TECHNISCHE UNIVERSITÄT DARMSTADT
DIGITAL HUMANITIES COOPERATION
UNIVERSITÄT KONSTANZ

Marie Revellio

Classics and the Digital Age.

Advantages and limitations of digital text analysis in classical philology

LitLab

Pamphlet #2

October 2015

Ed. Thomas Weitin

FORSCHUNGSSTELLE
DIGITAL HUMANITIES CENTER
UNIVERSITÄT KONSTANZ

Marie Revellio

Classics and the Digital Age

Advantages and limitations of digital text analysis in classical philology

Abstract

Die Klassische Philologie nahm computergestützte Methoden der Textanalyse früh als Chance wahr. Um einen Einblick in die jüngsten Entwicklungen der digitalen Textanalyse im Bereich der Latinistik zu geben wird eine Auswahl bestehender Textdatenbanken wie gängiger Analysetools vorgestellt, wobei insbesondere auf das Phänomen der Intertextualität als Untersuchungsfeld fokussiert wird. Zudem werden unmittelbar verknüpfte Themen wie die Digitalisierung und langfristige Erhaltung antiker Texte, der Status unterschiedlicher Text-Surrogate sowie die Notwendigkeit fremdsprachlicher Kenntnisse diskutiert.

Classical philology adopted computer-assisted research methods very early. In order to provide first insights into recent developments of digital Latin scholarship, several digital text collections and commonly used tools are introduced, focusing especially on research questions concerning the phenomenon of intertextuality. In addition, adjacent issues as digitization and long-term preservation of ancient texts, the status of different text surrogates and the need for language skills are discussed.

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie. Sie ist in der Zeitschriftendatenbank (ZDB) und im internationalen ISSN-Portal erfasst. Detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar. Alle Rechte, auch die des auszugsweisen Nachdrucks, der fotomechanischen Wiedergabe und der Über- setzung, vorbehalten. Dies betrifft auch die Vervielfältigung und Übertragung einzelner Textabschnitte, Zeichnungen oder Bilder durch alle Verfahren wie Speicherung und Übertragung auf Papier, Transparente, Filme, Bänder, Platten und andere Medien, soweit es nicht §§ 53 und 54 UrhG ausdrücklich gestatten.

©2015 Marie Revellio, marie.revellio@uni-konstanz.de

ISSN: 2628-4537

Ed. by Thomas Weitin

Marie Revellio

Classics and the Digital Age

Advantages and limitations of digital text analysis in classical philology

Latin philology encompasses the study of all kinds of texts written in Latin from its origin in 240 BC to contemporary Neo-Latin. A crucial part of the canonical subjects is up to 2000 years old and most often preserved in various artifacts, for instance in medieval manuscripts. Immediately the idea of a fogeyish and in some way old-fashioned philologist comes in one's mind, hunting in monastic libraries and shady archives for disappeared manuscripts.[1] Is this humanistic prototype still state-of-the-art? What happens with the Ancient World in the Digital Age? Is there anything like Classics 2.0?

Typically the beginning of humanities computing is bound to the Jesuit Father Roberto Busa, who created an extensive index of Saint Thomas Aquinas's *opera omnia*. Busa started working on the so-called *Index Thomisticus*[2] using punched cards and counting machines in the late 1940s.[3] Accordingly, in 2015 the field of Classics has already celebrated the diamond anniversary of the union from computers and Classics. With the advent of personal computers in the 1990s and the rise of the Internet digital research tools were easier accessible than ever, particular for private users, and are these days ubiquitous in even – let's say – 'classic' work in Classics.

Some subject-specific hallmarks might have supported the immediate adoption of computing in the study of Latin language and literature: The textual tradition of ancient Greek and Latin works is particular complex. They came down to us on different materials like stone,

[1] Cf. the vivid and to some extent mystified depiction of Poggio Bracciolini by Stephen Greenblatt in his book *The Swerve: How the World became modern*, New York, London: Norton 2011. Bracciolini, after the council of Constance in 1414 not any longer the papal secretary but a book hunter, found among other prominent works the to date only manuscript of Lucretius's *De rerum natura* in a German monastery and fits in the (long) persistent idea of humanists rescuing the classics out of medieval darkness.

[2] Robertus Busa: *Index Thomisticus. Sancti Thomae Aquinatis operum omnium indices et concordantiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur quaeque / consociata plurium opera atque electronico IBM automato usus digessit Robertus Busa*, Stuttgart-Bad Cannstatt: Frommann 1974. The printed version consists of 56 volumes, for the online version cf. http://www.corpusthomisticum.org/. All URLs were last accessed September 30th, 2015.

[3] Cf. Roberto Busa: „Foreword: Perspectives on the Digital Humanities." In: Susan Schreibman, Ray Siemens, John Unsworth (eds.): *A Companion to Digital Humanities*, Malden, Oxford: Blackwell 2004, p. xvi–xxi, p. xvii. Another prominent example of early computer-based research in Classics is the project of the Thesaurus Linguae Graecae (TLG), cf. the report of Theodore Brunner: „Classics and the Computer: The History of a Relationship." In: Jon Solomon (ed.): *Accessing Antiquity. The Computerization of Classical Studies*, Tucson, London: The University of Arizona Pres 1993, p. 10–33.

papyrus rolls or in medieval codices. Frequently, there are only fragments preserved, partly mixed with other texts covering the original provenance. This implicates two things: in order to handle the sources in a most suitable manner collaboration across many different disciplines like papyrology, codicology, paleography and epigraphy is inevitable. Furthermore, an essential task of classical philologists is to compile an accurate text version considering the preserved artefacts of a work. In order to make the editors preferred reading transparent and to assure intersubjective traceability, editions with a critical apparatus are arranged. Another challenge is the temporal distance to the historic author and his contemporary audience. In order to face this issue concordances, indices, thesauri and line-by-line commentaries (explaining words and phrases through quoting parallels from other texts) are prepared. Regarding the required searching, collecting, classification and ordering processes and the idea of treating a text as a *bag of words*, such traditional tasks are very suitable for computational support – especially in terms of accuracy, homogeneity, time and scale.

What does digital text analysis in Classics mean and what are the requirements? Which tools are commonly used? How do they work and which output do they provide? In the following a brief outline of common digital text collections and tools is given, in order to allow first insights into the field of digital Latin scholarship. Regarding the long history of computing in Classics this is not the place for an extensive full-scale presentation of all tools and projects.[4] Moreover, applying digital methods is per se as fragile as applying every other method, so the purpose of this work is to illustrate opportunities as well as limitations of digital Latin philology.

1. Hypothesis and Textual Data

To start from scratch, a research project usually begins with a research question and subsequent clearly stated hypotheses. Due to the fact that computers analyze texts as a sequence of alphanumeric characters (and spaces), only phenomena in some way connected to the textual

[4]Cf. for a concise overview of the field Alison Babeu: „Rome Wasn't Digitized in a Day": *Building a Cyberinfrastructure for Digital Classicists*, CLIR Publication No. 150 (2011), online available on http://www.clir.org/pubs/abstract/pub150abst.html; Sarah Buchanan: „The Emerging Tradition of Digital Classics." In: Samantha Hastings (ed.): *Annual Review of Cultural Heritage Informatics* 2014 (vol. 2), Lanham: Rowman & Littlefield 2015, p. 149-163. For information on further tools and recent projects cf. http://www.digitalclassicist.org/; for actual debates e.g. on digital editing cf. the discussion list of the Digital Classicists community.

surface can be focused on. As far as my experience goes, including a universities seminar given on digital text analysis in summer 2015, surprisingly many questions can in whole or at least in parts be traced back to such surface characteristics – contrary to common initial doubts.

A fundamental question of classical philology concerns the phenomenon of intertextuality. The term *intertextualité* was first utilized by Julia Kristeva in 1966 and the framing concepts were also elaborated by M. Bakhtin and R. Barthes.[5] The basic idea of this concept traces in many ways ancient reading practices of Greek and Latin texts, that is to find kinds of relations between texts, so-called *loci similes* – the most prominent correspondences on the word level. This similarity can be detected through comparison of two texts, which is traditionally done manually. However, a major benefit of computers is that they process fair amounts of texts in less time and in a more systematic way than ever possible for humans. A popular example that makes use of these advantages is software for plagiarism detection. The examination of references which are beyond these lexical characteristics, especially on the level of syntax, metrics, rhythm, sound, motives, topoi and work or even Œuvre structure is much more challenging. In case of computer-assisted analysis appropriate prepared texts with homogeneous annotations are needed. Therefore, a thrilling task is to determine the adequate digital text data for the analysis. As the degree of annotation determines the possible searches, different „surrogates“[6] of works are necessary: questions concerning syntactical correspondences need syntactical annotated text versions, questions concerning intertextual references in metrics require scanned text material.

Due to the early use of computer-assisted methods in Classics and the huge number of digitalizing projects in the past years, a vast amount of websites providing machine-readable texts exist. Sometimes only single works from a single author but mostly small corpora with particularly canonical works can be found on these websites. The quality of these texts is highly fluctuating for example in terms of typing errors, wrong or missing verse lines and markers of text locations. Additionally, the structure of the textual data varies between unstructured texts

[5]Kristeva mentioned the term first in her essay *Le mot, le dialogue et le roman* from 1966 edited in: Julia Kristeva: *Σημειωτικὴ. Recherches pour une sémanalyse*, Paris: Seuil 1969, 143-173, p. 146. For further readings on the origin of the concept and the relation to Classical studies cf. Yelena Baraz, Christopher van den Berg: „Introduction“In: *American Journal of Philology*, 134.1 (2013), p. 1–8.

[6]Martin Mueller: „Morgenstern's Spectacles or the Importance of Not-Reading.“ In: *Scalable Reading* (blog), 21.01.2013, https://scalablereading.northwestern.edu/2013/01/21/morgensterns-spectacles-or-the-importance-of-not-reading/. For further implications of the concept on digital text analysis methods cf. Thomas Weitin: „Thinking slowly. Literatur lesen unter dem Eindruck von Big Data.“ In: *Konstanz LitLingLab* Pamphlet No. 1 (2015), p. 10 seq.

in conventional formats like plain texts and structured texts in XML (= eXtensible Markup Language).[7] A few of the digital collections of Latin texts are briefly reviewed in the following. They can be divided in two groups: open source and licensed corpora.

The ***Latin Library*** is an open source collection and covers a wide range of epochs containing classical, Christian, medieval and Neo-Latin works.[8] The texts are drawn from varying websites or were scanned and sent in by contributors of the community. Often the specification of the used edition is lacking[9] and the texts contain several typing errors. The ***Packhard Humanities Institute (PHI)*** claims to hold a corpus of all works written in Latin till 200 AD and some of Late Antiquity.[10] These texts are fed in professionally, hold consistent verse and chapter references and clear specification of the editions used. In contrast, the aim of the project ***Musisque deoque*** is to include not only the text of one reference edition but also the variants of its critical apparatus.[11] It encompasses all Latin poetry from its origins to the Italian Renaissance. Another professional digital corpus holds the ***Perseus Digital Library*** presenting texts marked-up formerly in SGML and recently in XML technology according to the Text Encoding Initiative (TEI) standards.[12] These texts are enhanced by additional information on the edition and are prepared with several extra features like commentaries and translations or semantical, lexical and morphological information. The *Perseus Library* now joined the ***Open Greek and Latin Project***, which aims to compile every work written in ancient Greek or Latin from the archaic age to the modern.[13] In order to display the full textual tradition of a work, a further goal of this project is to represent various versions (text editions) of each work.

A basic licensed database is the ***Library of Latin Texts Series A (LLT-A)***, emerged

[7] XML is a machine-actionable and human-readable markup language to describe data and its structure, for instance the division of a text in paragraphs, which mimics the convenient reference codes in classical scholarship such as 'Ov. ars. 3,251'.

[8] http://www.thelatinlibrary.com/

[9] Cf. the source information for Horace: *'Carmina, Epistulae and Ars Poetica* posted by *full name of contributor'*.

[10] http://latin.packhum.org/

[11] http://www.mqdq.it, cf. also Massimo Manca, Linda Spinazzè, Paolo Mastandrea, Luigi Tessarolo, Federico Boschetti: „Musique Deoque: Text Retrieval on Critical Editions." In: *Journal for Language Technology and Computational Linguistic* 26.2 (2011), p. 129–140.

[12] http://www.perseus.tufts.edu/hopper/collections, cf. also David Smith, Jeffrey Rydberg-Cox, Gregory Crane: „The Perseus Project: a Digital Library for the Humanities." In: *Literary and Linguistic Computing* 15.1 (2000), p. 15-25. The Text Encoding Initiative develops XML standards for encoding texts in digital form, cf. http://www.tei-c.org/index.xml.

[13] http://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/

from CETEDOC Library of Christian Latin Texts.[14] It contains all written Latin literature till 1965 with special focus on Christian texts. For this database the text sources are indicated and the text quality is mostly accurate. Furthermore, there exist several collections maintained by publishers echoing the texts form their printed editions, even if the critical apparatus is missing these texts are very reliable. The *Bibliotheca Teubneriana Latina* ***(BTL)***, the *Patrologia Latina* ***(PL)*** and the *Loeb Classical Library* ***(LCL)*** should be mentioned in this context. Surprisingly, conventional locus numbering schemas, which are usually provided in the printed versions, are absent for instance in the electronic version of the *BTL* texts which complicates the correct labeling of passages.

In spite of the relatively limited canon of preserved Latin texts – in comparison with modern languages – to date not all texts produced in the long history of the Latin language can be found digital and in appropriate quality.[15] Creating one's own text file out of a print version can be done semi-automatically via scanning, OCR and ICR processing (= optical and intelligent character recognition to extract letters and word forms from scanned pictures), ideally converting the text data into XML documents conforming to the latest TEI standard and finally correcting them manually.[16] This last step is inevitable due to the actual technical standard and takes plenty of time. Necessary corrections concern for example misrecognized characters, troubles with the recognition of Greek letters or wrong and inconsistent TEI XML-tags. A major advantage of semi-automatic processes is that errors follow a certain regularity, which can therefore be detected more easily. Finally, the text file can be made available for the interested community, for instance on *GitHub* repositories as found for the *Open Greek and Latin Project*.

However, working with XML can be frustrating at first. Fortunately, the most commonly used XML-editor *oXygen*[17] and even more open source applications like *BaseX*[18] provide a

[14] http://clt.brepolis.net/llta/Default.aspx

[15] Insufficient quality concerns not only literary but also non-literary Latin texts as for example inscriptions. Besides the problem of typos several epigraphic databases do not note information as date and type of inscription or place of discovery consistently, as Miriam Bastian, a participant of my course „Digital text analysis"taught in summer 2015 at the University of Constance, noticed in her work on the *damnatio memoriae* of the emperor Domitian.

[16] For further reading on scanning and OCR technology for historical languages cf. Michael Piotrowski: *Natural Language Processing for Historical Texts*, Synthesis lectures on human language technologies 17, San Rafael: Morgan & Claypool 2012, p. 28 seqq., for a discussion of digitizing classical critical editions cf. 34 seqq.

[17] http://www.oxygenxml.com/

[18] http://basex.org/, cf. Christian Grün: *Storing and Querying Large XML Instances*, Konstanz 2010.

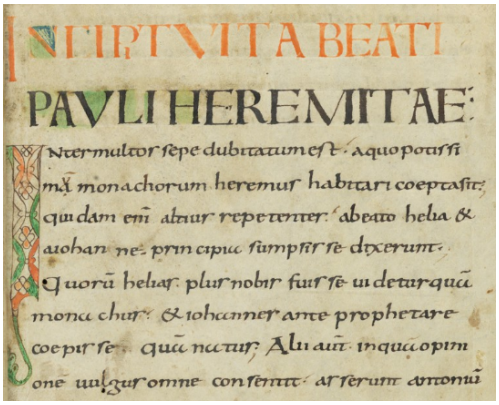
simple and user-friendly graphical interface. *BaseX* provides easy-to-use visualization of the data structure and further assisted XQuery possibilities. These are necessary to execute queries on the metadata presenting for instance the work structure and chapter numbering or to extract the very same in combination with the Latin text itself. Both applications support TEI XML and therefore EpiDoc, a TEI subset of markups to encode ancient documents.[19] The EpiDoc schema includes for example specialized tags for manuscripts materiality, erroneously transposed characters or scholia (comments around the text).

2. Digitization and Preservation or What is a „Text"?

The fact that several databases provide only the text of a single edition without the critical apparatus undermines the netted working practice of classical philology. In addition, digital critical editions often try to mimic their printed counterparts and do not (yet) utilize digital advantages like dynamic comparison of various editions.[20] Therefore, it remains thrilling whether digital critical editions can compete with their printed ancestors in everyday work. Furthermore, classical philologists do not have the text of a work. The more detailed one looks at a particular text passage the more unstable it gets: A single word form can begin to oscillate, when the history with all the *scribae* copying manually the text documents over the years is taken into account. So distinction of a work (all textual surrogates) and a text (a single variant) is taken for granted in the in-depth text analysis, as studying the own research history is an immanent part of the field.

[19] For the latest EpiDoc schema and guidelines cf.: http://sourceforge.net/p/epidoc/wiki/Home/

[20] Cf. the very variable design of the Library of Digital Latin Texts Project, which aims to establish an open source library with born-digital critical editions of Latin texts: http://digitallatin.org/library-digital-latin-texts



VITA BEATI PAVLI
MONACHI THEBAEI

1. 1. Inter multos saepe dubitatum est a quo potissimum monachorum eremus habitari coepta sit. Quidam enim altius repetentes, a beato Elia et Ioanne principia sumpserunt. Quorum et Elias plus nobis uidetur fuisse quam monachus, et Ioannes ante prophetare coepisse quam natus est [a]. 2. Alii autem, in quam opinionem uulgus omne consentit, adserunt Antonium huius propositi caput, quod ex parte uerum est. Non enim tam ipse ante omnes fuit, quam ab eo omnium incitata sunt studia. Amathas uero et Macarius, discipuli Antonii, e quibus superior corpus magistri sepeliuit, etiam nunc adfirmant, Paulum quemdam Thebaeum principem rei istius fuisse, non nominis, quam opinionem nos quoque probamus. 3. Nonnulli et haec et alia, prout

AB CD E G IK LM N O S T UW Deg. r v

1. 2 coepta sit : coeptum *W* ‖ 3 principia sumpserunt : principia sumsisse dixerunt *N* sumpsere principium *v* ‖ 4 plus propheta nobis *L* ‖ uidetur fuisse ~ *N* ‖ 5 prophetare coepisse : ~ *E* prophetare recoepisse *W* > *O* ‖ 7 adserunt : et *praem. B* adseuerant *D* adsu (mut) *C* asseruerunt *S* asserunt *W* ‖ 9 omnium > *B* ‖ incita *C* ‖ amathas : amatus *CLMS* amathias *N* ‖ 10 corpus magistri : ~ *Nv* + sui *L* ‖ 11 thebaidem (thebaidum m^2 *corr. M*) *LM* ‖ 12 rei istius ~ *N*

Fig. 1: The beginning of Jerome's vita of the hermit Paulus: A manuscript from c. 800 called N in a critical edition[21]

As to date, working with digital text analysis methods mostly means to work only on one (digital available) text variant, this condition should at least be kept in mind and included into the final conclusions.

Directly the question of preservation adjoins. In classical philology preservation concerns temporal distances of hundreds or even thousands of years, which should also be expected for the future. A still readable parchment codex made of goatskin from around 900 AC has to be compared with optical media as CD-ROM or DVD, which are suspected to last no more than some decades.[22] So, technical progress and even more institutional infrastructure and support is required to increase at least the probability of digital texts preservation and enable ongoing scholarly practice.

3. Tools

Typically, digital Latinists use more than one tool for a single project and often switch between them. The appropriate tools depend a lot on the research patterns derived from the hypotheses: examining the use of a special word category or metric peculiarities requires different tools. A

[21] MS and critical edition are as follows: St. Gallen, Stiftsbibliothek, Cod. Sang. 558, p. 3a. (http://www.e-codices.unifr.ch/de/list/one/csg/0558) and Pierre Leclerce, Edgardo Morales, Adalbert de Vogüé: *Jérôme. Trois vies de moines (Paul, Malchus, Hilarion)*, Source chrétiennes No. 508, Paris: Les Éditions du Cerf 2007.

[22] Cf. Marilyn Deegan, Simon Tanner: „Key issues in digital preservation." In: id. (eds.): *Digital preservation*, London: Facet 2006, p. 1-31, p. 14.

full-text search for single words in context or collocations is typically possible through the databases‘ user interfaces. Often they provide also tables with quantitative word counts and word distributions. In contrast, linguistic processing, such as lemmatization and morphologic or syntactic analyzes, needs more sophisticated tools. While in the beginning of Digital Classics the main aim was to provide electronic access to texts and to create simple retrieval possibilities, recently more complex information retrieval techniques such as text-mining, focusing on certain research areas as for instance on the phenomena of intertextuality, came to the fore.

3.1 Words in Context

To find out when a word first appeared in a given context one can look it up in printed concordances or execute a query to get personalized results in the *keyword in context (KWIC)* format. The ***Thesaurus Linguae Latinae (TLL)*** is the largest monolingual Latin dictionary covering the Latin language from its beginning till 600 AC.[23] It started in the end of 1890 and compiles all instances of a given word in most of the survived texts. Being not yet finished, it exists in print, CD and as an online version. In the latter you can search in lemma-articles for occurrences of single word forms and determine the query to single systematic categories.

PHI provides *Word Search* and *Concordance* facilities. As the *TLL* the *PHI* tool processes only sequences of letters. Therefore, logical operators and filters enable to define word breaks as well as to set the distance between matching results or to restrict searches to some authors and works only. The *Word Search* lists all results sorted by author and work in a *KWIC* format. Additionally, absolute and relative frequencies of letter sequences per author can be retrieved. At any time jumping into the continuous text is possible to ensure close reading of passages. The search functionalities of the ***LLT-A*** user interface are more advanced. It allows to search for lemma, to determine the period, author and title, to set restriction with wildcards and to determine the order of the searched forms.

With tools like these not only questions of semantics, change in language use and textual tradition but also questions concerning author or genre specific vocabulary use can be addressed. Another serious benefit is that a lot of instances for further in-depth study can be generated

[23] http://www.degruyter.com/db/tll, cf. for the archive‘s material and the article‘s parts *Thesaurus linguae latinae. Editus iussu et auctoritate consilii ab academiis societatibusque diversarum nationum electi, praemonenda de rationibus et usu operis*, Leipzig: Teubner 1990.

easily and systematically.

3.2 Linguistic Preprocessing of Latin Texts

Computers, as humans, do not speak any language without learning. Therefore, in order to quit the first level of processing language as alphanumeric signs and entering the word level, linguistic analysis and mark-up is necessary. A first step in morphologic analysis is to reduce the word to its stem by cutting away all flectional endings and in this way to retrieve the type (i.e. lemma) from the token (word form): the word forms *vult* and *veli*-s belong both to the lemma *volo*.

Automatic lemmatization of Latin texts can be done for example with ***Collatinus***.[24] It lemmatizes single word forms as well as whole texts by showing the supposed lemmata either in order of appearance in the text, alphabetically or in a list providing frequencies of the lemmata and forms. Plain text files can also be analyzed with ***LemLat***.[25] This tool provides not only all possible lemmata of every token but also allows some morphologic analysis of mood, tense, case, gender, number and person. The same morphosyntactic analysis is provided by ***Morpheus***, the morphological parsing and lemmatizing tool of the Perseus Project, which is implemented in the Perseus website as the *Latin Word Study Tool*.[26] *LemLat* and the *Latin Word Study Tool* allow XML export of the results for further processing by way of XQuery.

Annotations of the word categories – commonly labeled as noun, verb or preposition – can be done with a so-called *Part-of-Speech (PoS)* tagger. Since word forms can be ambiguous, like in *cur velis* (verb) *vivere* and *velis* (noun) *volamus*, the *PoS* depends also on the context of the word, especially in highly inflected languages as Latin with on top very flexible word order. Therefore *PoS*-tools often collect information on grammatical and semantic context (case, number or lemma), too. The language independent ***TreeTagger***[27] provides two parameters for Latin: a more simple tag set (providing 2 categories) trained on classical, late antique and medieval Latin and a more detailed tag set (providing 11 categories) trained only on the

[24] http://outils.biblissima.fr/collatinus/
[25] http://www.ilc.cnr.it/lemlat/
[26] http://www.perseus.tufts.edu/hopper/morph?redirect=true&lang=latin
[27] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/, cf. Helmut Schmid: „Probabilistic Part-of-Speech Tagging Using Decision Trees." In: *Proceedings of International Conference on New Methods in Language Processing* (1994).

medieval *Index Thomisticus*. Running the first without further training on classical texts is more accurate than the second, but both need manually correction and final disambiguation.[28]

To address questions of style, syntactically annotated texts are necessary, which can be found in *Treebanks*. For Latin texts there exist three manually annotated *Treebanks*: The ***Latin Dependency Treebank*** of the Perseus Project contains texts of classic prose and poetry.[29] The ***Proiel Treebank*** started working on translations of the New Testament and now encompasses also classical works of Caesar, Cicero and the late antique *Itinerarium Egeriae*.[30] The third one to be named is the ***Index Thomisticus Treebank*** with more than 16000 annotated sentences of three works by Thomas Aquinas.[31]

Linguistically preprocessed texts offer an additional layer of information and can be used in language independent tools like *Antconc* or the *Voyant tools*. As an example, consider working on Livy's *Ab urbe condita* investigating the characterization of generals through the authors stylization of their speeches to the troops. This could be done by firstly lemmatizing Livy's History of Rome with the *TreeTagger*, then examining the lemma lists of the generals' speeches with *Catma* and the *Voyant tools* in regard of vocabulary density and vocabulary use over the course of the speeches and finally comparing the results to the remaining text of Livy. In order to determine the type of speeches, the morphological analysis of the *TreeTagger* can be further used to focus on the incidence of verb forms and pronouns of the 1st and 2nd person only.[32]

[28] For an evaluation of the *TreeTagger* for the Latin language cf. David Bamman, Gregory Crane: „Building a Dynamic Lexicon from a Digital Library." In: *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (2008). For a review on other PoS-Tagger for medieval Latin cf. Steffen Eger, Tim vor der Brück, Alexander Mehler: „Lexicon-assisted tagging and lemmatization in Latin: A comparison of six taggers and two lemmatization methods." In: *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Science and Humanities* (2015), p. 105–113.

[29] http://nlp.perseus.tufts.edu/syntax/treebank/latin.html

[30] http://proiel.github.io/, cf. Dag Haug, Marius Jøhndal, Hanne Eckhoff, Eirik Welo, Mari Hertzenberg, Angelika Müth: „Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages." In: *Traitement automatique de langues* vol. 50.2 (2009) p. 17–45.

[31] http://itreebank.marginalia.it/. There have also been made attempts on automatic syntactic annotation with statistical parsers, for further reading cf. Barbara McGillivray: *Methods in Latin Computational Linguistics*, Leiden, Boston: Brill 2014, p. 28.

[32] Focussing on the speeches of Scipio Africanus the Elder and Fabius Maximus of book 28 only, this research was done by Laura Keller, a participant of the course „Digital text analysis". She investigated the mismatch of 'Scipios' initial claim *de me ipso taceo* (Liv. 28,27,13) and the following rather self-centered speech.

3.3 Text Mining

After the early adoption of digital methods in Classics the scholarly diligence turned mainly to modern languages like English and German, especially regarding more advanced information retrieval such as text data mining. However, there exist several tools to investigate more abstract levels of Latin texts. A typical feature of Latin poetry is, for instance, the absence of rhymes as they are known today. Instead a wide range of metric families, primarily borrowed from the Greeks, are used. ***Collatinus*** automatically analyzes the quantity of syllables with the help of a dictionary and displays a possible scansion. For ambiguous forms the tool displays all versions which have to be cleared manually for further use. The project ***Musisque Deoque*** focusses on poetry written in dactylic verses such as the epic hexameter of Vergil's *Eneid*. With this tool not only intertextual references in rhythm and sound can be investigated, but even accordances in metrical patterns like verse position, hiatuses or synalephas. As a result, the search for metrical and verbal references to e.g. the famous beginning of the *Eneid: Arma virumque cano, Troiae qui primus ab oris* (Verg. Aen. 1,1) is possible:

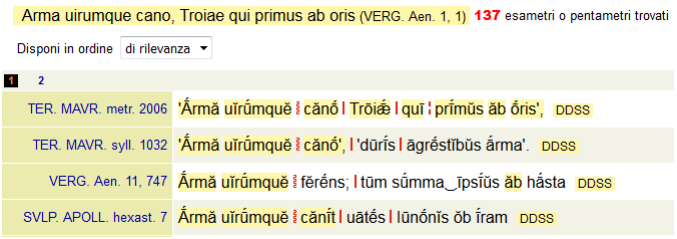


Fig. 2: Co-occurrence search for matched metrical patterns by *Musisque Deoque*

For intertextual investigations in terms of text reuse, i.e. on the lexical level, there are several tools available. ***Tesserae*** for example retrieves all Latin word forms of a source and a target text back to their lemma to find at least two similar words.[33] More than thousands of results are common, so refining the query with filters (e.g. to set stop words in order to neglect probably not meaningful results between *et* and *in*, or to control for the distance between the matching words) is useful to keep the results suitable for philological analysis. In order to facilitate further work a hierarchical order of the results is produced. An advantage concerning the origin of Latin literature is the ability of *Tesserae* to compare Latin with ancient Greek

[33] http://tesserae.caset.buffalo.edu/. Promising runs were made with Lucan's Pharsalia and the Eneid showing already known and also unobserved *loci similes*, cf. Neil Coffee, Jean-Pierre Koenig, Shakthi Poornima, Roelant Ossewaarde: „Intertextuality in the Digital Age." In: *Transactions of the American Philological Association* 142.2 (2012), p. 383–422.

texts. There is even an expansion to the level of synonyms and an experimental tool addressing thematic similarities between two texts lacking any lexical overlap. In contrast, the tool ***Citationgraph*** of the *eAqua* project provides research possibilities on text reuse beyond *loci similes*.[34] *Citationgraph* compares one text to the whole collection of the Ancient Greek *TLG* or, in demonstration mode, to the *BTL*. This allows the investigation of diachronic citation practice over several centuries. The *eAqua* project includes various visualizations and further tools like the text supplement tool, which provides automatic completion of corrupt inscriptions.

Another way of tracing similarities of texts and accordingly identifying the distinctiveness of an author's style concerns the field of stylistics. Computational stylistics focusses on countable features of a text in order to establish a textual 'footprint' of each author, which is especially interesting when attributing anonymous texts – of which plenty exist in classical philology. A prominent tool for this task is the ***stylo*** package for the statistical programming environment *R*.[35] *Stylo* focuses – in brief – on lists of the most frequent words in a text. These lists usually start with function words like *et, in, non, ut, ad*. The underlying hypothesis of this kind of authorship attribution states that the author's use of function words happens unconsciously.[36] The package analyzes the frequencies of the function words with statistical distance measures producing text-pairs of nearest neighbors, as shown in the following example of Tibullus.

Albius Tibullus is a roman love elegist of the Augustan age. The manuscript tradition transmits three books of elegies mentioning Tibullus as their author. The style of the first two books is homogenous and indeed written from the perspective of a speaker who calls himself *Tibullus* (1,3.9). But about the third book there is no consensus regarding the author(s) because it contains poems by a speaker called Lygdamus and by another called Sulpicia. Therefore, in 1838 Otto Gruppe suggested first that the so-called Lygdamus-Elegies (3,1-6) are not Tibullian, but may be written by Ovid because they closely resemble his lyric.[37] We tested this

[34] http://www.eaqua.net/

[35] Maciej Eder, Mike Kestemont, Jan Rybicki: „Stylometry with R: a suite of tools." In: *Digital Humanities 2013 Conference Abstracts*, University of Nebraska-Lincoln (2013), p. 487–489; R Core Team: *R. A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna (2014), cf. http://www.R-project.org/

[36] For the nexus of the author's style and postulated „linguistic homogeneity" in stylometric theory cf. Penelope Gurney, Lyman Gurney: „Authorship Attribution of the *Scriptores Historiae Augustae*." In: *Literary and Linguistic Computing* 13.3 (1998), p. 119-131, p. 121.

[37] Otto Gruppe arguments with similarity in use of collocations, conjunctions etc.; cf. Otto Gruppe: *Die römische Elegie. Erster Band, kritische Untersuchungen mit eingeflochtenen Übersetzungen*, Leipzig: Wigand 1838, p. 133 seq.

heavily discussed hypothesis with *stylo* by comparing Tibullian elegies from book 1 with the Lygdamus-Elegies and parts of Ovid's love elegy *Amores*. The cluster analysis in figure 3 shows the result indicating the nearest neighbors through vertical, their distances through horizontal lines:

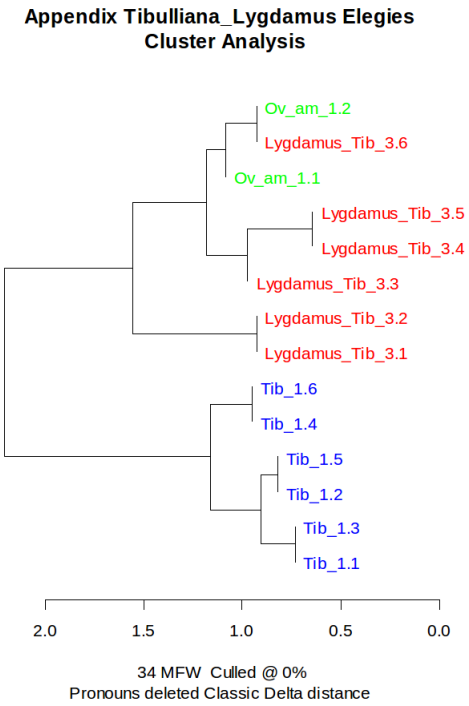


Fig. 3: The so-called Lygdamus-Elegies of the *Appendix Tibulliana* tested with *stylo*

In this setting the Lygdamus-Elegies indeed resemble more Ovid's *Amores* than Tibullus's elegies of the first book – this is not to say that Ovid is the veritable author. Although these first results support the Ovid hypothesis, further research has to be done. The most serious challenge of working on Latin poetry with *stylo* is to handle the sizes of the corpora in case of really short poems (frequently only some verses) because otherwise only the size affects the calculated distance.[38]

3.4 Language Proficiency

Text mining tools like *stylo* have little or nothing to do with the conventional understanding of literary studies that is based on reading as an approach to texts and literature. Perhaps thinking

[38] Working in the course „Digital text analysis" on the very short Sulpicia-Elegies of the *Appendix Tibulliana* Thomas Konrad extracted the bias of the file sizes as the major problem. Cf. for this issue also Maciej Eder: „Does Size Matter? Autorship Attribution, Small Samples, Big Problem." In: *Proceedings of Digital Humanities* (2010), p. 132–135.

of great libraries such as the ancient Library of Alexandria or the contemporary British Library, the following question was raised: „What do you do with a million of books?“[39] Seen in the context of the debate on close and distant reading[40] it is probably a more suggestive question asking for the answer: „You don't read them, because you can't.“[41] Nevertheless, in recent publications on digital literary studies, such as the eponymous Blackwell Companion edited by Ray Siemens and Susan Schreibman in 2007, a bunch of articles focuses precisely on the issue of reading.[42] That is, reading as well as reading skills and therefore language proficiency matters a lot in Digital Humanities. And in fact, there are several applications dedicated to assist in comprehensive reading of Latin and Greek texts or to improve classical language skills.

The *Perseus* tools encourage non-specialists to learn and improve their Greek and Latin skills, so that a wider audience is able to work with ancient texts. *Perseus* provides all texts along with dictionary information and makes it possible via the ***Vocabulary Tool*** to refine a custom vocabulary list of a text matching on the own masteries level. Even estimations on the texts complexity can be made: Text complexity is thereby understood as the vocabulary density, which is calculated as the ratio of the total word count of a work to the number of unique words. So comparing the vocabulary density score of Caesar's *De bello Gallico* (51.295 / 4.535 = 11,311) with Horace's *Carmina* (13.292 / 4.873 = 2,728), the continuous treating of the Gallic War in undergraduate classrooms is justifiable, at least with regard to limited vocabulary.

Language skills concerning syntax can be enhanced with tools like the ***Alpheios Reading tools***.[43] The set of tools is designed as a browser add-on so that it can be used on any website. It enables diagramming of sentences in the form of dependency trees with nodes (words) and arcs (dependency relations between head and dependent words). The dependency trees can be exported in XML for further analysis or usage in learning groups.

Analyzing the syntactic structure in order to read texts properly is a central task of Lati-

[39] Gregory Crane: „What Do You Do with a Million Books?“ In: *D-Lib Magazine* 12.3 (2006), online availabe on http://www.dlib.org/dlib/march06/crane/03crane.html.

[40] Cf. Franco Moretti: „Conjectures on World Literature.“ In: *New Left Review* 1 (2000), p. 54–68.

[41] Tanya Clement, Sara Steger, John Unsworth, Kirsten Uszkalo: *How Not to Read a Million Books*, 2008, online available on http://people.brandeis.edu/ unsworth/hownot2read.html

[42] Ray Siemens, Susan Schreibman (eds.): *A companion to Digital Literary Studies*, Malden, Oxford: Blackwell 2007.

[43] http://alpheios.net/

nists and of examination candidates. Therefore courses in translation skills are immanent to the universities curriculum. A translation seminar on Horace's *Carmina* revealed promising opportunities for the use of Blended Learning methods in the universities curriculum:[44] Working in rotating teams on the *Odes* of the fourth book and communicating via a system of web-based teaching and learning, the participants prepared their most valuable translation of all 15 *Odes* and the *Carmen saeculare*.

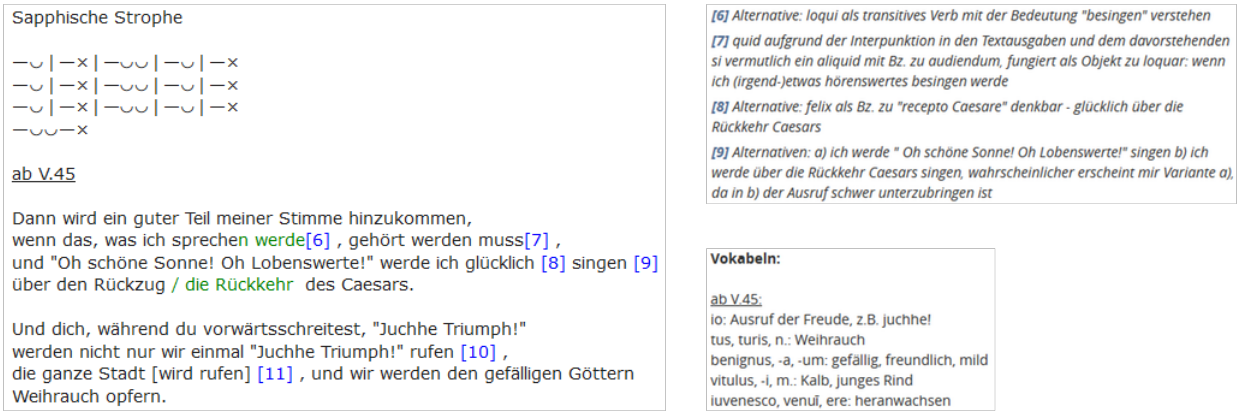


Fig. 4: Examples of the students work on Horace's *Odes*

The students compared notes on syntactic difficulties and thematic vocabulary with the *Perseus* tools, collected information on theme and metrics and situated the *Odes* in the context of other texts or the genre. The sustainability of the students' work was assured through further usability for the individual preparation of the reading list exam.

4. Observations on Working with Digital Methods in Classics

Classical textual study is about literary and non-literary Latin and ancient Greek texts. Applying digital methods to these texts requires to treat them in a most suitable way and therefore not only text quantity but also quality with regard to philological standards matters a lot to perform valid research – though, obtaining good textual data is extremely time-consuming even for small corpora. Of course, in-depth text critic and consideration of all textual variants is not necessary or even possible for every research question. However, the textual tradition has to be at least taken into account. Discussion of the status of different textual surrogates is necessary

[44] The seminar „Horace's *Odes*" also took place in summer term 2015 at the University of Constance.

and gets most virulent in respect of digital reference schemas for text variants. Furthermore, thinking carefully about the investigation setting and the methods of analysis is necessary and in the beginning a challenge: What investigates the application of the tool in fact and about what can valid and reliable statements consequently be made?

Of course, the issue of accessibility and preservation exists for printed and for digital text versions – though in different dimensions. Therefore, risk diversification over several media is probably a good solution to ensure the survival of ancient works, so that hunting for disappeared texts is hopefully not necessary but possible in the future. Besides, the enrichment in media as well as in analysis methods can effect an expansion of the established canon of Latin literature.

Regarding the interpretation of results two characteristics of digitally achieved results are striking: Firstly, for untrained literary scholars the visualizations require expertise in interpreting, otherwise this can be tempting to confirm personally preferred hypotheses. Secondly, the fair amount of results is enormous challenging: As digital analysis of, for instance, text reuse is done by automatic information retrieval, the results are not prone to human errors. But this means on the other hand that computer-based applications produce a lot of results, whereby really meaningful results are always surrounded by less or even not meaningful results. Restricting the search algorithm for more accuracy would always go in hand with loosing potentially relevant results. Therefore, a really thrilling task for the researcher is to find the needle in a haystack, the so-called 'noise'. Checking thousands and thousands of results is a great challenge for digital intertextuality research. Of course, in such masses of data failure is more likely. But this is a point where the value and importance of conventional work of philological scholars becomes most obvious. As a result, supplementing the toolbox of a Latinist with digital methods can support to facilitate the evaluation of conventional methods and therefore encourages to benefit of the potentials of both, conventional and digital methods of text analysis.