JDHCS 2010
Volume 1 Number 2

Unsupervised Detection and Visualisation of Textual Reuse on Ancient Greek Texts

Marco Büchler, Natural Language Processing Group, Institute of Mathematics and Computer Science, University of Leipzig
Annette Geßner, Ancient Greek Philology Group, Institute of Classical Philology and Comparative Studies, University of Leipzig
Thomas Eckart, Natural Language Processing Group, Institute of Mathematics and Computer Science, University of Leipzig
Gerhard Heyer, Natural Language Processing Group, Institute of Mathematics and Computer Science, University of Leipzig

Epigram

Lots of texts produce a large amount of text mining data that can easily be accessed by a powerful visual analytics component.

Introduction

Throughout the centuries most ancient texts have been lost or at least significantly altered due to the many dangers of frequent manual copying. These challenges include changes in dialect, the natural development of language, different textual interpretations, opinions, or beliefs and even censorship. In addition to comparing various old manuscripts, another way of handling this issue is to take a closer look at some text passages that have been quoted by other authors. Just as in modern literature, ancient authors have always made references to other works. By finding those textual reuses, it is possible to either verify or question the current edition of an ancient text. In this paper the linking of these text passages to each other is called a Reuse Graph G. Using a graph like this, it is possible to derive the importance of an author across the centuries by quantifying the degree of reuse of their text by other authors over time.

There have always been important ancient authors whose works and thoughts have been known to every educated reader and were frequently used by other authors. It is most likely that when the portion of oratory education grew, written sayings gained importance and thus text reuse grew significantly.[1]

Finding these quotations is not always a simple task for the modern reader, however, for quotation rules were not established until modern times. While it is difficult to describe the general methodology of text reuse in ancient Greek texts, it can be assumed at least for classical Greece that it was not common to name the referred author, as every educated reader would know whose thoughts had been brought into consideration.[2] Even when the referring author made clear that he had expressed another author's thoughts, there was no guarantee that this was done with exactly the same words.[3]

---

[1] van den Berg 2000, 14f.

[2] van den Berg 2000, 14f.

[3] Harbsmeier 2008, 48f.; Mülke 2008, 11f.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

This reason is why the detection of textual reuse is a very important task in Classical Studies, and much research has already been conducted in the past with a variety of computational approaches having been applied to ancient texts. Within this paper, an additional dimension to those text mining results is introduced: *Visual Analytics,* which is a new research field in computer science that is used to visualise massive datasets that are produced by technologies such as text mining. Using visualisations as shown in Figures 4a and 4b, a *micro view* (e. g. for ancient Greek philologists) is provided that can especially be used to investigate the discrepancies of one or more reused passages in terms of an original text. In further work, these systematically collected results can be reutilised to understand the variance interval of features for the detection of passages in a work written by an author without having the original texts – a dedicated task of authorship attribution. The visualisation shown in Figures 2a and 2b represents a *macro view* on textual reuse, it can be used to investigate more general aspects of textual reuse, such as for historians seeking to identify trends in *Neo-* or *Middle Platonism*.

Although it would be nearly impossible to read all ancient Greek literature and manually find all examples of text reuse, there have nevertheless been several attempts to do exactly this. Gerard Boter, for example, tried to collect all text reuse of Plato's work *res publica*.[4] Nonetheless this impressive work has also been criticised: "Users of this or any edition are warned that the textual variants presented by citations from Plato in later literature have not yet been as fully investigated as is desirable". This shortcoming, characterized by Kenneth Dover[5] is still existent and is unlikely to be corrected quickly with the help of traditional techniques of research.

Within this paper the terminology *reuse* or *textual reuse* has been deliberately chosen instead of *citation*, even if this term is quite often used in this context. This is caused by definitional ambiguity for in modern understanding a *citation*[6] is defined as a *textual reuse* that also includes a specific reference to a source text. If a proper reference is not given, the reuse is considered to be plagiarism.[7] Writing in ancient Greece and in other cultures like Chinese writing from the 19th century assumed a canon of well-known and accepted authors and their works. For this reason, specific authors and works are typically not referenced in all texts. From a modern perspective this would be considered plagiarism since the knowledge about a canon is not directly written down. To dissolve this ambiguity of citation, *reuse* or *textual reuse* has been chosen without any reference to source entries, especially since source references are irrelevant for unsupervised algorithms themselves.

General Overview and Related Work

In the field of text reuse much research has already been done and while it is impossible to address all relevant work some important aspects are summarised in this section. Scientifically, the linking of two text passages is formalized as a graph G=(V,E) consisting of a set V of vertices and a set E=VxV of edges between elements of V. The set V represents a non-overlapping corpus that is segmented into large linguistic units like sentences or paragraphs. This task can typically be done

[4] Boter 1989.

[5] Dover 1980, VII.

[6] Research Information Network 2009.

[7] Potthast et al. 2009.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

with linear costs of O(n). The set of edges E between two elements $v_i,v_j \in V$ represents pairwise links between two text passages. Computing those links is much more complex than defining the set V.

Trying to compute textual reuse by pairwise comparison is time-expensive due to a squared complexity of $O(n^2)$. While this is pragmatic when comparing smaller corpora such as the Dead Sea Scrolls with the Hebrew Bible,[8] with ancient Greek corpora like the *Thesaurus Linguae Graecae* (TLG)[9] that has about 5.5 million sentences, *3.025e13* comparisons would be necessary. Assuming that about 1000 comparisons can be done in a second, this process would approximately require a run time of almost 1000 years. Even if only all the sentences of an author such as Plato were compared with a corpus like the TLG, the processing time would still require about one year.

Reviewing more complex algorithms, most of them can be summarised as a two-step process:

- **Linking:** The first step is to link two passages as either directed or undirected. In a historical context it is often useful to highlight who has used texts from whom, but without metadata it is quite difficult to make a link directed. Typical approaches reduce the complexity in comparison to the above mentioned naive method of $O(n^2)$ to $O(n*log(n))$ which decreases the computation time dramatically.
- **Scoring:** After two text passages are linked, the second step is to score the similarity of the two linked passages.

In both of these steps, links of some passages are rejected. Depending on the text and the degree of textual reuse, there is a strong selection in the first step. With ancient Greek texts it can be observed that only one in 100 million possible linking candidates is considered an actual case of textual reuse. The scoring itself can be seen more as a fine-tuning (see section *Results and lessons learned*) that removes less similar sentences.

The *linking* step is divided into two strongly correlated sub-tasks, first a window size and then an algorithm need to be selected. While it depends on both the selected corpus and the research question, typically used observation windows include *sentences*,[10] *paragraphs*[11] and a *fixed word number window*.[12] For applications in the humanities, however, the choice of the window size will strongly depend on the following question: "*How was an author quoted?*". If there is a strong literal reuse then approaches using sentence segmentation or a fixed window are good choices. However, if a given piece of textual content is paraphrased or strongly mixed in with the referring author's own words, then a larger context like a paragraph is necessary; otherwise, the probability of a match decreases.

In the second step of the linking process, the link features are defined. Generally, there are three different clusters of approaches:

---

[8] Hose 2004.

[9] Pantella 2009.

[10] Hose 2004.

[11] Lee 2007.

[12] Mittler et al. 2009.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

1. **Words as features**: After all function words are removed, passages that have the same words are linked. The general idea of these approaches is to identify those passages of a text that have a significant common semantic density.[13]
2. **N-grams as features**: To extract textual reuse syntactically, several n-gram approaches for bi-grams and tri-grams exist. The key idea is to link units having a significant large overlap of n-grams.[14]
3. **Sub-graphs as features**: Graph-based approaches as shown in this paper deal with semantic relations between words. In the *Lexical Chaining* approach[15] that is often used for text summarisation[16] a *semantic construct* or a *semantic representation* of linguistic units is generated. When applying these approaches to a huge amount of text, an implicit feature expansion of paradigmatic word relations in terms of language evolution or different dialects is often observed. This is caused by the fact that these words are connected with the other words of a unit as well.

While the cluster of n-gram approaches is strongly focused on syntactical features, the approaches of both other clusters can also deal with textual reuse in a free word order.

To score a found link, a measure is used to compute the similarity of both linked units. Therefore the features themselves or the words of both units are taken to compute any kind of similarity. Besides, measures like overlap or the dice coefficient compute the similarity of two pairwise linked passages, while other measures like the city block metric, euclidean distance, or the Jenson-Shannon divergence[17] calculate the semantic distance between two units. The main difference between both clusters is that a similarity measure scores relevant links a high score whereas distance measures score a relevant link of two units as close as possible to zero.

Given a corpus C, a reuse graph G=(V,E) can be described by the following generalised algorithm.

```
1  V = segment_corpus(C) with v1, v2, ..., vn∈V, ∪vi=C and vi≠vj
2  for each vi∈V
3    Fi=train_features(vi);
4  for each vi∈V
5  for each fk∈Fi
6    ei=(vi,vj)∈E=select all vj containing feature fk
7  for each ei∈E
8    si=scoring(ei=(vi,vj)∈E; Fi; Fj);
9    if(si<threshold){E=E\{ei}}
```

Listing 1. Generalisation of a textual reuse algorithm consisting of 4 steps: Line 1: Segmentation of a corpus to linguistic units v_i (builds set V of a graph =(V,E)), lines 2-3: Training of features set F_i

[13] Mittler et al. 2009, Lee 2007.

[14] Hose 2004, Büchler 2008 & 2010.

[15] Waltinger et al. 2008.

[16] Yu et al. 2007.

[17] all Bordag 2007.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

for every unit v_i, lines 4-6: Linking process of units (builds initial set E of a graph =(V,E)) and lines 7-9: Scoring and removing of less significant edges (cleans set E of a graph =(V,E))

Properties and preprocessing on ancient Greek texts

All illustrated methods and results in this paper are based on the TLG, a comprehensive collection of Greek writers, including well-known authors like Diogenes, Galen and Plato. The corpus has been created and provided by the TLG research center at the University of California[18] and is today one of the most important resources when dealing with ancient Greek texts. The current version contains around 7200 works written by more than 1800 different authors in a time period of more than 1800 years. Since the origin of the digital corpus goes back to the 1970s, all of the texts and metadata are encoded in a binary format that is not a good basis for efficient text mining applications. Therefore a rather comprehensive tool chain of pre-processing steps was developed to deal with this issue.

Several specific tools were developed or adopted for this research (including an extractor for the text and its related metadata and a Beta Code to Unicode converter), largely due to the challenges of working with a strongly inflected language like ancient Greek and its different changes over this long period of time.

Sentence segmentation

As a first pre-processing step, a specially created rule-based sentence boundary detector splits the texts. To deal with extraneous information that is unimportant for the detection of text reuse (e.g. the markings of speaker roles), different lists of boundary marks are used in combination with abbreviation lists to enhance detection rate.

Tokenisation

Compared to modern languages, a more active tokenisation is applied. In addition to punctuation marks, all brackets of the Leiden Convention are removed.

As a result of this tokenisation, all TLG texts are segmented into 5,520,060 sentences with an average length of 13.51 words. Table 1 shows the resulting cumulative sentence length distribution.

Sentence length	<=5	<=10	<=15	<=20	<=25	<=30	<=35	<=40	>40
Cumulative distribution in %	29.63	51.82	68.40	79.39	86.48	90.99	93.92	95.64	100

Table 1. Cumulative distribution of sentence length in words

[18] Pantella 2009.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

Normalisation

Since the ancient Greek language is very rich in diacritical marks and several words exist in a variety of upper/lower case letter combinations, many different shapes of the same word can be found in the corpus. As an example the conjunction και can be found in the TLG in more than 15 different versions. Since many of these variants exist due to changes in writing or modern modifications of the original text (like the usage of lower case letters) a reuse detection based on these variants might ignore a large number of relevant text passages. A normalisation is therefore executed that internally reduces all words to a lower case representation and removes any diacritics. Table 2 shows the number of different spelling variants of some highly frequent words of the TLG.

word	τοῦ	πρὸς	τοῖς	κατὰ	τοῦτο	εἶναι	βασιλεία
Number of variants	15	8	8	21	10	15	14

Table 2. Number of word variants with identical normalised word form

Lemmatisation

Another class of variations of the same word is caused due to morphology, so all words were consequently analysed and internally reduced to their base form by using the morphological analyser Morpheus, which was developed by the Perseus Digital Library.[19] As Morpheus can also identify dialects, even dialectical variants are reduced to same base form.

Syntactical approach

The work in this paper presents both a syntactical approach that is based on a statistical n-gram expansion and a semantic approach that is based on a sentence segmentation as linguistic unit, the two of which are then compared in terms of their usability on the TLG corpus. This section will discuss the syntactical approach.

Starting with an n-gram of size 5, in every iteration all n-grams of length l of the previous iteration are taken to compute new statistically significant n-grams of size l+1. Statistically significant means that the new n-gram must have a log-likelihood score not smaller than 6.63 and a minimum n-gram frequency of 2. This step is iterated until no more n-grams can be computed.

Expanding significant n-grams in such a way has one benefit and one consequence. The benefit is that the longest common match of a reuse with the original text can be found. With this information available, visual access for philologists as shown in Figures 4a and 4b can be provided quite simply since the boundaries of an n-gram are determined by one of the following three causes: a) the beginning of a sentence, b) the end of a sentence or c) any kind of a differing word due to causes such as language evolution, dialect change, an inserted word or the boundaries of an embedded reuse within a larger sentence.

[19] Crane 2010.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

A negative consequence of this approach is that all common prefixes of the longest match that consists of at least 5 words are produced. Consequently, a post-processing step removes those prefixes. In addition, finding the prefix properties of those n-grams requires a frequency heuristic such as:

$$eps = \log_2\left(\frac{Frequency(x_1x_2...x_n)}{Frequency(x_1x_2...x_nx_{n+1})}\right)$$

Empirically, an epsilon between 0.1 and 0.2 yields the best results and only prefixes with a smaller score than *eps* are removed. A larger score indicates that there is at least one more unit referring to the same original text. However, this text passage may just have a smaller common longest n-gram match. If this formula were not applied such relevant links would be removed.

Intuitively, a conditional probability such as:

$$eps' = p(w_{n+1}|w_1w_2...w_n) = \frac{p(w_1w_2...w_nw_{n+1})}{p(w_1w_2...w_n)}$$

should be used instead of the aforementioned formula. However, in cases of $Frequency(w_1w_2...w_n) = Frequency(w_1w_2...w_nw_{n+1})$ a side effect of an artificial *eps* is computed. This is caused by the denominators N_n *and* N_{n+1} of both probabilities as shown in:

$$eps' = p(w_{n+1}|w_1w_2...w_n) = \frac{\frac{Frequency(w_1w_2...w_nw_{n+1})}{N_{n+1}}}{\frac{Frequency(w_1w_2...w_n)}{N_n}} \quad eps' = p(w_{n+1}|w_1w_2...w_n) = \frac{Frequency(w_1w_2...w_nw_{n+1})}{Frequency(w_1w_2...w_n)} * \frac{N_n}{N_{n+1}}$$

especially for smaller n-grams when both denominators can differ relatively strongly. For this reason, an artificial and non-constant error would be computed into eps.

Given a set of those longest matching n-grams, all sentences containing the same n-gram are pairwise compared for similarity. To compute the similarity of both linked units, the *dice coefficient* is used. Words of both sentences are then compared for a common overlap in relation to the words that could be overlapped.

Finally, the minimum number of n-grams of size 5 needs to be justified. In a humanities context it is important to make algorithmic models as simple as possible in order to increase acceptance. Thus, the size 5 is chosen, since every n-gram of this size is statistically significant. This can be concluded by computing an upper boundary for the statistical expectation of the n-gram using the independence assumption:

$$p(w_1w_2...w_n) = p(w_1) * p(w_2) * ... * p(w_n)$$

For simplification of this upper boundary, the probability of an n-gram is not assumed by n different but the same word (upper boundary). Having done this as in:

$$p(w_1w_2...w_n) = p(w_i)^n$$

the length *n* of this n-gram is gradually increased with the ultimate purpose being to get the left side of this equation in numerical problems. In detail this means that the probability $p(w_1w_2...w_n)$ is decreased by increasing the size of the n-gram until the statistically expected n-gram frequency $Frequency(w_1w_2...w_n)$ is smaller than 1.

To get the minimum length *n* of an n-gram satisfying this formula, the logarithm is drawn as in:

$$\frac{Frequency(w_1w_2...w_n)}{N_n} < p(w_i)^n$$

to derive the final equation:

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

$$1 < p(w_1)^n * N_n$$

$$log_2(1) < log_2\left(p(w_1)^n * N_n\right)$$

$$n > -\frac{log_2\left(N_n\right)}{log_2\left(p(w_i)\right)}$$

While N_n can be counted easily, a stronger focus needs to be given to the role of probability $p(w_i)$. In addition, function words play an important role when computing n-grams for natural language texts. In terms of the TLG corpus, the 100 most frequent words cover *40.1%* of all tokens and this has a strong influence on n-gram approaches. For simplification of the upper boundary, the mean of the 100 most frequent words—the 50th most frequent word—is applied on the derived formula having an probability $p(w_i)$=0.0029. As a result of this equation, $n=3.09$ is computed. Assuming the 50th most frequent word as an average probability for an upper boundary, n-grams of at least size 4 are required to fulfil the derived formula. Applying the most frequent and the 100th most frequent words to this formula as well, minimum n-gram sizes of *6.15* and *2.51* are calculated respectively.

Given these results, a minimum n-gram length of 4 would fit for most n-grams. A length of 5 is chosen, however, to increase the "statistical surprise". Given this minimum length of an n-gram, the minimum frequency of an n-gram is 2, since this is necessary for a textual reuse and further supports the "statistical" surprise. In contrast to typical computer science papers in the field of textual reuse, we have purposely decided to simplify the algorithmic model as described to only two parameters: minimum length of n-grams and minimum n-gram frequency. Taking into account that eAQUA[20] is a Digital Humanities project, all models need to be understandable by both classicists and computer scientists. Thus this simplification consolidates the model into one that is easy to understand while still satisfying all the statistical requirements of a quantitative approach.

Semantic approach

The syntactical approach discussed above is focused on literal textual reuse that can be applied on well-known authors such as Plato. Nonetheless, the assumption of simple copying is not the only possible method of textual reuse. This section accordingly focuses not on the method of textual reuse, but rather on the textual information that is being reused.

A piece of semantic information consists of words being associated. Formally, a piece of semantic information is described by sub-graph $G_{inf}=(V_{inf},E_{inf})$ of a semantic co-occurrence graph $G_{sem}=(V_{sem},E_{sem})$ with $V_{inf} \subset V_{sem}$ and $E_{inf} \subset E_{sem}$. A co-occurrence graph $G_{sem}=(V_{sem},E_{sem})$ is an association network of words of a corpus (as shown in Figures 1a and 1b) that is described by the set of unique words V_{sem} and the set of associations E_{sem} that are computed by a co-occurrence analysis.[21] Semantic co-occurrences are computed by observing all words in a dedicated window like a sentence

[20] Büchler et al. 2008.

[21] Evert 2005; Büchler 2008 & 2010.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

or a paragraph and measuring their statistical significance by measures such as the *log-likelihood ratio*[22] or *mutual information*.[23] A piece of semantic information $G_{inf}=(V_{inf},E_{inf})$ consists of a relatively small subset of words $V_{inf} \subset V_{sem}$ containing typically not more than 20 words and accordingly a strongly reduced set of associations $E_{inf} \subset E_{sem}$ as exemplified in Figures 1a and 1b.

To justify co-occurrences there are a number of different approaches, as described in detail in Büchler 2010. The most frequently used justification is attributed to the *Distributional Hypothesis*[24]: *the context surrounding a given word provides information about its meaning*. Given this definition, a context is a semantic profile of a word based on its co-occurrences.

The graph-based approach within this paper uses two main components:

1. **Co-occurrences**: Given a sentence-segmented corpus, a set $E_{sem,cooc}$ is computed by co-occurrence analysis.[25]
2. **Co-occurrence based similarity**: Given a set $E_{sem,cooc}$, a set $E_{sem,sim}$ is computed by contextual similarities of words having in $E_{sem,cooc}$ at least one common co-occurrence.[26] The set $E_{sem,sim}$ consists of two subsets $E_{sem,occur}$ and $E_{sem,not\text{-}occur}$ with the two properties $E_{sem,occur} \cap E_{sem,not\text{-}occur} = \varnothing$ as well as $E_{sem,occur} \cup E_{sem,not\text{-}occur} = E_{sem,sim}$. $E_{sem,occur}$ also contains all edges that occur directly in a lexical unit, while $E_{sem,not\text{-}occur}$ contains associations between words not directly occurring in a unit as a sentence. Based on the *Distributional Hypothesis,* this set mostly contains semantic relations as synonyms or cohyponyms.

Having computed both $E_{sem,cooc}$ and $E_{sem,sim}$, the intersection of both sets $E_{sem}=E_{sem,cooc} \cap E_{sem,sim}$ is built. Writing the same set of associations as $E_{sem}=E_{sem,cooc} \setminus E_{sem,not\text{-}occur}$ the impact of this intersection is more obvious. Building this intersection, a strongly collapsed graph $G_{sem}=(V_{sem},E_{sem})$ is generated, containing only associations between words that both occur together in a sentence and have a strong similar context. The collapsed graph G_{sem} contains clusters of subgraphs that will be called G_{inf} in this paper.

The intersection of $E_{sem,cooc}$ and $E_{sem,sim}$ can be viewed from two different points of view:

- $E_{sem,cooc}$**:** Caused by the intersection with $E_{sem,sim}$, all associations in $E_{sem,cooc}$ without significant contextual similarity are removed. Based on the fact that every word of a text reuse is part of the context of all other words, the contextual overlap needs to be significant.
- $E_{sem,sim}$**:** Caused by the intersection with $E_{sem,cooc}$, all associations in $E_{sem,sim}$ not occurring directly in the text are removed. Associations in $E_{sem,sim}$ are built by their overlap in their semantic profiles and not by their direct occurrence in the text. However, the common occurrence of two words is necessary if they are relevant in textual reuse.

---

[22] Dunning 1993.

[23] Church & Hanks 1989.

[24] Harris 1954.

[25] Buechler 2008.

[26] Bordag 2007.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

To explain this approach, a five sentence toy sample corpus is built based on a famous proverb of Wilson Mizner:

1. *Copy from one, it is plagiarism; copy from two, it is research.* [27]
2. *Plagiarism is not the same as copyright infringement.* [28]
3. *Plagiarism is to to copy from one but to copy from two is research.* [29]
4. *The concept of copyright originates with the Statute of Anne (1710) in Great Britain.* [30]
5. *In a legal context, an infringement Frers to the violation of a law or a right.* [31]

For simplification all words are handled in a case-insensitive way. Firstly, the co-occurrence graph $G_{sem}=(V_{sem},E_{sem})$ is computed. V_{sem} contains all words of this five sentence corpus. The set $E_{sem.cooc}$ contains, among others elements, some like *(copy, plagiarism)*, *(research, plagiarism)* and *(copyright, plagiarism)*. In contrast to that, the set $E_{sem.sim}$ has, among others elements, some such as *(copy, plagiarism)*, *(research, plagiarism)* and *(copyright, copy)*. The association between *copyright* and *copy* is built for example by the common co-occurrence with *plagiarism*. Intersecting $E_{sem.cooc}$ and $E_{sem.sim}$ finally to E_{sem}, associations such as *(copyright, infringement)* are removed since both words co-occur only in the 2nd sentence and have only *plagiarism* as a common content word. In the 4th and 5th sentences, however, completely disjoint contexts (based on content words) are given. On the other hand, an association such as *(copyright, copy)* is removed since both words do not occur in the same sentence (lexical unit).
As a final post-processing step, all function words and correlated associations are deleted. This could be done earlier in order to try and reduce the amount of data. Based on the TLG corpus, however, about 300 million co-occurrence are computed. Both removing stop words and ignoring words having a word frequency of 1 (Hapax legomenon) reduces only about 10% of the co-occurrences in each case. Consequently, there is no benefit to removing function words earlier in the process. The benefit of doing this as a post-processing step is that several sets of function words can be applied for selection without computing everything completely anew. Within this post-processing step it is only a selection of nodes and edges aiming to cluster the graph G_{sem}.

In Figure 1, two derivations of a sub-graph G_{inf} of G_{sem} are represented, having function words both included and excluded. In contrast to the syntactical approach explained earlier, both Figures emphasize the free word order of the semantic approach.

---

[27] Wilson Mizner

[28] http://en.wikipedia.org/wiki/Plagiarism

[29] Own rewritten text of the first sentence.

[30] http://en.wikipedia.org/wiki/Copyright

[31] http://en.wikipedia.org/wiki/Infringement

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

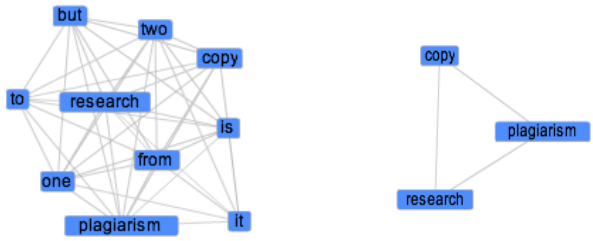

Figure 1. a) Left: Graph G_{inf} of the toy sample corpus including function words. b) Right: The same graph G_{inf} being reduced by all function words.

Comparing the graph-based with a simple word-based semantic approach, the difference can easily be given by an embedded reuse within another sentence. A word-based semantic approach extracts all content words as features. If a reuse is embedded into a large sentence then the content words of large sentences are implicit features of this reuse candidate too. However, the graph based approach removes those words by at least one of the both aforementioned properties.

Visualisation

Given a lot of unstructured data such as the ancient texts of the TLG corpus, a text mining approach typically produces more structured text mining data. In comparison to computer scientists evaluating algorithms, research in the humanities requires fully functional applications with easy access to lots of texts and massive sets of text mining data. The kind of access needed, however, depends on the particular research interests of the humanities scholar. Therefore in this paper two access methods are introduced for the field of Classical Studies. On the one hand, the view of *ancient Greek philologists* is labelled as the *micro view*. From this view, a focus on the variations of specific quotations is understandable. On the other hand, a *macro view* designed for *historians*, might focus on the changing usage of a quotation over a long time period. Investigating peaks in such a macro view, a single quotation does not matter since those peaks are given by significantly frequent sets of quotations implying an interest in this information during a particular time frame.

The macro view[32] is shown in Figure 2. The first graph visualizes the usage of Platonic quotations in time. Here what ancient philologists have pointed out about two important eras of Platonic philosophy can be seen clearly: The so-called Middle Platonism in the first and second century AD and the Neo-Platonism in the fourth and fifth century AD.

[32] http://www.eaqua.net:8080/portal/Citations.html?AuthorID=0059&WorkID=031

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

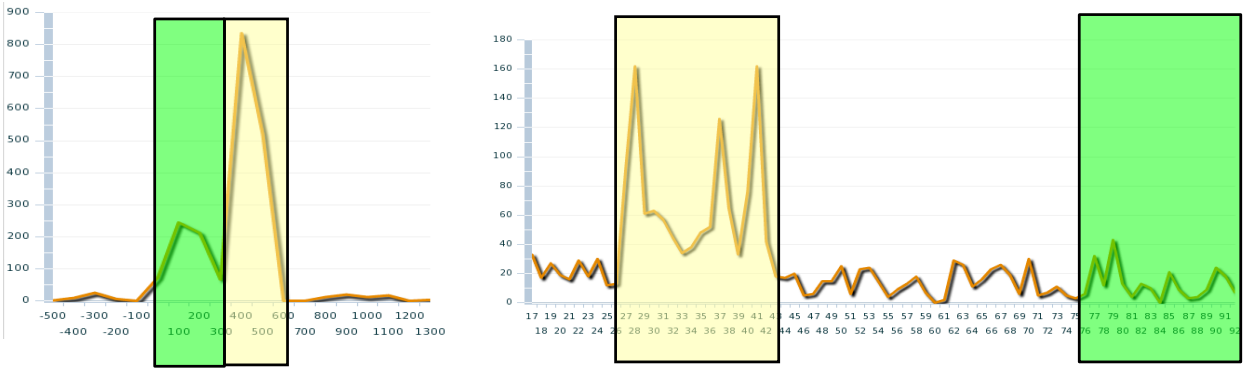

Figure 2. Macro view: Two of three screens of an interactive visualisation for quotation usage. a) Left: Century based distribution of literal quotations of Plato's Timaeus. b) Right: Text reuse distribution by Stephanus pages of Plato's Timaeus. The highest peak of the left picture is strongly correlated with the quotation usage of the pages 27 to 42 of the right picture: Neo-Platonism.

With help of a second visualisation (see Figure 2b) the most "famous" chapters of Plato's *Timaeus* can be determined by plotting how often single pages of his work have been reused.

As Figure 2 would likely be of stronger interest to historians, there is also a need for visualisations for researchers in the field of ancient Greek philology. As shown in Figures 3 and 4, a visualisation that highlights the differences in quotation usage is necessary. This is especially important if longer quotations are investigated.

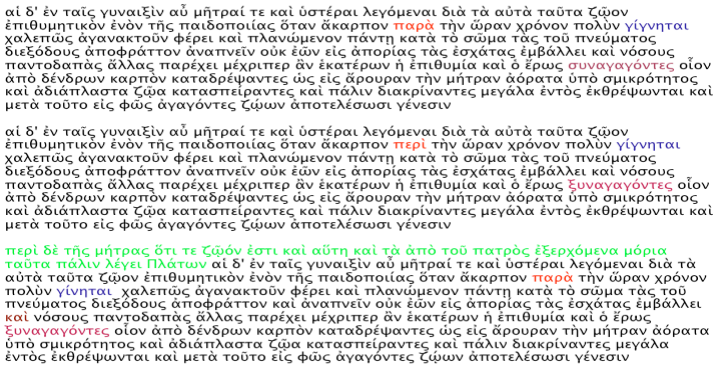

Figure 3. Original and two quoting sentences of Plato's Timaeus 91b7ff.

Comparing Plato's sentence about the female womb as a living being[33] with two quoting sentences without the help of any visualisation (see Figure 3), it is hard to recognise the differences between those long sentences. Only after having a very close look at them, can you point out small deviances like *περί* substituting *παρά,* the addition of a small word like *καί* or even just a missing single letter (*γινεται* instead of *γιγνεται*).

[33] *Timaeus* 91b7ff.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

This kind of reuse analysis is of great help, for it highlights the deviant words immediately by creating coloured branches that leave the blue main branch of the original sentence (see Figure 4).

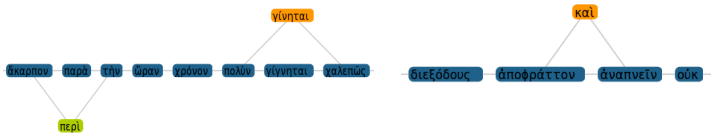

Figure 4. Highlighted differences of quotations (green, orange) in relation to original text of Plato (blue). a) Left: The orange word highlights the same word but including a language evolution of about 10 centuries. b) Right: An included word (orange) in the quotations is shown.

Results and lessons learned

Studying the usage of textual reuse, one research question is focused on the way text is reused. Figure 5 plots the percentage of references against the similarity by which the text is reused. Comparing reuse with similarity values of the range 0.8 to 1.0 in the green area, mostly literal reuse can be found. In terms of similarity measures between 0.8 and 0.9, the reuse percentage of perfect matches with a score of 1.0 is significantly smaller. This is caused by several influences as *language evolution*, *omitted* and *inserted words* or *different dialects* (see Figure 3).

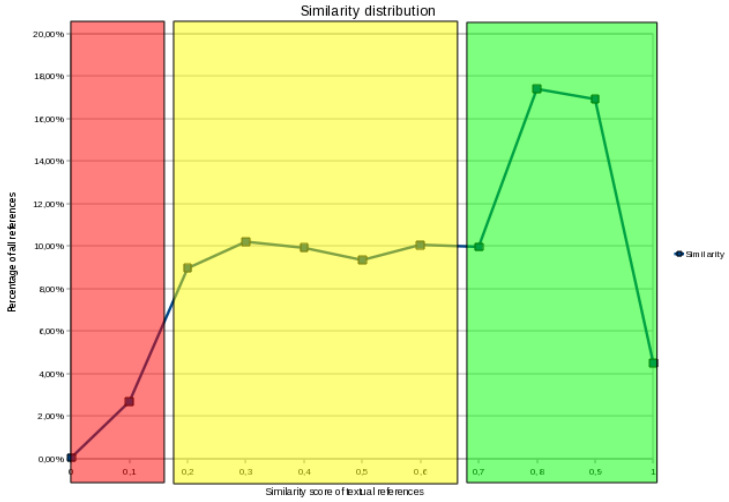

Figure 5. x-axes: Dice similarity scores. y-axis: Percentage of all found reuse candidates. Green indicates literally reused texts. Red is mostly noise which can be removed for sure. The yellow area is the undecidable range strongly depending on the genre of the underlying text.

Whereas the red area is noise, it is impossible to decide by an unsupervised algorithm whether the yellow area is a reuse or not. Comparing the results for Plato with some other authors, it is far easier to find quotations in philological texts. In historiography, for example, there are many more commonly used phrases like "*on land and on the sea*" or "*and they marched up to*", phrases that can't reasonably be considered as examples textual reuse. In contrast, a similarity threshold of 0.3 can be

applied with a high precision for philological texts, but on texts such as those of the Atthidographers 80% is noise.

Relating to the section *Syntactical approach*, it could be initiated with a minimum n-gram size of 4 if philological texts like those of Plato need to be analysed. The chance of linking sentences that are only aligned by a phrase, however, increases dramatically. The analysis of this paper demonstrates the fact that with ancient texts the quality of results mostly depends on the linking step of listing 1. As this is caused by the embedded reuse of text in a larger sentence, a relevant similarity score of as low as 0.3 can be observed.

To avoid having relevant links have small similarity scores, an additional task of finding the boundaries of a textual reuse could help. However, this is almost impossible with paraphrased texts. Even with a literal case of reuse, this task is often too difficult to be automated as shown by the following four beginnings of one original and five found quotations overall:

αἱ	*μῆτραί τε καὶ ὑστέραι λεγόμεναι ...*
αἱ δὲ ἐν ταῖς γυναιξὶν	*μῆτραί τε καὶ ὑστέραι λεγόμεναι ...*
αἱ δ' ἐν ταῖς γυναιξὶν αὖ	*μῆτραί τε καὶ ὑστέραι λεγόμεναι ...*
αἱ δ' ἐν ταῖς γυναιξὶ	*μῆτραί τε καὶ ὑστέραι λεγόμεναι ...*

Due to this problem, a deeper analysis with units smaller than a sentence is currently not explored in this paper. The segmentation of sense units within a sentence, however, can be built by the graph based approach.

Given a similarity threshold of 0.2 as shown in the red area of Figure 5, about 11.5% of all found reuse candidates are rejected as examples of textual reuse. Since all found candidates have an n-gram of at least 5 words in common, it's important to focus on which reuse candidates are selected by this threshold. Taking the similarity threshold of 0.2 and the assumption that reused texts have similar length, the dice coefficient can be simplified. Taking into account that at least 5 word are in common, there are sentences necessary consisting of more than 25 words to reject a link candidate by this similarity threshold. In relation to Table 1 – distribution of sentence length – only 13.52% of all sentences have a length of more than 25.

Interpreting the plots of Figure 6, it appears that authors tend to reuse text more freely the more distant they are from Plato in time. On the one hand it became harder and harder to get a proper manuscript of an original text, and on the other hand the Greek language itself evolved over time.

In contrast to the syntactical approach, the main benefit of the introduced semantic approach is the ability to ignore a set of words as function words. Based on this insight the number of phrase based misalignments of text passages can be reduced.

Regarding Figure 3, an implicit expansion of language evolution or different dialects as shown on the example *γινεται* instead of *γιγνεται* is done, and thus additional hits can be found.

Since a semantic representation per unit is generated, however, some new mismatches of text passages are linked based on the same key words, especially with longer sentences.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

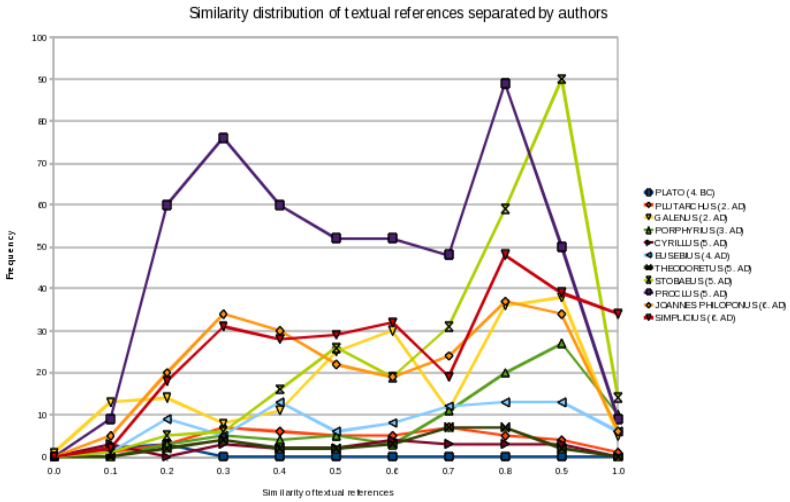

Figure 6. Different similarity distributions for some authors quoting Plato. x-axis: Degree of similarity computed by dice coefficient. y-axis: Number of quotations with within the defined similarity slice.

Summary

This paper demonstrates the approach of the eAQUA project for identifying textual reuse on ancient Greek texts with the help of text mining algorithms and visualisation of the results.

The regarding desideratum of research has been illustrated, the term *textual reuse* has been defined, two approaches have been outlined, and the genesis of the applied search algorithms has been explicated. Additionally, in the course of addressing preprocessing issues, four separate steps have been highlighted: sentence segmentation, tokenisation, normalisation and lemmatisation. Furthermore, this paper attempted to provide the reader with an understanding of the project's internal syntactical and semantic approach.

Another focal point was the research possibilities enabled by the visualisation of various results. The *micro view*, primarily aimed at classical philologists, provides a thorough examination of the text itself, especially regarding the variations in a text's "version history". The *macro view* delivers an overview of information that is connected to the search results: it looks at how many authors/works reused a particular text passage and how this changed depending on the epoch. Therefore the potential importance of this visualisation for ancient literature and history is quite clear. In addition, the benefits of the methodological approaches and visualisations presented here have been compared to manual research methods. Nonetheless, potential problems have also been addressed, such as the challenge of spotting textual reuse in historiographic texts, which are far more strongly pervaded by phrases (*on land and on the sea*) that can not be counted as textual reuse despite the similar wording. Future research will focus attention on these issues.

References

Bordag, S. 2007. Elements of knowledge-free and unsupervised lexical acquisition. PhD diss., Leipzig University.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

Boter, Gerard. 1989. *The textual tradition of Plato's Republic*. Leiden: Brill.

Büchler, Marco. 2008. *Medusa: Performante ttxtstatistiken auf großen textmengen: Kookkurrenzanalyse in theorie und anwendung*. Saarbrücken: VDM Verlag.

Buechler, Marco. Medusa Release Page. January 2006. http://aspra25.informatik.uni-leipzig.de/medusa/.

Buechler, Marco, Gerhard Heyer, and Sabine Gründer. 2008. eAQUA - Bringing modern text mining appraoches to two thousand years old ancient texts. *4th IEEE International Conference on e-Science*. Indianapolis: Indiana University.

Church, Kenneth Ward, and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16, no. 1 (March 1990): 22-29.

Crane, George R. 2009. Perseus Digital Library. http://www.perseus.tufts.edu/hopper/ (accessed July 12, 2009).

Dunning, Ted. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19, no. 1 (March 1993): 61-74.

Evert, S. 2005. The statistics of word concurrences: Word pairs and collocations. PhD diss., Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Harbsmeier, Martin. 2009. Zitat oder paraphrase? Zwei Homer-Stellen in Platons *Politeia*. In Ursula Gärtner, Ute Tischer (Hgg.), *13. Aquilonia. Beiträge präsentiert zum 13. Jahrestreffen der Klassischen Philologie in Ostdeutschland*, Frankfurt, 17-50.

Harris, Zelig S. 1954. Distributional structure. *Word* 10 (1954): 146-162.

Hose, Ron. 2004. Investigation of sentence level text reuse algorithms. http://www.cs.cornell.edu/BOOM/2004sp/ProjectArch/DeadSea/index.html (accessed October 29, 2009).

Lee, John. 2007. A computational model of text reuse in ancient literary texts. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics:* 472-479. Prague: Association of Computational Linguistics:

Mittler, Barbara, Jennifer May, Peter Gietz, and Anette Frank. 2009. HRA4 QuotationFinder - Cluster Asia and Europe. *Universität Heidelberg*. http://www.asia-europe.uni-heidelberg.de/de/forschung/heidelberg-research-architecture/hra-projects/quotationfinder.html (accessed January 11, 2010).

Mülke, Markus. 2008. *Der autor und sein text: Die verfälschung des originals im urteil antiker autoren*. Berlin: Walter de Gruyter.

Plato. 1980. *Symposium*. Ed. Kenneth Dover. New York: Cambridge University Press.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

Pantella, Maria. 2009. Thesaurus Linguae Graecae. University of California, Irvine. http://www.tlg.uci.edu (accessed July 12, 2009).

Potthast, Martin, Benno Stein, Andreas Eiaselt, Alberto Barrón Cedeño, and Paulo Rosso. 2009. Overview of the 1st international competition on plagiarism detection. *3rd PAN Workshop: Uncovering Plagiarism, Authorship and Social Software Misuse:* 1-9. San Sebastian: CEUR Workshop Proceedings.

Research Information Network. 2009. Communicating knowledge: how and why researchers publish and disseminate their findings. Research Information Network. http://www.rin.ac.uk/our-work/communicating-and-disseminating-research/communicating-knowledge-how-and-why-researchers-pu (accessed January 11, 2010).

van den Berg, Wim. 2000. Autorität und schmuck: Über die funktion des zitates von der antike bis zur romantik. In *Instrument Zitat: über den literarhistorischen und institutionellen Nutzen von Zitaten und Zitieren*, by Klaus Beekman and Ralf Grüttemeier: 11-36. Amsterdam: Rodopi.

Waltinger, Ulrich, Alexander Mehler, and Gerhard Heyer. 2008. Towards automatic content tagging: Enhanced web services in digital libraries using lexical chaining. Eds. José Cordeiro, Jouaquim Filipe, and Slimane Hammoudi. *4th International Conference on Web Information Systems and Technologies: 231-236*. Funchal: INSTICC Press.

Yu, Lei, Jia Ma, Fuji Ren, and Shingo Kuroiwa. 2007. Automatic text summarization based on lexical chains and structural features. *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing:* 574-578. Qingdao: SNPD.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago