# THE DIGITAL CLASSICIST 2013

BULLETIN OF THE INSTITUTE OF CLASSICAL STUDIES SUPPLEMENT 122

DIRECTOR & GENERAL EDITOR: JOHN NORTH

DIRECTOR OF PUBLICATIONS: RICHARD SIMPSON

# THE DIGITAL CLASSICIST 2013

## EDITED BY STUART DUNN AND SIMON MAHONY

**INSTITUTE OF CLASSICAL STUDIES**
**SCHOOL OF ADVANCED STUDY**
**UNIVERSITY OF LONDON**

**2013**

The cover image is of a torso of Pothos (Roman 1st century BC – 1st century AD) in the Museu Calouste Gulbenkian, Lisbon, Portugal.


ISBN 978-1-905670-49-9



Designed and computer typeset at the Institute of Classical Studies.

Printed by Short Run Press Limited, Bittern Road, Exeter EX2 7LW.

This volume is dedicated to the memory of two people whose untimely death marks a great loss, both personally and to our communities.

Elaine Matthews (died 26 June 2011): one of our esteemed contributors, ambassador and advocate of the Digital Humanities and the place there for Classics, thank you for all your many contributions to scholarship, to this volume, and your generous words on the cover of the earlier *Digital Classicist* (Ashgate 2010) volume.

Gerhard Brey (1954-2012): a valued friend, colleague, and collaborator with whom we shared intellectual ideas as well as coffee and biscuits. Gerhard was always willing to seek out new areas of 'interest' and so could be willingly called upon to review chapters in this and the earlier Ashgate volume.

# TABLE OF CONTENTS

## ACKNOWLEDGEMENTS

The editors would like to thank the Institute for Classical Studies, and specially the Deputy Director and Administrator Olga Krzyszkowska, for their continued support and generosity in hosting and supporting the Digital Classicist seminars. Thanks are also due to all the members of our community who have presented papers at our seminars and conference panels as well as all those who have come along to listen and support these events.

We are grateful to the following scholars for comments, criticism, and advice on individual chapters in this volume; it is through their input that we are able to ensure the high quality of the final work: Chris Blackwell; Gabriel Bodard; John Bodel; Kalina Bontcheva; Gerhard Brey; Hugh Cayless; Graeme Earl; Michael Fulford; Sebastian Heath; Tim Hill; Kathryn Piquette; Dot Porter; Julian Richards; Matteo Romanello; Charlotte Roueché; Melissa Terras; Notis Toufexis; Charlotte Tupman; Michelle Wienhold.

# ABSTRACTS

Andrew Bevan *Travel and interaction in the Greek and Roman World. A review of some computational modelling approaches* pp. 3-24

Inferring dynamic past behaviours from the static archaeological record is always a challenge, but computational and quantitative techniques can be helpful. In particular, they can provide useful insight on patterns of movement and interaction, by better characterising existing archaeological evidence, suggesting simple models of mobile decision-making or proposing expected patterns against which the observed record can be compared. This paper reviews the range of modelling options now available for understanding the movement and interaction behind the archaeological and historical record. There are increasing opportunities not only to pick and choose between different modelling approaches, but also to integrate them in a more theoretically and practically satisfactory way.

Vince Gaffney, Phil Murgatroyd, Bart Craenen, and Georgios Theodoropoulos *'Only individuals': moving the Byzantine army to Manzikert* pp. 25-43

Traditionally, history has frequently emphasized the role of the 'Great Man or Woman', who may achieve greatness, or notoriety, through the consequences of their decisions. More problematic is the historical treatment of the mass of the population. Agent-based modelling is a computer simulation technique that can not only help identify key interactions that contribute to large scale patterns but also add detail to our understanding of the effects of all contributors to a system, not just those at the top. The Medieval Warfare on the Grid project has been using agent-based models to examine the march of the Byzantine army across Anatolia to Manzikert in AD 1071. This article describes the movement model used to simulate the army and the historical sources on which it was based. It also explains why novel route planning algorithms were required in order to surmount problems with standard solutions.

Elton Barker, Leif Isaksen, Nick Rabinowitz, Stefan Bouzarovski, and Chris Pelling *On using digital resources for the study of an ancient text: the case of Herodotus's 'Histories'* pp. 45-62

Involving the collaboration of researchers from Classics, Geography, and Archaeological Computing, and supported by funding from the AHRC, *Hestia* aims to enrich contemporary discussions of space by developing an innovative methodology for the study of an ancient narrative, Herodotus's *Histories*. Using the latest digital technology in combination with close textual study, we investigate the geographical concepts through which Herodotus describes the conflict between Greeks and Persians. Our findings nuance the customary

topographical vision of an east versus west polarity by drawing attention to the topological network culture that criss-crosses the two, and develop the means of bringing that world to a mass audience via the internet.

-In this chapter we discuss three main digital aspects to the project: the data capture of place-names in Herodotus; their visualization and dissemination using the web-mapping technologies of GIS, Google Earth, and Timemap; and the interrogation of the relationships that Herodotus draws between different geographical concepts using the digital resources at our disposal. Our concern will be to set out in some detail the digital basis to our methodology and the technologies that we have been exploiting, as well as the problems that we have encountered, in the hope of contributing not only to a more complex picture of space in Herodotus but also to a basis for future digital projects across the Humanities that spatially visualize large text-based corpora. With this in mind we end with a brief discussion of some of the ways in which this study is being developed, with assistance from research grants from the Google Digital Humanities Awards Program and JISC.

Marco Büchler, Annette Geßner, Monica Berti, and Thomas Eckart

*Measuring the Influence of a Work by Text Re-Use* pp. 63-79

Over the centuries an incredible amount of ancient Greek texts have been written. Some of these texts still exist today whereas other works are lost or are available only as fragments. Without considering intentional destruction, one major question remains: why did some texts remain and others get lost? The aim of this chapter is to investigate this topic by trying to determine the influence of certain ancient Greek works through detecting text re-use of these works. Text re-use measures if and how an author quotes other authors and in this chapter we differentiate between *re-use coverage* and *re-use temperature*.

Tobias Blanke, Mark Hedges, and Shrija Rajbhandari

*Towards a virtual data centre for Classics* pp. 81-90

A wide variety of digital resources have been created by researchers in the Classics. These tend to focus on specific topics that reflect the interests of their creators; nevertheless they are of utility for a much broader range of research, and would be more so if they could be linked up in a way that allowed them to be explored as a single data landscape. However, while the resources may be reusable, the variety of data representations and formats used militates against such an integrated view. We describe two case studies that address this issue of interoperability by creating virtual resources that are independent of the underlying data structures and storage systems, thus allowing heterogeneous resources to be treated in a common fashion while respecting the integrity of the existing data representations.

Ryan Baumann *The 'Son of Suda On-line'* pp. 91-106

The Son of Suda On-Line (SoSOL) represents the first steps towards a collaborative, editorially-controlled, online editor for the Duke Databank of Documentary Papyri (DDbDP). Funded by the Andrew W. Mellon Foundation's Integrating Digital Papyrology Phase 2

(IDP2), SoSOL provides a strongly version-controlled front-end for editing and reviewing papyrological texts marked up in EpiDoc XML.

Elaine Matthews and Sebastian Rahtz

*The Lexicon of Greek Personal Names and classical web services* pp. 107-24

This chapter documents the data resources of the long-term classical research project, *The Lexicon of Greek Personal Names* (LGPN), published in six volumes since 1987. It explains and demonstrates the web interfaces and services which now make available online the bulk of the LGPN, providing both powerful searching tools for scholars and an interface to allow other systems to link to LGPN data. Making the data available online provides direct, unmediated access to the material and supports exploitation of the data for further research both individual and collaborative.

We describe the work that went into creating the Lexicon, detail the granularity of the data structures, and explain the history of the project's record management. We then move onto the work undertaken in recent years to provide an archival XML-based format for the Lexicon's long-term preservation, and show how this has allowed us to build new web services, including exposure of Resource Description Framework (RDF) metadata, using the ontology of the CIDOC Conceptual Reference Model (CRM) ontology for semantic web applications.[1]

Simon Mahony: HumSlides *on Flickr: using an online community platform to host and enhance an image collection* 125-46

Moving a teaching and research image collection from an analogue to a digital medium for delivery brings with it many advantages but at the same time it also presents many new problems and ones probably not previously considered. This chapter discusses the move of a departmental slide collection, firstly to a proprietary in-house format, and then subsequently to the online community platform Flickr. It draws on the experience and model of the Library of Congress in partnership with Flickr and *The Commons*, as well as initiatives at Oxford and at New York University, and in doing so critically analyses and evaluates the possibilities for the future development of this collection. It asks why this collection is not currently being used to its potential and examines how the development of a user community would help to enrich the collection and ensure long term sustainability and future growth.

[1] CIDOC CRM is an ISO standard (21127:2006) that 'provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation' <http://www.cidoc-crm.org/>.

One of the great potentials of the internet is its capacity to aggregate and unify information from diverse sources. Information in the Classics, and data generated by classicists, is inherently fragmented, and organized according to different standards. This paper describes a project at King's College London which sought to provide a set of aggregating services to humanities scholars. www.arts-humanities.net provides a platform, a library, and a taxonomy to organize and present data: we describe its facilities for supporting a multi-source dataset tracing the paths of Romano-British inscriptions, both in space and conceptually. Itinerant geographies of metrical versus text inscriptions are discussed, including how these can be published in a variety of non-conventional platforms, such as Twitter. We argue that, in the future, these platforms will come to play a critical role in the wider scholarly discourse of the Classics.

## ABBREVIATIONS

| | |
|---|---|
| ABM | Agent-based modelling |
| ADS | Archaeology Data Service |
| AHRC | Arts and Humanities Research Council |
| API | Application Programming Interface |
| APIS | Advanced Papyrological Information System |
| AWIB | Ancient World Image Bank |
| CC | Creative Commons |
| CI | Collective Intelligence |
| CIDOC | International Council of Museums |
| CRM-CIDOC | CIDOC Conceptual Reference Model |
| CSV | Comma-separated data fields |
| DANS | Data Archiving and Networked Services |
| DARIAH | Digital Research Infrastructure for the Arts and Humanities |
| DDbDP | Duke Databank of Documentary Papyri |
| DPI | Dots per inch |
| DVCS | Distributed Version Control Systems |
| EDM | Europeana Data Model |
| GAP | Google Ancient Places |
| GIS | Geographical Information Systems |
| HEA | Higher Education Academy |
| HEFCE | Higher Education Funding Council for England |
| HESTIA | Herodotus Encoded Space-Text-Image Archive |
| HGV | *Heidelberg Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens* |
| IAph | *Inscriptions of Aphrodisias* |
| IDP | Integrating Digital Papyrology |
| ISAW | Institute for the Study of the Ancient World |
| JDI | Image Digitization Initiative |
| JISC | Joint Information Systems Committee |
| JSON | JavaScript Object Notation |
| KML | Keyhole Markup Language |
| LaQuAT | Linking and Querying Ancient Texts |
| LCCW | Longest Common Consecutive Words |
| LGPN | *The Lexicon of Greek Personal Names* |
| LoC | The Library of Congress |
| MDID | Madison Digital Image Database |
| OAI-ORE | Open Archives Initiative Object Reuse and Exchange |
| OER | Open Education Resources |
| OGSA-DAI | Open Grid Service Architecture–Data Access and Integration |
| OGSA-DQP | Distributed Query Processing |
| PDF | Portable Document Format |

| | |
|---|---|
| PN | Papyrological Navigator |
| PRM | Probabilistic Road Map |
| RDF | Resource Description Framework |
| RIB | *Roman Inscriptions of Britain* |
| SGML | Standard Generalized Markup Language |
| SOL | Suda On-Line |
| SoSOL | Son of Suda On-Line |
| SQL | Structured Query Language |
| SVG | Scalable Vector Graphic |
| TEI | Text Encoding Initiative |
| TLG | *Thesaurus Linguae Graecae* |
| URI | Uniform Resource Identifiers |
| V&A | Victoria and Albert Museum |
| VRE | Virtual Research Environment |
| WFS | Web Feature Service |
| WMS | Web Map Service |
| WYSIWYG | What-You-See-Is-What-You-Get |
| XML | Extensible Markup Language |

# MEASURING THE INFLUENCE OF A WORK BY TEXT RE-USE

MARCO BÜCHLER, ANNETTE GEßNER,
MONICA BERTI, THOMAS ECKART

*1. Introduction*

Before the invention of modern book printing it was a very difficult, onerous, and expensive task to copy a text: the text itself had to be obtained, writing material was expensive, and even good scribes took a long time to make good copies by hand. It seems evident that everybody would think quite carefully about which books to copy and which not, and we thus have to assume that only certain kinds of books have been preserved until today. So what could have been the criteria behind these decisions?[1]

There are different reasons why a work was copied often enough to survive until today. One of them was good fortune: for example, it was conserved on a papyrus in the sands of Egypt or was 'accidentally' passed on as a palimpsest and was not destroyed in a fire, as in the Library of Alexandria. Another reason could be that this work had not been forbidden for ideological reasons (for instance in times of iconoclasm) and many other reasons could be added here as well.[2]

But we also have to consider the possibility that the significant influence of a work was one of the main reasons for its uninterrupted tradition. A work must have held the genuine interest of enough people who considered it to be worth copying and could also afford to do so. While it cannot always be determined why one work has been transmitted and another one has not, we have to assume that a work passed down until today had a certain impact that made people copy it again and again throughout the centuries. Those works must have been important, valuable, and/or useful to those who decided to copy them, or requested a copy. So, if the tradition of a work is long enough that it still exists in some form today, this is in itself quite significant and presents a question worth researching.

Classicists have long tried to determine the influence of a work by measuring its reception by citation indices throughout the centuries. The most common method used is to examine the texts of authors whose works have survived to look for traces of quotations and text re-use in their works. The major problem with this approach is that the manual collection of every passage that bears evidence of text re-use (especially of lost works) is a very demanding and time-consuming task. In order to obtain faster and hopefully more

[1] For an interesting discussion of this question see: H. A. Cayless, *Ktêma es aei: digital permanence from ancient perspective*, in *Digital research in the study of classical antiquity*, ed. G. Bodard and S. Mahony (London 2010) 139-50.

[2] See: H. Hunger, *Handschriftliche Überlieferung in Mittelalter und früher Neuzeit, Paläographie*, in *Einleitung in die griechische Philologie,* ed. H. G. Nesselrath (Stuttgart and Leipzig 1997).

complete results, classicists can now use a variety of tools at least partially to automate this kind of research. The approach described in this chapter is to provide an automatic search for textual re-use[3] and then to visualize the results, taking different aspects of re-use into consideration in order to make them easier to interpret.

Any discussion of textual re-use must also address the larger question of so-called 'fragments' of lost authors and works. The term 'fragment' is applicable to a wide range of ancient evidence, which includes archaeological ruins, epigraphical and papyrological documents, and many other pieces of the material record. By 'fragments', however, we mean not only the material remains of ancient writings, but also quotations of lost texts preserved through other texts. A huge number of quotations of lost texts have been gathered together in print collections, enabling scholars to reconstruct lost works and depict the personalities of 'fragmentary authors'.[4]

The importance of gathering quotations (fragments) of lost works is due to the fact that a significant majority of ancient texts have been lost. Nonetheless we can reconstruct this inestimable cultural patrimony thanks to traces of text re-use preserved in later works. At the same time, collecting fragments of lost authors also permits us to provide a useful measure of the shifting boundaries of canon formation over time.[5] Moreover, working with quotations of lost works serves as an extraordinary methodological exercise in attempting to discover patterns that could be useful within the fields of allusion discovery, plagiarism detection, and text re-use. Finally, gathering fragments of ancient works and representing them digitally is a fundamental exercise: model is built that can also be very useful for tracking modern quotations and, in particular, for use in multi-million book libraries such as Google Books or the Internet Archive.[6]

## *2. Related work and state of the art*

In the field of text re-use, much research has already been conducted and, while it is impossible to address all relevant work here, some important aspects are summarized 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 such as a sentence or a paragraph. This task can typically be done with a linear cost of $O(n)$. The set of

[3] By 'text re-use' we mean a textual congruence, which can be a good indicator for finding quotations. Nevertheless until proven it remains uncertain if this re-use is indeed a (direct or even indirect) quotation or a text passage, which has been quoted by later authors.

[4] For more on this see: M. Berti, M. Romanello, A. Babeu and G. Crane, 'Collecting fragmentary authors in a digital library (Greek fragmentary historians)', in *Proceedings of the 9th ACM/IEEE-CS joint conference on digital libraries* (Austin, TX 2009) 259-62, and M. Berti, 'Fragmentary texts and digital libraries', in *Philology in the age of Corpus and Computational Linguistics,* ed. G. Crane, A. Lüdeling, and M. Berti, CHS Publication (forthcoming).

[5] See: G. Most, ed., *Collecting fragments - Fragmente sammeln,* (Göttingen 1997).

[6] O. Kolak and B. N. Schilit, 'Generating links by mining quotations', in *Proceedings of the nineteenth ACM conference on hypertext and hypermedia (HT 2008). Pittsburgh, Pennsylvania*, (New York, NY 2008) 117-26.

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 re-use by pairwise comparison is quite time-expensive due to the squared complexity of $O(n^2)$. This approach is useful for comparing smaller corpora such as the Dead Sea Scrolls with the Hebrew Bible.[7] But using it with an ancient Greek corpus like the CD-ROM Version E of the *Thesaurus Linguae Graecae* called *TLG*-E,[8] which has about 5.5 million sentences, would require *3.025e13* comparisons. 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 a single 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 summarized as a four-step process:

- *Fingerprinting*: Every re-use unit such as a paragraph or a sentence first needs to be fingerprinted. In detail, this means that the re-use unit can be quantified by a set of syntactical features such as any *n-gram approach* and semantic methods like *semantic clustering*. Within this chapter we decided for reasons of reliability to utilize a syntactical approach. Following this, two questions must be answered: first, does an overlapping or a non-overlapping approach make the most sense for this kind of problem, and, second, does a static or a non-static n-gram size fit best for the investigated question? To explain in greater detail, overlapping features means that a sentence or a paragraph is quantified by pairwise overlapping n-grams (shingling). Furthermore, this implies that every word is part of at least two n-gram features. In contrast to this, non-overlapping fingerprinting means that every word is exclusively part of one feature. For this research, we decided to choose an overlapping and non-static fingerprinting approach that is named *Longest Common Consecutive Words* (*LCCW*). Depending on both the research question and the problem, syntactical features are sometimes preferred, while in other cases fingerprints that are more semantic are required (see section *Measuring influence by hypertextuality*).
- *Selection*: Quoting a text passage always implies purposeful re-use and this means that both the original text and the subsequent quotation have similar patterns as far as the fingerprints are concerned. For this reason, not all possible fingerprint values are necessary. Imagine that two identical re-use units of twenty words exist that are then compared by a bi-gram fingerprinting. Without any selection all of the nineteen possible bi-gram features would link these two sentences with each other. It would be necessary, however, to have just one link between these sentences. In order to avoid any missing links by too strong a

[7] R. Hose, *CS490 Final report: investigation of sentence level text reuse algorithms,* Boom 2004 Bits On Our Minds: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.90.9835> [accessed on 2nd July 2011].

[8] *TLG* Consortium, *Thesaurus linguae Graecae,* CD-ROM Disk E, University of California,Irvine, released in February 2000.

selection process, four or five significant features would represent the re-use units perfectly.

- *Linking*: This step links 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 directed link. 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 next step is to score the similarity of the two linked passages.

In both of these last two steps, links of some passages are rejected. Depending on the text and the degree of textual re-use, there is often a strong selection in the linking step. Several experiments on different corpora and languages have shown in the past that only one in 100 million possible linking candidates is considered as an actual case of textual re-use.[9] The scoring itself can be seen more as a fine-tuning that removes less similar sentences.

The *fingerprinting* 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 re-use, 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 fingerprinting process, the link features are defined. Generally, there are three different clusters of approaches:

- *Words as features*: after all the function words, such as articles and conjunctions, 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]

[9] See: M. Büchler, *Informationstechnische Aspekte des historischen Wissenstransfers*. (Engl. *Computational aspects of historical knowledge transfer*). (PhD thesis submitted Leipzig University 2013).

[10] Hose, *Final report* (n. 7 above)

[11] John Lee, 'A computational model of text reuse in ancient literary texts', in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, (Prague 2007) 472-79.

[12] B. Mittler, J. May, P. Gietz and A. Frank, *QuotationFinder - Cluster Asia and Europe - Uni Heidelberg* (2009): <http://www.asia-europe.uni-heidelberg.de/de/forschung/heidelberg-research-architecture/hra-projects/quotationfinder> [accessed on 11th January 2010].

[13] Mittler *et al*., *QuotationFinder* (n. 12 above), Lee, *Computational model* (n. 11 above).

- *N-grams as features*: to extract textual re-use 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]
- *Sub-graphs as features*: graph-based approaches, as shown in this chapter, deal with semantic relations between words. In the *Lexical Chaining* approach[15] that is often used for text summarization,[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 re-use 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. Measures like the *Dice coefficient* compute the *similarity* of two pairwise linked passages by commonly used (and overlapping) words, while other measures like the *city block metric*, *Euclidean distance,* or the *Jenson-Shannon divergence* calculate the *semantic* distance between two units.[17] The main difference between these two types of measures is that a *similarity* measure scores relevant links with a high score, whereas *distance* measures score a relevant link of two units as close as possible to zero.

Given a corpus *C,* a re-use graph $G=(V,E)$ can be described by the following generalized algorithm:

1. V = segment_corpus(C) with $v_1, v_2, ..., v_n \in V$, $\cup v_i = C$ and $v_i \neq v_j$
2. **for each** $v_i \in V$
3. $F_i$=train_features($v_i$);
4. **for each** $v_i \in V$
5. **for each** $f_k \in F_i$

[14] Hose, *Final report* (n. 7 above); M. Büchler, G. Heyer, and S. Gründer, *Bringing modern text mining approaches to two thousand year old ancient texts, e-Humanities – an emerging discipline*. Workshop in the 4th IEEE International Conference on e-Science (2008); M. Büchler, *Medusa release homepage – a statistical engine for natural language processing matters*. <http://mbuechler.e-humanities.net/medusa/, 2005-2011> (2011).

[15] U. Waltinger, A. Mehler, G. Heyer, 'Towards automatic content tagging: enhanced web services in digital libraries using lexical chaining', in *4th Int. Conf. on Web Information Systems and Technologies (WEBIST '08), 4-7 May, Funchal, Portugal*, ed. J. Cordeiro, J. Filipe and S. Hammoudi (Barcelona 2008) 231-36.

[16] L. Yu, J. Ma, F. Ren, S. Kuroiwa, 'Automatic text summarization based on lexical chains and structural features', in *snpd, vol. 2, Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, (Qingdao 2007) 574-78.

[17] For all these see: S. Bordag, *Elements of knowledge-free and unsupervised lexical acquisition,* (Unpubl. PhD thesis, Leipzig University 2007).

6. $e_i$=($v_i$,$v_j$)∈$E$=select all $v_j$ containing feature $f_k$

7. **for each** $e_i$∈$E$

8. $s_i$=scoring($e_i$=($v_i$,$v_j$)∈$E$; $F_i$; $F_j$);

9. if($s_i$<*threshold*){$E$=$E$\{$e_i$}}

*Listing 1: Generalization of a textual re-use 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 $F_i$ 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))*

Within the humanities, text re-use has always been an important field of research, especially in classical philology, and there has often been a strong focus on fragmentary works. The importance of providing technical tools for conducting research in this area has been a subject of increasing importance recently as millions of texts have been digitized. Typically, classicists had to conduct manual searches of textual concordances in databases such as the *TLG*, and the main drawback is that this work takes a lot of time and can be incomplete.

Regarding the fragments of lost works, we presently have many printed collections of Classical fragmentary authors. In particular, we have many collections of the fragments of Greek historians.[18] These collections have allowed scholars to reconstruct the characteristics and personalities of otherwise-lost authors. One of the major limits of printed collections of fragmentary authors is that one can only see the quotation *extracted* from the context in which it has been preserved. New digital models for collecting and representing fragments permit us to see the text of the fragments inside their *contexts* of transmission and according to different editions. In this multilevel structure, we can reconstruct the whole tradition of a text from its 'original' form to its reception and transmission across the centuries.[19]

Digital libraries and hypertextual models allow us to rethink the fundamental question of the relation between the fragment and its context of transmission, including representing and expressing every element of print conventions in a more dynamic and interconnected way. A fragment is in itself a perfect model of hypertext and in a digital library the fragment can be linked to the whole text of the source in which it is preserved. In this way it is possible to see the excerpt directly inside its context of transmission, avoiding the misleading idea of an independent material existence of fragmentary texts.[20]

[18] See: Berti *et al*., 'Collecting fragmentary authors' (n. 4 above) and M. Berti, 'Fragmentary texts' (n. 4 above).

[19] See: Berti *et al*., 'Collecting fragmentary authors' (n .4 above) and M. Berti, 'Fragmentary texts' (n. 4 above).

[20] See: Berti *et al*., 'Collecting fragmentary authors' (n. 4 above) and M. Berti, 'Fragmentary texts' (n. 4 above).

| Author | Work | Century | Genre |
|---|---|---|---|
| HELLANICUS | Fragmenta | 5 BC | Hist. |
| XENOPHON | Memorabilia | 5 BC | Hist. |
| ZENO | Testimonia et fragmenta | 4 BC | Hist. |
| PLATO | Timaeus | 5 BC | Phil. |
| EPHORUS | Fragmenta | 4 BC | Phil. |
| ARISTOTELES et CORPUS ARISTOTELICUM | De anima | 4 BC | Phil. |
| Flavius JOSEPHUS | Contra Apionem (= De Judaeorum vetustate) | 1 AD | Hist. |
| PLUTARCHUS | Pompeius | 1 AD | Hist. |
| APPIANUS | Mithridatica | 1 AD | Hist. |
| GALENUS | Ad Glauconem de medendi methodo libri ii | 2 AD | Phil. |
| CELSUS | Ἀληθὴς λόγος | 2 AD | Phil. |
| ALEXANDER | De anima | 2 AD | Phil. |

Table 1: An overview of the selected authors and works in this paper (names as in the *TLG-E*).

### *3. Investigated works*

In order to test our approach, we have chosen twelve authors from two different time periods and two different genres (see Table 1). The most important criterion for choosing these works was comparability. One of the requirements was that the works had to have a similar length (*i.e.* number of tokens). Then they needed to belong to similar genres and time-spans and also had to include fragmentary authors. We have chosen philosophy and historiography as genres not only because of our genuine interest in these two literary fields, but also because we had observed some differences between the quotation of philosophical and historical texts.[21] For time-spans we have chosen the fifth and fourth centuries BC and the first and second centuries AD since these centuries seemed to contain enough interesting authors with works in the two selected genres as well as with almost the same text-length (*ca*. 15,000 to 30,000 tokens). The works chosen are not supposed to be considered the most important or influential ones of their time and genre, but to offer some variety in order to compare the results.

### *4. Data used and pre-processing*

#### *4.1 Data used*

All illustrated methods and results are based on the *Thesaurus linguae Graecae* Version E,[22] a comprehensive collection of Greek writers, including many well-known authors like Plato and Sophocles and coverage from Homer's time to the fall of Constantinople in the

[21] See Büchler, *Informationstechnische Aspekte* (n. 9 above).

[22] *TLG* Consortium, *Thesaurus Linguae Graecae,* CD-ROM Disk E, University of California,Irvine, released in February 2000.

fifteenth century. This corpus has been created and provided by the *TLG* research centre at the University of California, Irvine, and today it is one of the most important digital resources when dealing with ancient Greek texts. The version used for this study contains around 7200 works written by more than 1800 different authors over a time period of more than 1800 years. Since the origin of this digital corpus goes back to the 1970s, all textual data and metadata are encoded in a binary format (*TLG*-E) that is not a suitable basis for efficient Text Mining applications. A rather comprehensive tool chain of pre-processing steps therefore had to be built.

*4.2 Pre-processing*

Several specific tools were either developed or adapted, including an extractor for the binary input data, a Beta Code to Unicode converter as well as different tools for dealing with problems concerning a strongly inflected language such as ancient Greek and its various changes over a long period of time.

Step 1: Sentence segmentation
As a first pre-processing step, a newly developed, rule-based sentence boundary detection algorithm splits the text. To deal with various extraneous information that is unimportant to the detection of textual re-use (such as the marking of speaker roles), different lists of boundary marks are used in combination with abbreviation lists to enhance the sentence boundary detection rate.

Step 2: Tokenization
As the next step, all tokens were extracted from the segmented sentences. In comparison with modern languages as modern English or German, a more active *tokenization* process was needed for dealing with ancient Greek. Specifically, this means that more irrelevant parts of the input material had to be removed to gain usable text. In addition to punctuation marks, all brackets of the Leiden Conventions were also removed.

As a result of these first two steps, all *TLG* works are segmented into about 4.96 million sentences with an average length of 13.51 words. Table 2 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 2: Cumulative distribution of sentence length in words

Step 3: Normalization
Since the ancient Greek language contains a large number of diacritical marks and many words also exist in a variety of upper/lower-case letter combinations, many different forms of the same word type can be found in the corpus. As an example, the

conjunction καί exists in the *TLG*-E in more than fifteen different versions.[23] Since many of these variants exist due to changes in writing or modern modifications of the original text (such as the usage of lower case letters), a re-use detection process based on these variants might ignore a huge portion of relevant text passages. Therefore, a normalization process is executed that reduces all words internally to a lower-case representation and removes any diacritics. Table 3 shows the number of different spelling variants for some high frequency words of the *TLG*.

| Word | *τοῦ* | *πρός* | *τοῖς* | *κατὰ* | *τοῦτο* | *εἶναι* | *βασιλεία* |
|---|---|---|---|---|---|---|---|
| Number of variants | 15 | 8 | 8 | 21 | 10 | 15 | 14 |

Table 3: Number of word variants with identical normalized word form

Step 4: Lemmatization
Another class of variations of the same word are due to morphology. Therefore, all words have been analysed and internally reduced to their base form by using the morphological analyser Morpheus, which was developed by the *Perseus Digital Library* (*Perseus*).[24] As Morpheus can also identify dialects, even dialectical variants are reduced to the same base form.

### *4.3 Pre-processing as an ongoing process*

The pre-processing of highly structured text, especially on a corpus that covers a very long period of time, is not a task that can simply be done once and then be considered as complete. Because of various differences between ancient Greek and modern languages, many standard Natural Language Processing (NLP) tools failed when used on ancient Greek text: tokenization and sentence segmentation proved particularly difficult. For this reason, existing tools had to be substituted by specialized replacements, various parameters had to be adjusted, and dedicated tools had to be created to deal with special phenomena. Thus the quality of the whole pre-processing system had to be evaluated regularly by classicists and the task turned out to be an ongoing process.

## *5. Re-use methodology*

Text re-use is represented in a formal re-use graph *G=(V,E)* having *V* as the set of re-use units (such as the sentences of a text corpus) and *E* as the set of pairwise edges between elements of *V*. Within this study, we decided to use the *Longest Common Consecutive Words* fingerprinting (*LCCW*) with a selection of at least 5 words in a row and with feature frequency of at least 2. We have chosen to use the dice coefficient as the similarity measure with a threshold of 0.4 that provides good results. This threshold is low enough not to ignore embedded quotations and is high enough typically to ignore phrases such as ‘in the Name of our Lord Jesus Christ’.

[23] *καὶ, Καὶ, καί, Καί, και, Καῖ, Και, καῖ, καΐ, καὶ, καί, Καϊ, Καΐ, Κᾶι, Κάι.*

[24] See the Digital Classicist wiki entry for Morpheus: <http://wiki.digitalclassicist.org/Morpheus>.

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 between a text re-use and the original text can be found. With this information available, visual access for philologists can be provided quite simply since the boundaries of an n-gram are determined by one of the following three causes:

- the beginning of a sentence,
- the end of a sentence, or,
- any kind of a differing word due to causes such as language evolution, dialect change, an inserted word or the boundaries of an embedded re-use within a larger sentence.

A negative consequence of this approach is that all common prefixes of the longest match that consist of at least five words are produced.[25] Consequently a post-processing step removes these prefixes. In addition, finding the prefix properties of those n-grams requires a frequency heuristic such as:

$$eps = \log_2 \left( \frac{Frequency(x_1 x_2 \ldots x_n)}{Frequency(x_1 x_2 \ldots x_n x_{n+1})} \right)$$

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 less common longest n-gram match. 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.

The reasons for deciding to use *LCCW* as the text re-use fingerprinting method for this study are twofold. First, by way of using this most restrictive fingerprinting, we can assume a higher level of precision. Since it is impossible to validate all detected hypertextual edges in *E* manually, it made sense to us to choose this very restrictive technique. By way of the similarity threshold of 0.4, however, the re-use detection remains aware of embedded quotations. Secondly, using *LCCW* keeps the amount of data at an acceptable level.

### *6. Measuring influence by hypertextuality*

Within this study we introduce two parameters of measuring hypertextuality: *re-use coverage* $C_{LCCW}$ and *re-use temperature* $T_{LCCW}$, whereas *LCCW* indicates the fingerprinting methodology.

[25] The minimum threshold of five is chosen by statistical properties of n-grams. With respect to the audience, however, we skipped a detailed description.

Given a set *V* of re-use units such as a sentence, *re-use coverage* $C_{LCCW}$ measures the ratio between quoted re-use units of a work in relation to the total amount of re-use units of this work. It is, however, not of interest if a re-use unit is quoted more than once or not.

The *re-use temperature* $T_{LCCW}$, however, measures the frequency *f* for a dedicated re-use unit and scales it as described by the following formula:

$$\mathcal{T}_{LCCW} = log_{10}(f) + 1$$

Table 5 illustrates the behaviour of this formula in a bit more detail. The logarithmic scaling of *f* down-scores peaks significantly. If $T_{LCCW}$ increases by 1, ten times more quotations can be observed (see Table 5).

| **Re-use temperature** $T_{\mathrm{LCCW}}$ | 0,00 | 1,00 | 1,50 | 2,00 | 2,50 | 3,00 | 3,50 |
|---|---|---|---|---|---|---|---|
| **frequency of re-use f** | 0,00 | 1,00 | 3,16 | 10,00 | 31,62 | 100,00 | 316,23 |
| **color** | black | blue | violett | red | dark orange | orange | yellow |

Table 5: $T_{LCCW}$ vs. f. Translation table of $T_{LCCW}$(f) for some significant re-use temperatures. Furthermore, $T_{LCCW}$(f) can be mapped to colours that are used in Figure 1.

Whereas $C_{LCCW}$ measures the 'breadth' of re-use, *i.e.* how many parts of the work have been re-used at least once, $T_{LCCW}$ indicates the 'depth' of re-use, called 'temperature' by way of measuring the *re-use frequency* of a dedicated re-use unit. By way of ordering all re-use units as they occur in the text, a temperature map as shown in figure 1 can be generated. At the x-axis these ordered re-use units are shown. They are normalized to relative positions. A text position of 0.4 represents a text re-use unit at 40% of this work. The y-axis is solely to add one dimension, so that the plots in figure 1 are a rectangle instead of a single line. The colours represent the *re-use temperature* $T_{LCCW}$ as indicated in Table 5.

Before interpreting these results, it has to be taken into consideration that the chosen text re-use discovery algorithms find passages that use *exactly* the same words as the chosen text, *i.e.* they are very literal textual concordances. The majority of located passages would likely be considered to be intentional references. Due to no limitations of the time span in which the re-use is found, this method not only shows by whom the chosen authors (see table 1) have been re-used, but also what older text-passages these authors re-used themselves. Nevertheless, by demonstrating intertextuality, the results are of great help for classicists trying to answer our research question.

The results shown in figure 1 have to be interpreted by a classicist. Works like the *De Anima* of Aristotle and Celsus' *Ἀληθὴς λόγος* seem to have been referenced very broadly, due to the fact that at least one re-use has been found for almost every part of these works. But this high amount of referencing could be for a variety of reasons: for instance, it could

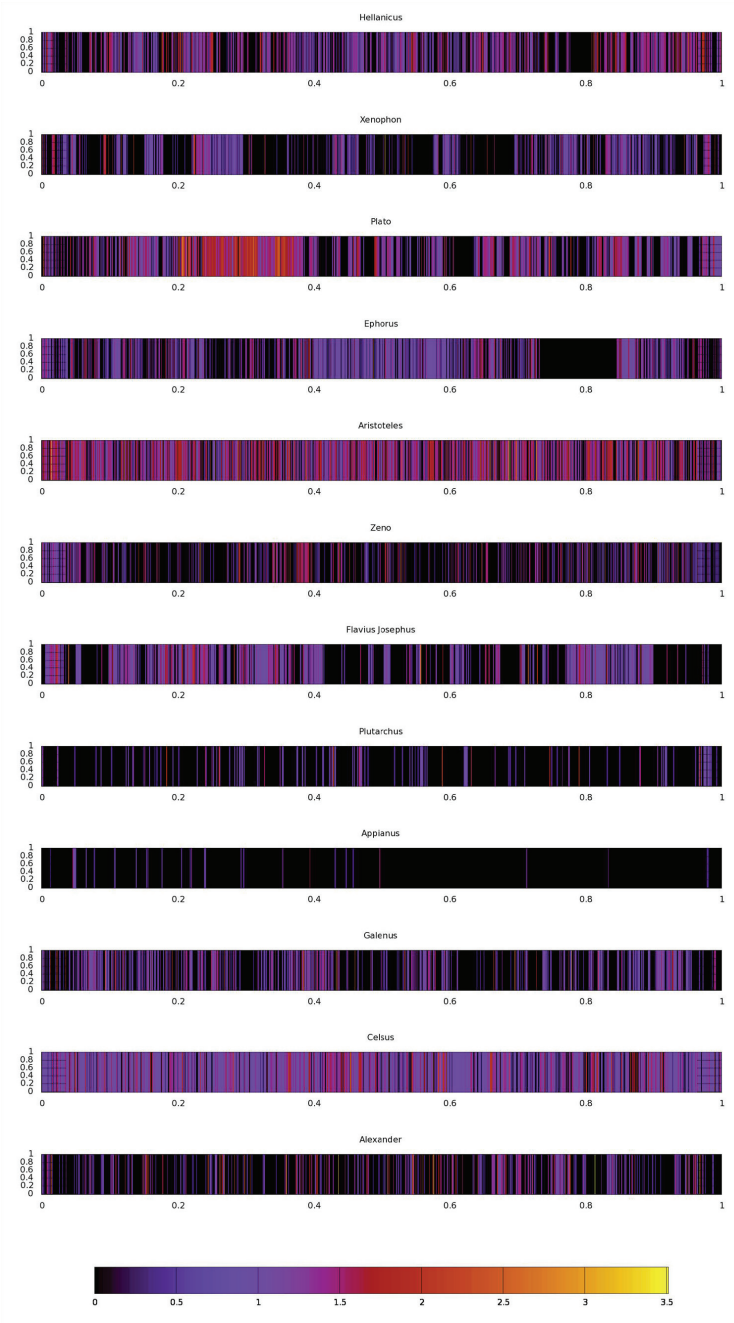


Figure 1: Re-use temperature $T_{LCCW}$ for all works from table 1. The x-axis represents the re-use units (sentences). A score of 0.2 represents a text position of 20% within this work. The colours represent the re-use temperature $T_{LCCW}$.

be the result of anthologists, who decided to add some parts of these texts into their collection. While the fact that a work is chosen to be part of an anthology itself seems to indicate the relative importance of a work, is it really enough to measure its larger influence? We think that, in order to measure the impact of a work, we should consider factors beyond how *many* parts of this text have been re-used and should also explore if there are some passages of this text that became so 'popular' that they were used much more *often* than others (*i.e.*, they had a (very) high temperature). As we know from famous sayings such as 'I know that I know nothing' and 'To be or not to be', the popularity of an author or work can frequently be demonstrated by only a few words that are cited very often. Table 6 demonstrates this concept, showing the top three works with the highest numbers of passages found for every temperature.

As this table demonstrates, it is quite possible to state that Aristotle's work can indeed be considered very influential, because it has not only been re-used *broadly* but also because it ranks in the top three in every category of temperature. Celsus has some high temperatures too, but passages of his work are not as frequently quoted as Aristotle. It is comparable with the *Timaeus* of Plato, which is not so much *broadly* referenced, but has lots of high temperatures for distinctive parts of the text.

By comparing Plato with Hellanicus in figure 1, the difference between a work of Plato that still exists and a fragmentary work can be shown. Whereas for Hellanicus only smaller clusters of re-use can be observed (which depend significantly on the order of the fragments), one strongly quoted cluster exists in Plato's *Timaeus* between the text positions 0.2 and 0.4.

Perhaps the most interesting result observed is that the work of Alexander of Aphrodisias, which is not broadly cited at all, has one of the highest temperatures of all chosen works. This could mean that single-text passages of this work have had a significant influence on other authors.

Furthermore, philosophical texts tend to be quoted more broadly, whereas historical texts are quoted in a more focused and frequent fashion. Table 6 supports this point. While a high text re-use temperature of 2.5 finds historians as part of the top three, they disappear significantly from the top three as the text re-use temperature is gradually decreased. On the other hand, a philosopher such as Alexander of Aphrodisias is, significantly, within the top three. When decreasing the text re-use temperature, however, his work decreases in terms of its score. In detail, this means that Alexander seems to have been quoted in some text passages very frequently wherever his work has an atypical quotation usage for less frequent re-used text passages. A possible reason for these few highly quoted passages could be due to the fact that this work of Alexander of Aphrodisias is a commentary on the work *De Anima* written by the famous philosopher Aristotle, which includes quoting passages from this work, which can be very often re-used phrases themselves.

One significant result of the temperature maps of Figure 1 and the data from table 6 is that philosophical texts tend to be re-used more literally than historical texts. This is likely due to the fact that in philosophical texts certain ideas and thoughts of famous philosophers are discussed and thereby repeated much more often and more accurately than in historiography. In table 7 this fact is highlighted in detail. Table 7 aggregates all the authors from table 1 within the genres of philosophy or history on the one hand. On the other hand, all authors are grouped into two clusters based on their living time. Comparing both columns

representing the genre, a significant difference of 0.22 (0.5475-0.3268) between philosophy and history is observed. This indicates a more literal quotation style of philosophical rather than historical texts. On the other hand, by comparing both rows of dating clusters, a similar quotation style is observable. Whereas, in the fifth and fourth centuries BC coverage of 0.52 is possible it is only 0.35 for the first and second centuries AD.

| | | Re-use temperature | | | | | |
|---|---|---|---|---|---|---|---|
| | | *3,5* | *3* | *2,5* | *2* | *1,5* | *1* |
| **Authors** | *HELLANICUS* | 0,00000 | 0,00000 | 0,00000 | 0,01004 | 0,08233 | 0,53614 |
| | *XENOPHON* | 0,00000 | 0,00045 | 0,00223 | 0,00848 | 0,03527 | 0,35536 |
| | *PLATO* | 0,00000 | 0,00000 | 0,00426 | 0,04691 | 0,16311 | 0,61301 |
| | *EPHORUS* | 0,00000 | 0,00000 | 0,00000 | 0,00179 | 0,03461 | 0,50776 |
| | *ARISTOTELES et CORPUS ARISTOTELICUM* | 0,00160 | 0,00480 | 0,01279 | 0,03437 | 0,24860 | 0,74820 |
| | *ZENO* | 0,00000 | 0,00000 | 0,00561 | 0,01402 | 0,05329 | 0,30208 |
| | *Flavius JOSEPHUS* | 0,00000 | 0,00000 | 0,00102 | 0,00917 | 0,04383 | 0,47299 |
| | *PLUTARCHUS* | 0,00000 | 0,00000 | 0,00000 | 0,00444 | 0,01110 | 0,13762 |
| | *APPIANUS* | 0,00000 | 0,00000 | 0,00000 | 0,00000 | 0,00102 | 0,03176 |
| | *GALENUS* | 0,00000 | 0,00075 | 0,00373 | 0,01343 | 0,03731 | 0,37537 |
| | *CELSUS* | 0,00000 | 0,00000 | 0,00166 | 0,02244 | 0,08728 | 0,86367 |
| | *ALEXANDER* | 0,00074 | 0,00297 | 0,00742 | 0,01782 | 0,05197 | 0,25390 |

Table 6: Selected re-use temperature $T_{LCCW}$ for all works of table 1. Those cells that are marked with grey background colour are part of the top three authors at this temperature.

| | | Genre | | |
|---|---|---|---|---|
| | | *Philosophy* | *History* | |
| **Century** | *5th and 4th* | 0.61112549 | 0.42769148 | 0.5209896 |
| | *1st and 2nd* | 0.48431877 | 0.25644378 | 0.35233665 |
| | | 0.54751773 | 0.32680458 | |

Table 7: A contingency table of re-use coverage of dating by century and genre.

*7. Further work*

This chapter has examined two text re-use properties out of the much larger set that could be investigated. Both *re-use coverage* and *re-use temperature* are properties that can be extracted easily. An ever more detailed consideration of this topic, however, would necessitate including additional data, such as the dating of authors and texts. In this chapter we did not make a distinction if one of our selected authors quoted another one or if the author himself was quoted. The key issue is that the necessary date information is of too poor a quality and this makes this type of work almost impossible. By improving this information, however, we will separate both *re-use coverage* and *re-use temperature* by the additional dimension of the degree of an own contribution or a quotation.

Furthermore, we could show that the same algorithm – such as the *Longest Common Consecutive Words* – works differently on two genres of the same language. On one hand, we can apply different algorithms. On the other hand, it seems to be obvious that in different genres the re-use style tends to be significantly different. For this reason one further

dimension such as the degree of closeness to the original makes sense. More generally, this dimension corresponds to the *Kolmogorov complexity* – that is an algorithmic distance between an input and an output. In the context of reception importance, the degree of change is of interest, since, if an author is quoted more literally, this quotation can be weighted higher than less literal ones, since the degree of re-using on purpose is much more significant.

Since classical scholars especially require more than just the *LCCW* algorithm, it is part of our current research plan to include further algorithms. In detail, there are currently active developments on the *TRACER* Java library that provide much more functionality.[26] The software library is not designed as a monolithic algorithm, but as a *6 level architecture* where, for each level, at least one implementation has to be selected to build the complete algorithm. Currently, there already exist about 120,000 possible combinations that have recently been evaluated in detail. This includes not only text re-use algorithms but also the extraction of *canonical references*[27] in order to apply the metrics *re-use temperature* and *re-use coverage* to those kinds of quotation traces as well.

## *8. Conclusion*

Classicists have always considered trying to determine the influence of an ancient work by measuring its hypertextuality as an effective approach. Undertaking this work by means of text-mining methods thus presents the next logical step to take. And indeed it has led to very interesting results and the visualization proved to be very useful.

Measuring influence has made it clear that we have to consider different aspects of a text's impact on other authors. It seems best to divide the found results into different categories:

- Works referenced broadly with high temperatures (Aristotle, Plato – both philosophers);
- Works referenced broadly, but not with high temperatures (Celsus);
- Works with single passages of high temperatures, but not referenced broadly (Alexander, Xenophon, Hellanicus, Zeno);
- Works which meet none of these criteria (Appianus, Plutarch – both writing about history);
- Works somewhere in between.

This research question, however, cannot be answered simply by looking at the figures and tables. Research and interpretation by scholars will always play a large role in deciding

[26] The *TRACER* library (http://etraces.e-humanities.net/TRACER) is a text re-use engine that will be available by a Creative Commons licence in summer 2013. Due to the complexity of TRACER, it is planned to initiate in summer 2013 a series of teaching courses that will be announced on the aforementioned link.

[27] For more on this see: M. Romanello, M. Berti, A. Babeu, G. Crane, 'When printed hypertexts go digital: information extraction from the parsing of indices', in *HT 09. Proceedings of the 20th ACM Conference on Hypertext and Hypermedia,* (Turin and New York 2009) 357-58.

how influential an author or work could have been, as will the data selected and used for analysis. For example, we cannot say that Plutarch was an author without any influence on other authors, but we can affirm that he was not referenced quite literally and/or those referencing works got lost. Therefore, it is necessary to search to see if he had an impact on other authors in different ways. One of the reasons for the few re-uses of Plutarch could be the fact that historiographical works do not seem to be quoted as close to the original texts as, for instance, philosophical texts, which is an interesting fact in itself.

Natural Language Processing Group, Institute of Mathematics and Computer Science, University of Leipzig, Germany [mbuechler|teckart]@e-humanities.net

Ancient Greek Philology Group, Institute of Classical Philology and Comparative Studies, University of Leipzig, Germany agessner@e-humanities.net
Dipartimento di Studi Umanistici, Università di Roma Tor Vergata Department of Classics, Tufts University monica.berti@uniroma2.it monica.berti@tufts.edu

*References*

Berti, M., M. Romanello, A. Babeu, and G. Crane, 'Collecting fragmentary authors in a digital library (Greek fragmentary historians)', in *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries (JCDL 2009)* (Austin, TX 2009) 259-62.

Berti, M., 'Fragmentary texts and digital libraries', in *Philology in the age of Corpus and computational linguistics*, ed. G. Crane, A. Lüdeling, and M. Berti, CHS Publication (forthcoming).

Beta Code *Thesaurus Linguae Graecae – the beta code manual*, online publication [accessed: 25th October 2010]: <http://www.tlg.uci.edu/encoding>.

Bordag, S., *Elements of knowledge-free and unsupervised lexical acquisition,* (Unpubl. PhD thesis, Leipzig University 2007).

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

Büchler, M., G. Heyer, and S. Gründer, *Bringing modern text mining approaches to two thousand year old ancient texts, e-Humanities – an emerging discipline*. Workshop in the 4th IEEE International Conference on e-Science (2008).

Büchler, M., *Medusa release homepage – a statistical engine for natural language processing matters*: http://mbuechler.e-humanities.net/medusa/, 2005-11 (2011).

Büchler, M., *Informationstechnische Aspekte des historischen Wissenstransfers*. (Engl. *Computational aspects of historical knowledge transfer*). (Unpubl. PhD thesis, to be submitted at Leipzig University 2013).

Cayless, H. A., *Ktêma es aei: digital permanence from ancient perspective*, in *Digital research in the study of classical antiquity*, ed. G. Bodard and S. Mahony (London 2010) 139-50.

Hose, R., *CS490 Final Report: investigation of sentence level text reuse algorithms,* Boom 2004 Bits On Our Minds [accessed: 2nd July 2011]: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.90.9835>.

Hunger, H., *Handschriftliche Überlieferung in Mittelalter und früher Neuzeit, Paläographie*, in *Einleitung in die griechische Philologie,* ed. H.G. Nesselrath (Stuttgart and Leipzig 1997).

Kolak, O., and B. N. Schilit, 'Generating links by mining quotations'*,* in *Proceedings of the nineteenth ACM conference on hypertext and hypermedia (HT 2008). Pittsburgh, Pennsylvania*, (New York, NY 2008) 117-26.

Lee, J., *A computational model of text reuse in ancient literary texts*, in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, June 2007*, Association for Computational Linguistics 2007) 472–79.

Mittler, B., J. May, P. Gietz, and A. Frank, *QuotationFinder - Cluster Asia and Europe - Uni Heidelberg* [accessed: 11th January 2010]: <http://www.asia-europe.uni-heidelberg.de/de/forschung/heidelberg-research-architecture/hra-projects/quotationfinder>.

Most, G., ed., *Collecting fragments - Fragmente sammeln* (Göttingen 1997).

*Perseus Digital Library*, online publication [accessed: 23rd October 2010]: <http://www.perseus.tufts.edu/hopper>.

Romanello, M., M. Berti, A. Babeu, G. Crane, 'When printed hypertexts go digital: information extraction from the parsing of indices'*,* in *HT 09. Proceedings of the 20th ACM Conference on Hypertext and Hypermedia,* (Turin and New York 2009) 357-58.

*TLG* Consortium, *Thesaurus Linguae Graecae,* CD-ROM Disk E, University of California,Irvine, released in February 2000.

Waltinger, U., A. Mehler, G. Heyer, 'Towards automatic content tagging: enhanced web services in digital libraries using lexical chaining', in *4th Int. Conf. on Web Information Systems and Technologies (WEBIST '08), 4-7 May, Funchal, Portugal*, ed. J. Cordeiro, J. Filipe and S. Hammoudi (Barcelona 2008) 231-36.

Yu, L., J. Ma, F. Ren, S. Kuroiwa, 'Automatic text summarization based on lexical chains and structural features', in *snpd, vol. 2, Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, (Qingdao 2007) 574-78.