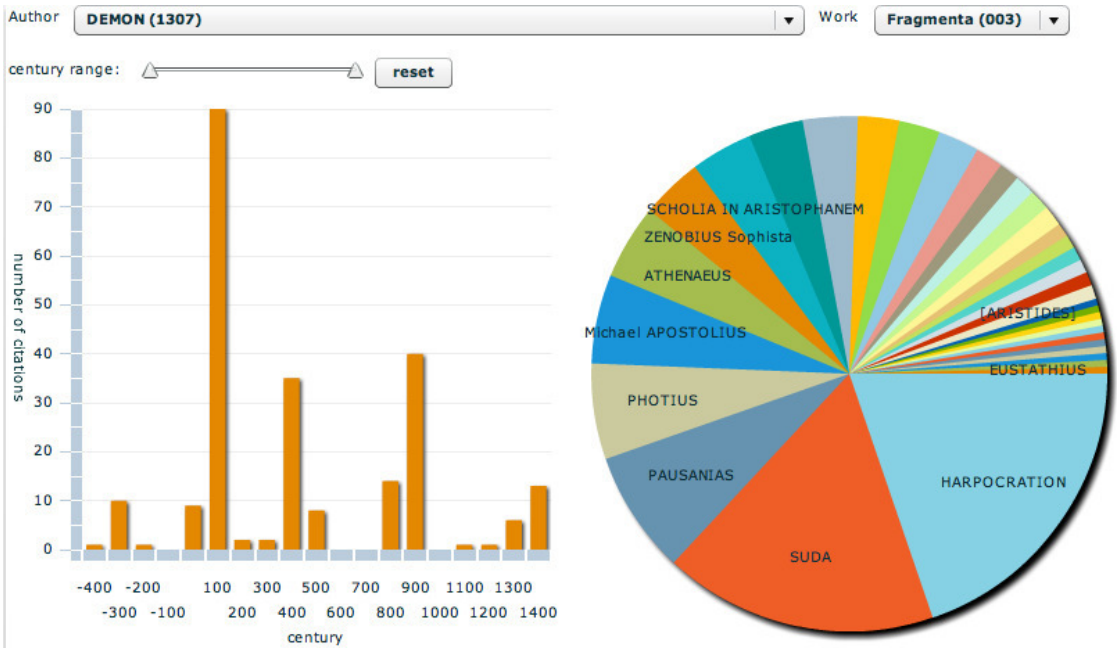

Figure 12

The profile of Demon in figure 12 yields interesting details about this quite mysterious Atthidographer. Not many references to his works have been survived and the most important Atthidographer Philochorus wrote his Atthis against Demon's. From category a) we find Athenaeus and Pausanias. The b) authorities are the major exploiters of Demons work. One may find also some proxies from e) but the most important medieval author was Michael Apostolius whose major work was about Greek proverbs. This indicates – and a further inquiry into the b) evidence had confirmed this – that Demon was more used for his work on proverbs than for his Atthis. However the number of citations of the first century AD is extraordinary.

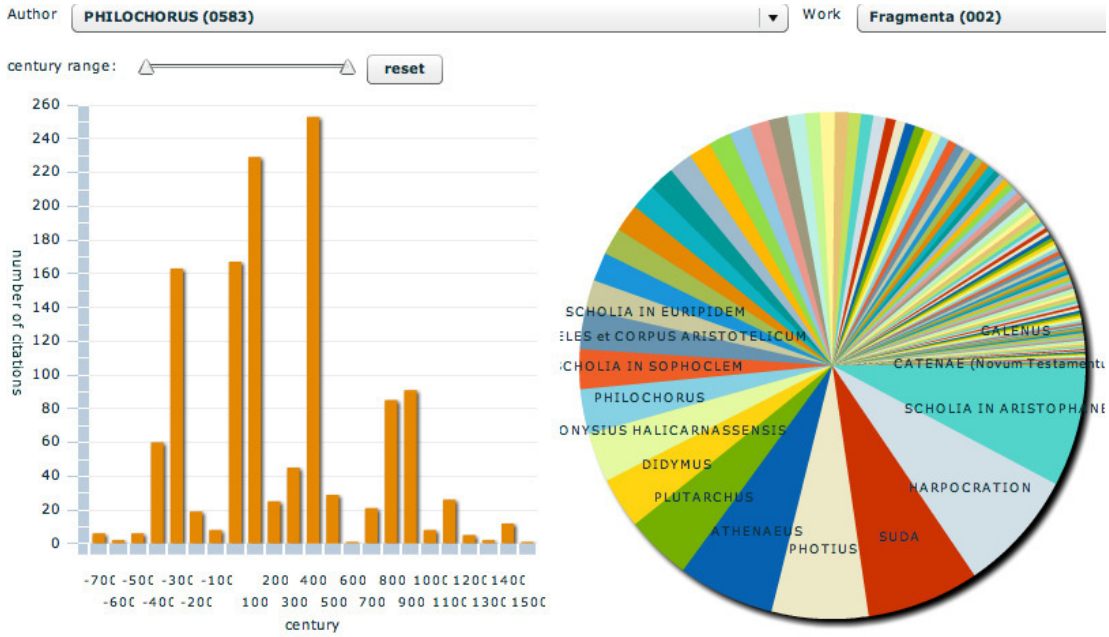

Figure 13

In figure 12 the profile of Philochorus shows a wide distribution across the centuries indicating that his work had been held famous throughout the time and at the first glance one is

inclined to deny the theory of a loss of his work few decades after its completion.[5] The main references belong to the categories a), b) and e) with Didymus as a famous scholiast whose thoughts about Demosthenes' speeches luckily have been preserved on papyrus (P. Berol. 9780). He cites mainly Philochorus as evidence for his arguments. Dionysius from Harlicarnassus is a further roman authority that we did not find among the other profiles.

One might conclude that by interpreting the instantly produced profiles of tradition it is hard to maintain the statement that all Atthides were somehow similar to one another.[6] The differences in the tradition are striking. We must rather think of the Atthides as differing significantly in terms of content and style than has been assumed hitherto. Perhaps only because they all dealt with Athens or Attica, they have been grouped under a single category.

With some slight changes the same tool that has been used to create the tradition profiles can create profiles of the sources of the Atthidographers.

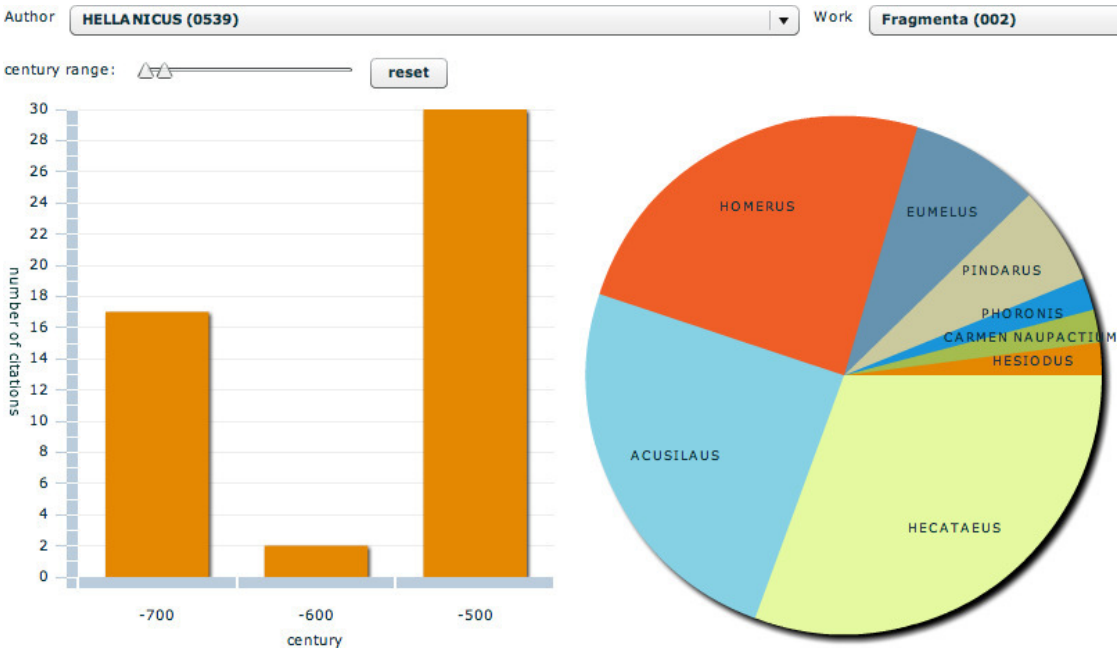

Figure 14

Figure 14 shows how this is possible. The chronological histogram of the references can be limited concerning the start and end date of the inquiry. In order to see possible sources of a certain author it is necessary to know the date when the authority in question has flourished. In our case, Hellanicus of Lesbus is focussed on. He lived during the fifth century and counts as the very first Atthidographer despite Pausanias' claims that Clidemus was the first (Pausan. 10.15.1). So the *terminus ante quem* must be set to the fifth century BC. All authors citing Hellanicus later than that date are thus omitted. The authors who are left must be somehow connected with Hellanicus' work. As has been stated earlier, the so-called fragments from the collections of fragmentary authors tend to be highly interpolated with other authors and the same pieces are assigned multiple times. That's why the results of the Citationsgraph are as it were polluted. On the other hand, these interferences might show that the authors mentioned at the same time also wrote about the same things. The sources of Hellanicus' Atthis are often disputed. From temple chronicles to oral tradition ranges the field of speculations about the origin of his work.[7] These inventions are mandatory since the *communis opinio* supposes Hellanicus to have written in the same manner as Clidemus and the others up to Philochorus, which means that he interwove mythical and historical information. What we can see in the profile shows us sources that are crucial for mythical parts.

[5] Jacoby 1954, pp 239.
[6] Jacoby 1949 & 1954 passim.
[7] See e.g. Jacoby 1949, pp. 215.

Well-known classics like Homer or Pindar, epics and poets as well as authors of theogony. In fact almost all known references to an Atthis of Hellanicus are made in context with some myth. Still we have evidence that he wrote also about his own days (e.g. Thuc. 1.97 = F Gr Hist 323a T8, or F26 and F27). The profile rather supports the theory that Hellanicus used available information and transformed it slightly to adapt it to new necessities of his patrons. This is a common practice in Greek historiography. However the theory cannot be proved based only on the profile but nevertheless the indication tackles common theories.

The paper has shown that it is very useful to re-examine the long lasting evidence of our research in ancient history when new methods and tools are available. The obstacles that occur when they get applied have not been omitted either. With the eAqua text mining tools, co-occurrence graph and Citationsgraph, it is less complicated to do this than before. However, further ideas must be developed in order to deal with the integration of these methods so they fit to the demands of the research. And so, quicker than ever, surveys and comparisons can be compiled revealing all the information for any author or any word that has to be analysed. Commonly uphold theories can thus be easily revised without putting too much effort in trying to reproduce already existing knowledge just for evaluation purposes and so the readiness for revisions even of huge evidence increases.

André Bünte
Lehrstuhl für Alte Geschichte, Historisches Seminar, Universität Leipzig
abunte@uni-leipzig.de

Literature

Costa, V. (2005) *Filocoro di Atene. I: I frammenti dell'Atthis*. Tivoli: Tored.

Dunning, T. (1993) Accurate Methods for the Statistics of Surprise and Coincidence. In: *Computational Linguistics* 19(1): 61-74.

Jacoby, F. (1949) *Atthis*. Oxford: Clarendon Press.

Jacoby, F. (1954) “A Commentary on the Ancient Historians of Athens (Nos. 323a-334), vol. 1:Text.” In: *Die Fragmente der Griechischen Historiker - Teil 3. Geschichte von Städten und Völkern (Horographie und Ethnographie)* ; b Suppl., Leiden: Brill.

Lenz, C. G. / Siebelis, C. G. (Hg.) (1811) *Philochori Atheniensis librorum fragmenta. Accedunt Androtionis Ἀτθίδος reliquiae*. Leipzig: Schwickert.

Lenz, C. G. / Siebelis, C. G. (Hg.) (1812) *Phanodemi, Demonis, Clitodemi atque istri atthidon et reliquorum librorum fragmenta - Accedit prolusio scholastica de atthidon scriptoribus, et additamentum ad Philochori fragmenta*. Leipzig: Schwickert.

Mulleri, C. / Mulleri, T. (Hg.) (1841) “Fragmenta Historicorum Graecorum - Hecataei, Charonis ..., Apollodori bibliotheca cum fragmentis.” In: *Fragmenta Historicorum Graecorum - Apollodori bibliotheca cum fragmentis*, 1:384-417. Paris: Firmin-Didot.

Rayson, P. / Garside, R. (2000) Comparing corpora using frequency profiling. In *proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*: 1-6.

Schubert, Ch. (2010) Formen der griechischen Historiographie: Die Atthidographen als Historiker Athens. In: *Hermes* 138(3): 259-275.

Worthington, I. (2010) *Brill's New Jacoby*. Leiden: Brill.