

Citation Detection and Textual Reuse on Ancient Greek texts

Marco Büchler, Annette Geßner

eAQUA Project

Natural Language Processing Group,
Institute of Mathematics and Computer Science,
University of Leipzig, Germany
mbuechler@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 (Plato Symposium, Cambridge, 1980, VII) is still existent and is unlikely to be corrected quickly with the help of traditional techniques of research. Textual reuse plays an important role in research of Classical Studies. Similar to modern publications authors are using texts of others as source for the own work. However, in ancient texts stronger word by word citations can be observed.

Within the eAQUA project we investigate the reception of Plato as a case study of textual reuse on ancient Greek texts. Our research in eAQUA is carried out in three steps. First we extract word by word citations. This is being done by combining n-gram overlappings 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. On the one hand, a positional inverted list is used for selecting only reuse candidates with a small set of non common 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 about 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. For that, an intelligent pre-clustering of relevant reuse candidates is needed. Such a divide & conquer strategy dramatically reduces the complexity. 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 is 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 improve the results for strongly inflected languages like Ancient Greek.

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). Taken Plato's *Timaeus*, it can be clearly identified that both mentioned phases of Plato's reception (figure1 – left) are based on different “chapters” of the *Timaeus*.

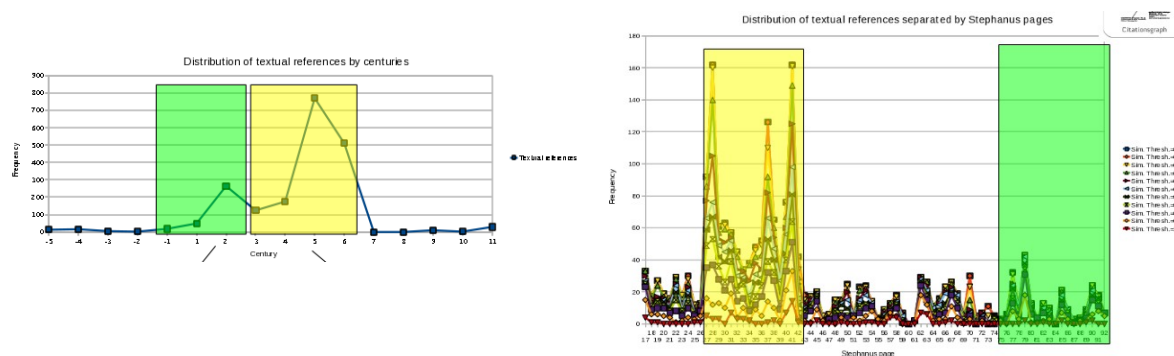


Fig. 1: left: Century based distribution of literal citations of Plato's *Timaeus*. Right: Citation distribution by Stephanus pages of Plato's *Timaeus*

As the 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 of highlighting the differences in citation usage is necessary. This is especially important if longer citations are investigated.

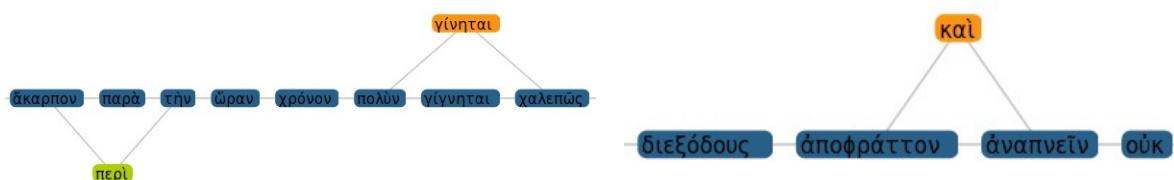


Fig. 1: Highlighted differences of citations (green, orange) in relation to original text of Plato (blue). Left: The orange word highlights the same word but including a language evolution of about 10 centuries. Right: An included word (orange) in the citation is shown.

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 relevance to build modern representative corpora which is necessary since especially web corpora typically contains several duplicates of the same text.

Within the Chicago Colloquium we will discuss the technical realisation and efficiency of the above mentioned dimensions and apply them in the field of Plato's aftermath. Based on substantial experience of an ongoing collaboration between researchers of Classical Studies and Computer Science we shall also reflect on the different approaches to working with text.