Detection of Citations and Textual Reuse on Ancient Greek Texts and its Applications in the Classical Studies: eAQUA Project

Büchler, Marco
mbuechler@eaqua.net
Natural Language Processing Group, Institute of Mathematics and Computer Science, University of Leipzig, Germany

Geßner, Annette
agessner@eaqua.net
Ancient Greek Philology Group, Institute of Classical Philology and Comparative Studies, University of Leipzig, Germany

Heyer, Gerhard
gheyer@eaqua.net
Natural Language Processing Group, Institute of Mathematics and Computer Science, University of Leipzig, Germany

Eckart, Thomas
teckart@eaqua.net
Natural Language Processing Group, Institute of Mathematics and Computer Science, University of Leipzig, Germany

"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 (Dover, 1980) is still existent and is unlikely to be corrected quickly by traditional research techniques. Textual reuse plays an important role in Classical Studies research. Similar to modern publications, classical authors used the texts of others as sources for their own work. In ancient texts, however, a less stronger form of word by word citation can be observed. Additionally, the complexity of ancient resources disallows fully manual research.

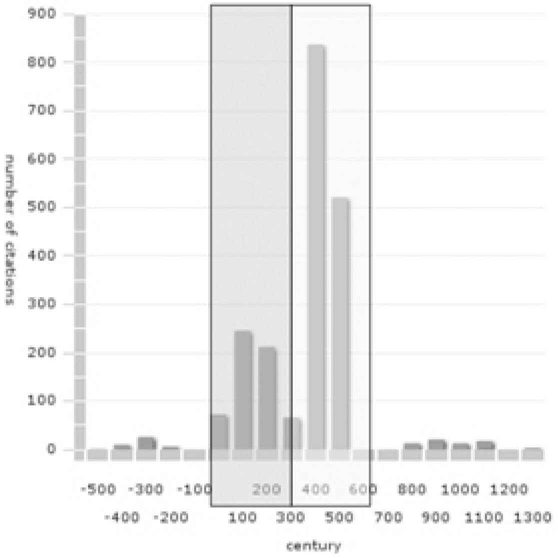
From a bird's eye view there are different points of view to the problem of textual reuse implying different research interests (Büchler and Geßner, 2009):

- A **Computer Science** perspective focuses on algorithms (*technical view*): Which algorithm is better than others? The scope of this research is wide ranging and also relates to plagiarism detection in modern texts like theses at universities (Potthast et al., 2009).
- A **Historian** is interested in more complex correlations (*macro view*). For this kind of work a dedicated user interface is necessary to figure out relations between e.g. chapters of a book and their citation usage on a timeline.
- The research interests of a **Classical Philologist** focus on the textual differences between the original text and its variants in citations (*micro view*). These varying requirements necessitate designing different user interfaces for these three kinds of researchers.

Within the eAQUA project we are investigating the reception of Plato as a case study of textual reuse in ancient Greek texts. Our research is carried out in two steps. On the *technical level*, we firstly extract word by word citations. This is achieved by combining syntactical ngram overlappings (Hose, 2009 and Büchler, 2008) and significant terms for several of Plato's works. In the second step the constraints on syntactic word order are relaxed. This is done by combining text mining and information retrieval techniques. A graph based approach is then introduced that can deal with free word order citations. The key concept is not syntactically based, but focuses on the semantic level to extract the relevant *core information* of a used citation. Then the information is represented as a formal graph that is similar to the *Lexical Chaining* approach (Waltinger et al. 2008) that is often used for text summarisation (Yu et al. 2007). On the one hand syntactical and semantic approaches are only used to select reuse candidates with a small set of uncommon matching words within a citation. On the other hand, a complete pairwise comparison of all of the nearly 5.5 million sentences in the TLG corpus would require approximately 1000 years due to the squared complexity of O(n^2) that was used for example to compare the Dead Sea Scrolls with the Hebrew Bible (Hose, 2009). For this reason, an intelligent pre-clustering of relevant reuse candidates is needed. Such a divide and conquer strategy reduces the complexity dramatically. Whilst the second step only increases the degree of free word order, in the third step the algorithm is expanded by similarly used words like *go* and *walk*. Those candidates are computed by similar cooccurrence profiles. The three levels briefly described above are only one dimension of reuse exploration. Other relevant dimensions that will be discussed are the *degree of preprocessing* as well as the *visualisation* of textual reuse in terms of citations.

In the field of preprocessing the main focus lies on *tokenisation* (more active tokenisation is needed with ancient texts than on modern languages), *normalisation* (reducing all words internally to a lower-case representation without diacritics) and *lemmatisation* (reducing all words internally to a word's base form). This dimension can speed up the algorithm and also improves the results for strongly inflected languages like Ancient Greek.

Leaving the technical point of view of computer scientists, the research of Classicists includes both an application of a *macro view* for Historians as well as one for the *micro view* of Classical Philologists. The visualisation dimension of textual reuse is important since text mining approaches typically generate a huge amount of data that can't be explored manually. This is shown in Fig. 1. Whilst the light grey area marks Neoplatonism (about 5. AC) the grey ranges highlight Middle Platonism (about 2. AC). Taking Plato's *Timaeus*, one can clearly identify that both phases of Plato's reception (see Fig. 1 – top) are based on different "chapters" of *Timaeus* (bottom).



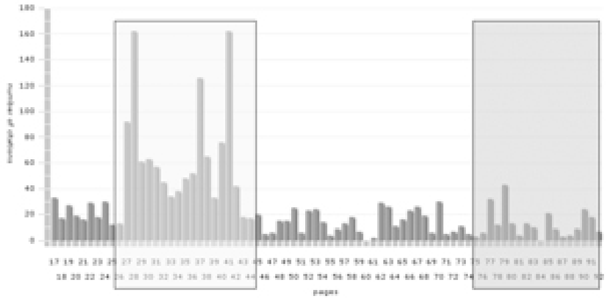


Fig. 1. Macro view: Screen of an interactive visualisation for citation usage. Citation distribution by Stephanus pages of Plato's Timaeus. The highest peak of the first picture is strongly correlated with the citation usage of the pages 27 to 42 of the second picture: Neo Platonism.

As Fig. 1 is of stronger interest for Historians, there is also a requirement for a visualisation for researchers from the field of Classical Greek Philology. As shown in Fig. 2, a visualisation highlighting the differences in citation usage is necessary. This is especially important if longer citations are investigated.

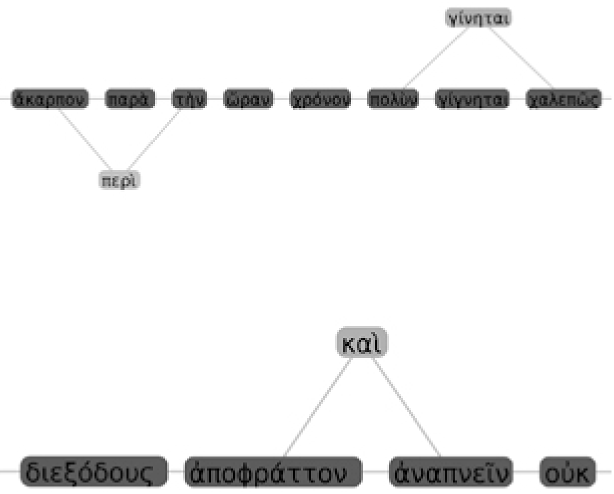


Fig. 2. Micro view: Highlighted differences of citations (green, orange) in relation to original text of Plato (blue). Top: The orange word highlights the same word but including a language evolution of about 10 centuries. Bottom: An included word (orange) in the citation is shown.

Additionally, it will be demonstrated how to detect different editions of the same original text. Such completely unsupervised approaches are important to investigate the scientific landscape of text digitisation. Furthermore, the relation to modern plagiarism detection will be given as well as the importance of building modern representative corpora since especially web corpora typically contain several duplicates of the same text.

In the evaluation section different results related to the comparison of various approaches on several text genres will be shown. An example of those results is given by contrasting citations of Plato's work with the textual reuse of the Atthidographers. Whilst citations of Plato can be extracted quite well by the syntactical approach even with very low similarity thresholds, the same approach works with an accuracy smaller than 20% for textual reuse of the Atthidographers.

Finally, results of a still in progress manual evaluation will be presented relating to the question of how and why a passage was cited.

References

Büchler, M. (2008). *Medusa: Performante Textstatistiken auf großen Textmengen: Kookkurrenzanalyse in Theorie und Anwendung*. Saarbrücken: VDM Verlag Dr. Müller.

Büchler, M. and Geßner, A. (2009). 'Citation Detection and Textual Reuse on Ancient Greek texts'. *2009 Chicago Colloquium on Digital Humanities and Computer Science.* Argamon, S. (ed.). Chicago.

Dover, K. (1980). *Plato: Symposium.* Cambridge: Cambridge University Press.

Hose, R. (2009). *CS490 Final Report: Investigation of Sentence Level Text Reuse Algorithms*. `http://www.cs.cornell.edu/BOOM/2004sp/ProjectArch/DeadSea/index.html` (accessed 29 October 2009).

Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A. and Rosso, P. (2009). 'Overview of the 1st International Competition on Plagiarism Detection'. *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09).* Stein, B., Rosso, P., Stamatatos, E. Koppel, M. and Agirre, E. (ed.). CEUR-WS.org, pp. 1-9.

Waltinger, U., Mehler, A. und Heyer, G. (2008). 'Towards Automatic Content Tagging: Enhanced Web Services in Digital Libraries Using Lexical Chaining'. *4th Int. Conf. on Web Information Systems and Technologies (WEBIST '08), 4-7 May, Funchal, Portugal.* Cordeiro, J. and Filipe, J. and Hammoudi, S. (ed.). Barcelona: INSTICC Press, pp. 231-236.

Yu L., Ma, J., Ren, F. and Kuroiwa, S. (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 (SNPD 2007).* V. 2, pp. 574-578.