JDMDH Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages

Marco Büchler, Laurence Mellerin

▶ **To cite this version:**

Marco Büchler, Laurence Mellerin. JDMDH Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages. Journal of Data Mining and Digital Humanities, 2017. halshs-03621102

HAL Id: halshs-03621102
https://shs.hal.science/halshs-03621102

Submitted on 3 Apr 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

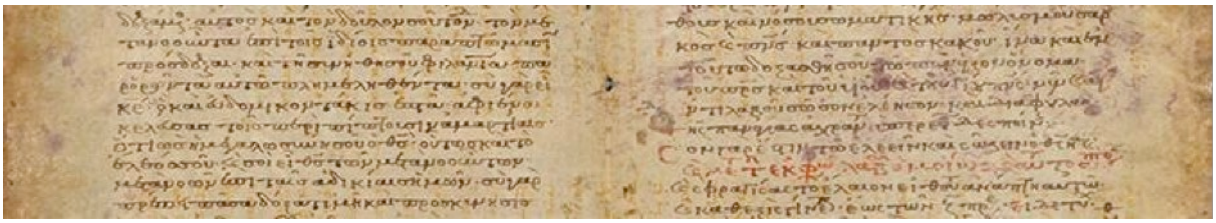
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Intertextuality in Ancient Languages

JDMDH Special Issue on

Computer-Aided Processing of Intertextuality in Ancient Languages

Edited by Marco BÜCHLER (University of Göttingen & State- und University Library Göttingen, Germany) and Laurence MELLERIN (Sources Chrétiennes, HiSoMA, Lyon, France), 2017.



This special issue originates in the International workshop on computer aidedprocessing of intertextuality in ancient languages, held in Lyon (2nd-4th June 2014), coorganized by HiSoMA (UMR 5189, Lyon), LIRIS (UMR 5205, Villeurbanne) and the University of Göttingen and SUB Göttingen (e-TRAP), with the support of the National Research Agency (ANR Biblindex) and the Partner University Fund (*PUF*).

This workshop was initiated as the conclusive meeting of the ANR project BIBLINDEX, which aims at establishing an exhaustive statement of the biblical references found in the texts of the Late Antiquity and the Middle Ages. Were gathered computer scientists and digital humanists. The sessions presented the state of art regarding concepts and technics used to process quotations and text reuses in ancient languages.

Thanks to the editorial system of the JDMDH, the proceedings of this workshop have been open to other contributions also dealing with intertextuality, linguistic preprocessing and the preservation of scholarly research results, specifically applied to corpora in Ancient Languages and for which few online resources exist (Ancient Greek, Latin, Hebrew, Syriac, Coptic, Arabic, Ethiopic, etc.).

Part 1: Towards a Digital Ecosystem: NLP. Corpus infrastructure. Methods for Retrieving Texts and Computing Text Similarities

- Methods for the detection of intertexts and text reuse, manual (e.g. crowd-sourcing) or automatic (e.g. algorithms);
- Infrastructure for the preservation of digital texts and quotations between different text passages;
- Linguistic preprocessing and data normalisation, such as lemmatisation of historical languages, root stemming, normalisation of variants, etc.

1) Identification of Parallel Passages Across a Large Hebrew/Aramaic Corpus

Authors: Shmidman, Avi and Koppel, Moshe and Porat, Ely

We propose a method for efficiently finding all parallel passages in a large corpus, even if the passages are not quite identical due to rephrasing and orthographic variation. The key ideas are the representation of each word in the corpus by its two most infrequent letters, finding matched pairs of strings of four or five words that differ by at most one word and then identifying clusters of such matched pairs. Using this method, over 4600 parallel pairs of passages were identified in the Babylonian Talmud, a Hebrew-Aramaic corpus of over 1.8 million words, in just over 30 seconds. Empirical comparisons on sample data indicate that the coverage obtained by our method is essentially the same as that obtained using slow exhaustive methods.

2) Interactive Tools and Tasks for the Hebrew Bible

Authors: Winther-Nielsen, Nicolai

Computer-aided processing of intertextuality offers new promising tools for the visualization of the sources and for the performance of tasks that would barely be possible to do without a digital solution. This contribution explores how the corpus of the Hebrew Bible created and maintained by the Eep Talstra Center for Bible and Computer can support new ways in which we can learn from our ancient texts as modern knowledge workers. It first describes how the corpus was used for the development of Bible Online Learner as a persuasive technology to enhance language learning with, in, and around a database that drives interactive tasks for learners. The achievements obtained through this project so far are very promising. Interactive corpus-technology also has an important bearing on the task of textual criticism as an increasingly specialized area of research that depends on the availability of digital resources. Commercial solutions developed by software companies like Logos offer advanced digital scholarly resources from the German Bible Society as a useful alternative to often inaccessible and expensive print versions. Corpus-driven learning and new digital resources will also allow scholars to do new academic tasks in textual criticism and interpretation, and we already now see promising tools for text categorization, analysis of translation shifts, and interpretation

emerge as the potential models for the future. The main goal in the future will be to provide easier and more affordable global access for these new tools.

3) Preprocessing Greek Papyri for Linguistic Annotation

Authors: Vierros, Marja and Henriksson, Erik

Greek documentary papyri form an important direct source for Ancient Greek. It has been exploited surprisingly little in Greek linguistics due to a lack of good tools for searching linguistic structures. This article presents a new tool and digital platform, " Sematia " , which enables transforming the digital texts available in TEI EpiDoc XML format to a format which can be morphologically and syntactically annotated (treebanked), and where the user can add new metadata concerning the text type, writer and handwriting of each act of writing. An important aspect in this process is to take into account the original surviving writing vs. the standardization of language and supplements made by the editors. This is performed by creating two different layers of the same text. The platform is in its early development phase. Future developments, such as tagging linguistic variation phenomena as well as queries performed within Sematia, are discussed at the end of the article.

4) Text Alignment in Ancient Greek and Georgian: A Case-Study on the First Homily of Gregory of Nazianzus

Authors: Pataridze, Tamara and Kindt, Bastien

This paper discusses the word level alignment of lemmatised bitext consisting of the Oratio I of Gregory of Nazianzus in its Greek model and Georgian translation. This study shows how the direct and empirical observations offered by an aligned text enable an accurate analysis of techniques of translation and many philological parameters of the text.

5) Recurrent Pattern Modelling in a Corpus of Armenian Manuscript Colophons

Authors: Van Elverdinghe, Emmanuel

Colophons of Armenian manuscripts are replete with yet untapped riches. Formulae are not the least among them: these recurrent stereotypical patterns conceal many clues as to the schools and networks of production and diffusion of books in Armenian communities. This paper proposes a methodology for exploiting these sources, as elaborated in the framework of a PhD research project about Armenian colophon formulae. Firstly, the reader is briefly introduced to the corpus of Armenian colophons and then, to the purposes of our project. In the third place, we describe our methodology, relying on lemmatization and modelling of patterns into automata. Fourthly and finally, the whole process is illustrated by a basic case study, the occasion of which is taken to outline the kind of results that can be achieved by combining this methodology with a philologico-historical approach to colophons.

6) From manuscript catalogues to a handbook of Syriac literature: Modeling an infrastructure for Syriaca.org

Authors: Gibson, Nathan P. and Michelson, David A. and Schwartz, Daniel L.

Despite increasing interest in Syriac studies and growing digital availability of Syriac texts, there is currently no up-to-date infrastructure for discovering, identifying, classifying, and referencing works of Syriac literature. The standard reference work (Baumstark's Geschichte) is over ninety years old, and the perhaps 20,000 Syriac manuscripts extant worldwide can be accessed only through disparate catalogues and databases. The present article proposes a tentative data model for Syriaca.org's New Handbook of Syriac Literature, an open-access digital publication that will serve as both an authority file for Syriac works and a guide to accessing their manuscript representations, editions, and translations. The authors hope that by publishing a draft data model they can receive feedback and incorporate suggestions into the next stage of the project.

7) Integrated Sequence Tagging for Medieval Latin Using Deep Representation Learning

Authors: Kestemont, Mike and De Gussem, Jeroen

In this paper we consider two sequence tagging tasks for medieval Latin: part-of-speech tagging and lemmatization. These are both basic, yet foundational preprocessing steps in applications such as text re-use detection. Nevertheless, they are generally complicated by the considerable orthographic variation which is typical of medieval Latin. In Digital Classics, these tasks are traditionally solved in a (i) cascaded and (ii) lexicon-dependent fashion. For example, a lexicon is used to generate all the potential lemma-tag pairs for a token, and next, a context-aware PoS-tagger is used to select the most appropriate tag-lemma pair. Apart from the problems with out-of-lexicon items, error percolation is a major downside of such approaches. In this paper we explore the possibility to elegantly solve these tasks using a single, integrated approach. For this, we make use of a layered neural network architecture from the field of deep representation learning.

8) Measuring and Mapping Intergeneric Allusion in Latin Poetry using Tesserae

Authors: Burns, Patrick J.

Most intertextuality in classical poetry is unmarked, that is, it lacks objective signposts to make readers aware of the presence of references to existing texts. Intergeneric relationships can pose a particular problem as scholarship has long privileged intertextual relationships between works of the same genre. This paper treats the influence of Latin love elegy on Lucan's epic poem, Bellum Civile, by looking at two features of unmarked intertextuality: frequency and distribution. I use the Tesserae project to generate a dataset of potential intertexts between Lucan's epic and the elegies of Tibullus, Propertius, and Ovid, which are then aggregrated and mapped in Lucan's text. This study draws two conclusions: 1. measurement of intertextual frequency shows that the elegists contribute fewer intertexts than, for example, another epic poem (Virgil's Aeneid), though far more than the scholarly record on elegiac influence in Lucan would suggest; and 2. mapping the distribution of intertexts confirms previous scholarship on the influence of elegy on the Bellum Civile by showing concentrations of matches, for example, in Pompey and Cornelia's meeting before Pharsalus (5.722-815) or during the affair between

Caesar and Cleopatra (10.53-106). By looking at both frequency and proportion, we can demonstrate systematically the generic enrichment of Lucan's Bellum Civile with respect to Latin love elegy.

9) A Hackathon for Classical Tibetan

Authors: Almogi , Orna and Dankin , Lena and Dershowitz , Nachum and Wolf , Lior

We describe the course of a hackathon dedicated to the development of linguistic tools for Tibetan Buddhist studies. Over a period of five days, a group of seventeen scholars, scientists, and students developed and compared algorithms for intertextual alignment and text classification, along with some basic language tools, including a stemmer and word segmenter.

Part 2: Managing different types of text re-uses

This part focuses on the conceptual definitions, the modelling of the unstable idea of "quotation" and the XML-TEI encoding to implement for its characterization.

1) TEI-encoding of text reuses in the BIBLINDEX Project

Authors: Hue-Gay, Elysabeth and Mellerin, Laurence and Morlock, Emmanuelle

This paper discusses markup strategies for the identification and description of text reuses in a corpus of patristic texts related to the BIBLINDEX Project, an online index of biblical references in Early Christian Literature. In addition to the development of a database that can be queried by canonical biblical or patristic references, a sample corpus of patristic texts has been encoded following the guidelines of the TEI (Text Encoding Initiative), in order to provide direct access to quoted and quoting text passages to the users of the https://www.biblindex.info platform.

2) Encoding (inter)textual insertions in Latin "grammatical commentary"

Authors: Bureau, Bruno and Nicolas, Christian and Pinche, Ariane

The ancient commentaries provide a large sample of quotations from classical or biblical texts for which Latin gramamrians developed a complex system of insertion of quoted texts. The paper examines how to encode these places using XML Tei, and focuses on difficult cases, such as inaccurate quotations, or quotations of partly or wholly lost texts.

3) A Classification of Manuscripts Based on A New Quantitative Method. The Old Latin Witnesses of John's Gospel as Text Case

Authors: Pastorelli, David

A new method for grouping manuscripts in clusters is presented with the calculation of distances between readings, then between witnesses. A classification algorithm (" Hierarchical Ascendant Clustering "), achieved through computer-aided processing, enables the construction of trees illustrating the textual taxonomy obtained. This method is applied to the Old Latin witnesses of the Gospel of John, and, in order to provide a study of a reasonable size, to a chapter as a whole (chapter 14). The result basically confirms the text-types identified by Bonatius Fischer, founder of the Vetus Latina Institute, while it invalidates the classification adopted by the current edition of the Vetus Latina of the Gospel of John.

Part 3: Visualisation of intertextuality and text reuse

1) Version Variation Visualization (VVV): Case Studies on the Hebrew Haggadah in English

Authors: Cheesman, Tom and Roos, Avraham,

The 'Version Variation Visualization' project has developed online tools to support comparative, algorithm-assisted investigations of a corpus of multiple versions of a text, e.g. variants, translations, adaptations (Cheesman, 2015, 2016; Cheesman et al., 2012, 2012-13, 2016; Thiel, 2014; links: www.tinyurl.com/vvvex). A segmenting and aligning tool allows users to 1) define arbitrary segment types, 2) define arbitrary text chunks as segments, and 3) align segments between a 'base text' (a version of the 'original' or translated text), and versions of it. The alignment tool can automatically align recurrent defined segment types in sequence.Several visual interfaces in the prototype installation enable exploratory access to parallel versions, to comparative visual representations of versions' alignment with the base text, and to the base text visually annotated by an algorithmic analysis of variation among versions of segments. Data can be filtered, viewed and exported in diverse ways. Many more modes of access and analysis can be envisaged. The tool is language neutral. Experiments so far mostly use modern texts: German Shakespeare translations. Roos is working on a collection of approx. 100 distinct English-language translations of a Hebrew text with ancient Hebrew and Aramaic passages: the Haggadah (Roos, 2015).

2) Visualizing linguistic variation in a network of Latin documents and scribes

Authors: Korkiakangas , Timo and Lassila , Matti

This article explores whether and how network visualization can benefit philological and historical-linguistic study. This is illustrated with a corpus-based investigation of scribes' language use in a lemmatized and morphologically annotated corpus of documentary Latin (Late Latin Charter Treebank, LLCT2). We extract four continuous linguistic variables from

LLCT2 and utilize a gradient colour palette in Gephi to visualize the variable values as node attributes in a trimodal network which consists of the documents, writers, and writing locations underlying the same corpus. We call this network the "LLCT2 network". The geographical coordinates of the location nodes form an approximate map, which allows for drawing geographical conclusions. The linguistic variables are examined both separately and as a sum variable, and the visualizations presented as static images and as interactive Sigma.js visualizations. The variables represent different domains of language competence of scribes who learnt written Latin practically as a second-language. The results show that the network visualization of linguistic features helps in observing patterns which support linguistic-philological argumentation and which risk passing unnoticed with traditional methods. However, the approach is subject to the same limitations as all visualization techniques: the human eye can only perceive a certain, relatively small amount of information at a time.

Part 4: Project presentations

1) QuotationFinder - Searching for Quotations and Allusions in Greek and Latin Texts and Establishing the Degree to Which a Quotation or Allusion Matches Its Source

Authors: Herren, Luc

The software programs generally used with the TLG (Thesaurus Linguae Graecae) and the CLCLT (CETEDOC Library of Christian Latin Texts) CD-ROMs are not well suited for finding quotations and allusions. QuotationFinder uses more sophisticated criteria as it ranks search results based on how closely they match the source text, listing search results with literal quotations first and loose verbal parallels last.

2) Digital Greek Patristic Catena (DGPC). A brief presentation

Authors: Paparnakis, Athanasios and Domouchtsis, Constantinos

The project is to develop a database, which is planned to include all available information on the use of the Bible in the patristic works of Migne's Patrologia Graeca. Utilization of the data will be available through a web page equipped with necessary tools for developing data mining techniques and other methods of analysis. The main aim of the project is to revive the catenae, the ancient exegetical tool for biblical interpretation.

3) Dealing with all types of quotations (and their parallels) in a closed corpus: The methodology of the Project The literary tradition in the third and fourth centuries CE: Grammarians, rhetoricians and sophists as sources of Graeco-Roman literature

Authors: Rodríguez-Noriega, Lucía

The Project The literary tradition in the third and fourth centuries CE: Grammarians, rhetoricians and sophists as sources of Graeco-Roman literature (FFI2014-52808-C2-1-P) aims to trace and classify all types of quotations, both explicit (with or without mention of the author and/or title) and hidden, in a corpus comprising the Greek grammarians, rhetoricians and " sophists " of the third and fourth centuries CE. At the same time, we try to detect whether or not these are first-hand quotations, and if our quoting authors (28 in all) are, in turn, secondary sources for the same citations in later authors. We also study the philological (textual) aspects of the quotations in their context, and the problems of limits they sometimes pose. Finally, we are interested in the function of the quotation in the citing work. This is the first time that such a comprehensive study of this corpus is attempted. This paper explains our methodology, and how we store all these data in our electronic card-file.

4) Editing New Testament Arabic Manuscripts in a TEI-base: fostering close reading in Digital Humanities

Authors: Clivaz, Claire and Schulthess, Sara and Sankar, Martial

If one is convinced that " quantitative research provides data not interpretation " [Moretti, 2005, 9], close reading should thus be considered as not only the necessary bridge between big data and interpretation but also the core duty of the Humanities. To test its potential in a neglected field – the Arabic manuscripts of the Letters of Paul of Tarsus – an enhanced, digital edition has been in development as a progression of a Swiss National Fund project. This short paper presents the development of this edition and perspectives regarding a second project. Based on the Edition Visualization Technology tool, the digital edition provides a transcription of the Arabic text, a standardized and vocalized version, as well as French translation with all texts encoded in TEI XML. Thanks to another Swiss National Foundation subsidy, a new research project on the unique New Testament, trilingual (Greek-Latin-Arabic) manuscript, the Marciana Library Gr. Z. 11 (379), 12th century, is currently underway. This project includes new features such as " Textlink " , " Hotspot " and notes: HumaReC.

5) Bioinformatics and Classical Literary Study

Authors: Chaudhuri, Pramit and Dexter, Joseph P.

This paper describes the Quantitative Criticism Lab, a collaborative initiative between classicists, quantitative biologists, and computer scientists to apply ideas and methods drawn from the sciences to the study of literature. A core goal of the project is the use of computational biology, natural language processing, and machine learning techniques to investigate authorial style, intertextuality, and related phenomena of literary significance. As a case study in our approach, here we review the use of sequence alignment, a common technique in genomics and computational linguistics, to detect intertextuality in Latin literature. Sequence alignment is distinguished by its ability to find inexact verbal similarities, which makes it ideal for identifying phonetic echoes in large corpora of Latin texts. Although especially suited to Latin, sequence alignment in principle can be extended to many other languages.

6) Computer - Assisted Processing of Intertextuality in Ancient Languages

Authors: Hedges, Mark and Jordanous, Anna and Lawrence, K. Faith and Roueché, Charlotte and Tupman, Charlotte

The production of digital critical editions of texts using TEI is now a widely-adopted procedure within digital humanities. The work described in this paper extends this approach to the publication of gnomologia (anthologies of wise sayings) , which formed a widespread literary genre in many cultures of the medieval Mediterranean. These texts are challenging because they were rarely copied straightforwardly ; rather , sayings were selected , reorganised , modified or re-attributed between manuscripts , resulting in a highly interconnected corpus for which a standard approach to digital publication is insufficient. Focusing on Greek and Arabic collections , we address this challenge using semantic web techniques to create an ecosystem of texts , relationships and annotations , and consider a new model – organic , collaborative , interconnected , and open-ended – of what constitutes an edition. This semantic web-based approach allows scholars to add their own materials and annotations to the network of information and to explore the conceptual networks that arise from these interconnected sayings .

As this special issue allows continuous updates, it is still possible to add a contribution if you are working on these topics.

7) Using the Text Alignment Network for Scholarship on Intertextuality

Authors: Kalvesmaki, Joel

The Text Alignment Network (TAN) is a suite of XML encoding formats intended to serve anyone who wishes to encode, exchange, and study translations, paraphrases, adaptations, quotations, and other varieties of text reuse. This article briefly introduces TAN, and in the spirit of the special issue of this journal focuses on the syntax of its intertextual pointers, which are styled to be both human-readable and -interoperable. Because TAN is at present an experimental format, this report notes progress, promise, and future prospects.

8) Processing Tools for Greek and Other Languages of the Christian Middle East

Authors: Kindt, Bastien

This paper presents some computer tools and linguistic resources of the GREgORI project. These developments allow automated processing of texts written in the main languages of the Christian Middel East, such as Greek, Arabic, Syriac, Armenian and Georgian. The main goal is to provide scholars with tools (lemmatized indexes and concordances) making corpus-based linguistic information available. It focuses on the questions of text processing, lemmatization, information retrieval, and bitext alignment.

Contact

Marco Büchler: mbuechler(at)etrap(dot)eu

Laurence Mellerin: laurence.mellerin(at)mom(dot)fr

See http://jdmdh.episciences.org/volume/view/id/158