UNIVERSITÄT LEIPZIG

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

Marco Büchler[*], Annette Geßner[+], Gerhard Heyer[*], Thomas Eckart[*]

eAQUA Project

[*]Natural Language Processing Group,
Institute of Mathematics and Computer Science,
University of Leipzig, Germany
[mbuechler|gheyer|teckart]@eaqua.net

[+]Ancient Greek Philology Group,
Institute of Classical Philology and Comparative Studies,
University of Leipzig, Germany
agessner@eaqua.net

Abstract: „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, VII) is still existent and is unlikely to be corrected quickly by traditional research techniques of research. Textual reuse plays an important role in research of Classical Studies. Similar to modern publications authors have been using texts of others as source for the own work. However, in ancient texts stronger word by word citations can be observed. Additionally, the complexity of Ancient resources disallow a full manual research.
From a bird's eye perspective there are different points of view to the problem of textual reuse implying different research interests (Buechler et al. 2009):

- A **Computer Science** perspective focuses on algorithms (*technical view*): Which algorithm ist better than others? The scope of this research is widely ranged e. g. up to plagiarism on modern texts like thesis on 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 the timeline.
- The research interests of a **Classical Philologist** focus on the textual difference between the original text and its variants from the citations (*micro view*). Caused by those requirements there are different necessities in designing user interfaces for this kind of researchers.

Within the eAQUA project we investigate the reception of Plato as a case study of textual reuse on ancient Greek texts. Our research is carried out in three steps. On the *technical level*, we firstly extract word by word citations. This is being done by combining syntactical n-gram overlappings (Hose, Buechler 2008) and significant terms for several of Plato's works. In the second step the constraints on syntactic word order are being relaxed. This is being done by combining text mining and information retrieval techniques. A graph based approach is introduced which 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 being similar to the

Lexical Chaining approach (Waltinger et al. 2008) which is often used for text summarisation (Yu et al. 2007). On the one hand syntactical and semantic approaches are only used for selecting reuse candidates with a small set of uncommon matching words within a citation. On the other hand, a complete pairwise comparison of all about 5.5 million sentences in the TLG corpus would need approximately 1000 years caused by squared complexity of $O(n^2)$ which was used e. g. to compare the Dead Sea Scrolls with the Hebrew Bible (Hose). For that reason, an intelligent pre-clustering of relevant reuse candidates is needed. Such a divide & 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 co-occurrence profiles. The three levels shortly 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 on 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 focus both an application of a *macro view* for Historian as well as another one for the *micro view* of Classical Philologist. The visualisation dimension of textual reuse is important since text mining approaches typically compute a huge amount of data which can't be explored manually. This is shown in figure 1. Whilst the yellow area marks the Neoplatonism (about 5. AC) the green ranges highlight the Middle Platonism (about 2. AC). Taking Plato's Timaeus, one can clearly identify that both mentioned phases of Plato's reception (figure1 – left) are based on different "chapters" of the Timaeus.

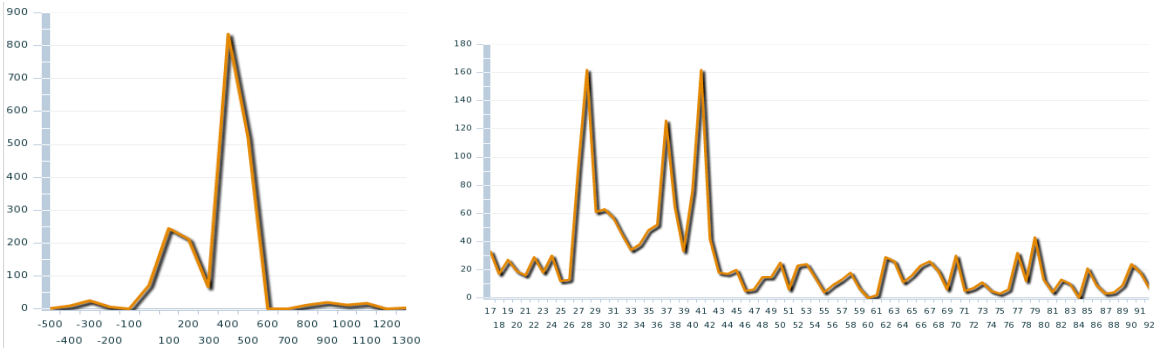

Fig. 1: Macro view: Two of three screens of an interactive visualisation for citation usage. left: Century based distribution of literal citations of Plato's Timaeus. Right: Citation distribution by Stephanus pages of Plato's Timaeus. The highest peak of the left picture is strongly correlated with the citation usage of the pages 27 to 42 of the right picture: Neo Platonism.

As figure 1 is of stronger interest for Historians, there is also a requirement of a visualisation for researchers from the field of Classical Greek Studies. As shown in figure 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). Left: The orange word highlights the same word but including a language evolution of about 10 centuries. Right: 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 scope to modern plagiarism detection will be given as well as the relevance to build modern representative corpora which are necessary since especially web corpora typically contain several duplicates of the same text.

In the evaluation section different results related to comparison of different approaches on different text types like literary classification will be shown. An example of those results is to be given by contrasting citations of Plato's work with the textual reuse of the Atthidographers. Whilst citations of Plato can be extracted quite good 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.

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

References

[Buechler 2008] Büchler, M. *Medusa: Performante Textstatistiken auf großen Textmengen: Kookkurrenzanalyse in Theorie und Anwendung*, Vdm Verlag Dr. Müller, 2008. ISBN-10: 3639011252

[Buechler et al. 2009] Büchler, M., Geßner, A. *Citation Detection and Textual Reuse on Ancient Greek texts*, In S. Argamon (eds), 2009 Chicago Colloquium on Digital Humanities and Computer Science, Chicago (Nov. 2009).

[Dover 1980] Dover, K., *Plato: Symposium*, Cambridge University Press: Cambridge, 1980.

[Hose] Hose, R. *CS490 Final Report: Investigation of Sentence Level Text Reuse Algorithms*. http://www.cs.cornell.edu/BOOM/2004sp/ProjectArch/DeadSea/index.html Last accessed: Oct, 29th 2009.

[Potthast et al. 2009] Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A. and Rosso, P. *Overview of the 1st International Competition on Plagiarism Detection*. In Benno Stein, Paolo Rosso, Efstathios
Stamatatos, Moshe Koppel, and Eneko Agirre, editors, SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), pages 1-9, September 2009. CEUR-WS.org. ISSN 1613-0073.

[Waltinger et al. 2008] Waltinger, U., Mehler, A. und Heyer, G., *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, José Cordeiro and Joaquim Filipe and Slimane Hammoudi (Eds.), INSTICC Press, Barcelona pp.231-236, 2008.

[Yu et al. 2007] Yu L., Ma, J., Ren, F., Kuroiwa, S., *Automatic Text Summarization Based on Lexical Chains and Structural Features*, snpd, vol. 2, pp.574-578, Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD 2007), 2007.