Word Lists, Concordances, Text Comparison, and Morphology in Latin and Ancient Greek Texts

Robert Maier

Katharina Geisler-Str. 16, 85356 Freising, Germany
robert@maierphil.de

Abstract. The CD-ROM editions of Latin and Ancient Greek texts by Packard Humanities Institute (PHI) and Thesaurus Linguae Graecae (TLG) have been within the most appreciated digitalization projects in classical philology since the Nineties. The initial philosophy of these projects was to make the texts available in an open and well documented format and to leave it to interested people to create software in order to deal with these texts. While at the very beginning of the electronic evolution in classical philology the visual representation of texts and word searches within the entire literature were the main challenge now advanced text mining methods are becoming more and more important. Several software packages have been developed within the past 18 years which involved almost different methods and offered various research opportunities. Some of the most advanced methods like creation of concordances, morphologic and stylistic searches and text comparison shall be discussed in depth.

Keywords: concordance, text comparison, morphology, Latin, Ancient Greek, TLG, PHI

1 Introduction

Due to the huge digitalization projects of Packard Humanities Institute (PHI) [1] and Thesaurus Linguae Graecae (TLG) [2] that were started in the Eighties most classical Latin and Greek texts are available on CD-ROM. These projects were started in the Eighties and first published in the early Nineties and have to be considered the beginning of the digital age of classical philology. The PHI and TLG CD-ROMs offered the possibility to perform word searches through the whole ancient literature for the first time.

2 Historical Overview

We should have a short look at how philologists worked before getting into contact with electronic text editions. The central instances were lexica and encyclopedias. Most of these works have their roots in the 19th century or even earlier. In that time

one of the most famous projects in classical philology was initiated, the Thesaurus Linguae Latinae (ThLL)[1]. The work on this lexicon was started in 1894 and it was planned to be a concordance of the entire classical Latin literature from the beginning to Isidor of Sevilla (6th century). Since computers and similar instruments did not exist the methods were completely different to how a modern electronical lexicon would be prepared today. The Thesaurus Linguae Latinae was – and still is – a printed lexicon. A very important difference between printed and electronic texts is that it is nearly impossible to change or insert text parts in printed material without creating a new edition. This essential problem was responsible for working on the Thesaurus Linguae Latinae letter by letter. Therefore text portions were excerpted and record cards created for every word that occurred within the texts containing the word and its context. These record cards were then sorted alphabetically by hand. The biggest handicap was that the whole text corpus had to be determined before the work on the Thesaurus could begin because it would have been very difficult or impossible to insert new texts or new editions.

3 PHI and TLG: License Model, Text Format and Software Development

Initially the PHI and TLG license model was a subscription system. The user could hire the CD-ROM for several years and pay for this period of time. Software developers could get the CD-ROMs for free. Neither PHI nor TLG delivered software to work with the CD-ROM editions. In the past few years the license models have changed significantly. Nowadays TLG offers their text corpus online with basic functionalities for text visualization and research while PHI is distributing the CD-ROM edition for free.

The original text format used for the CD-ROM editions was open and documented to give interested programmers the chance to develop their own software able to deal with the texts. This lead to a series of very different software projects with different license models offering different function sets to work with the texts.

PHI and TLG used a very compact text format. The text portions are stored in the so called Beta Code, a pure 7 bit ASCII format, where Greek texts were transcribed by Latin characters. The so called level information that contains data on author, work and work structure is stored within the texts using the ASCII codes from 0x80 to 0xFF. Therefore level information is separated very well from text portions. In most cases it follows the common division types, e.g. “book”, “chapter”, “line”, and is therefore structured in a hierarchical way. Exceptions of the normal behavior can be found especially in the PHI inscription collection but there are also some slight format differences between PHI and TLG texts.

The two principal CD-ROM editions are very different in length. While the PHI CD-ROM of classical Latin texts amounts to about 7,360,000 words (350,000

[1] Thesaurus Linguae Latinae at Bayerische Akademie der Wissenschaften, http://www.thesaurus.badw-muenchen.de/.

different forms) the TLG CD-ROM contains nearly 10 times more text, i.e. about 72,000,000 words (1,100,000 different forms)[2].

The text files are organized in blocks of 8kB each in order to accelerate access which was important when working on a standard PC in the early Nineties. References to each text block could be retrieved from an index file that accompanied each text file. Additional information on the authors, editions, text classification etc. is stored separately in special data base files and can be used independently.

Beta Code is a simple but efficient code to transcribe ancient Greek texts using Latin characters. Capitals are preceded by an asterisk [Γ→*G, γ→G], accents are represented by slashes and the equal sign, [ά→A/, ὰ→A\, ᾶ→A=], spiritus signs by parentheses [ἀ→A), ἁ→A(], the iota subscriptum by the pipe sign [ᾳ→a|]. Other characters and format information which cannot be represented by the basic encoding are inserted as a signifier that indicates the type of the character or format information (&, $, # …) followed by a number (e.g. Ϡ [Greek letter Sampi]→#5). Uncertain or corrupted text passages are denoted by special punctuation signs. Those passages are very difficult to handle in search and concordance routines. An important practical aspect of Beta Code is that it needs significantly less space than uncompressed Unicode[3] and that it is still readable.

Table 1. Example for the beginning of a text file (Plutarch, 1st work: Theseus). Level information in grey characters, format information in normal characters, text portions are highlighted.

Byte no.	Hexadecimal view	ASCII view
0x0000	EF 80 B0 B0 B0 B7 FF EF	ï€°°°·ÿï
0x0008	81 B0 B0 B1 FF EF 82 D4	°°±ÿï‚Ô
0x0010	E8 E5 F3 FF AF F4 FF 40	èåóÿ¯ôÿ@
0x0018	40 40 7B 31 24 32 30 2A	@@{1$20*
0x0020	51 2A 48 2A 53 2A 45 2A	Q*H*S*E*
0x0028	55 2A 53 20 2A 4B 2A 41	U*S *K*A
0x0030	2A 49 20 2A 52 2A 57 2A	*I *R*W*
0x0038	4D 2A 55 2A 4C 2A 4F 2A	M*U*L*O*
0x0040	53 24 7D 31 20 A1 40 2A	S$}1 ¡@*
0x0048	28 2F 57 53 50 45 52 20	(/WSPER
0x0050	45 29 4E 20 54 41 49 3D	E)N TAI=
0x0058	53 20 47 45 57 47 52 41	S GEWGRA
0x0060	46 49 2F 41 49 53 2C 20	FI/AIS,
0x0068	57 29 3D 20 2A 53 4F 2F	W)= *SO/
0x0070	53 53 49 45 20 2A 53 45	SSIE *SE
0x0078	4E 45 4B 49 2F 57 4E 2C	NEKI/WN,

[2] These data have been retrieved with the concordance tools of LECTOR 2007.

[3] The uncompressed Beta Code version of the Genesis occupies about 234kB, the uncompressed Unicode version about 442kB. The ODT version needs 163kB.

White characters on dark background represent the level information, black letters on grey the text portions and black characters on white are additional format signs. This short section represents the following text portion:

0007, 001, Thes, t

ΘΗΣΕΥΣ ΚΑΙ ΡΩΜΥΛΟΣ

0007, 001, Thes, 1, 1, 1

Ὥσπερ ἐν ταῖς γεωγραφίαις, ὦ Σόσσιε Σενεκίων,

Before the transition to Unicode, Greek texts were displayed using different ancient Greek TrueType fonts which used an 8 Bit ASCII representation of Greek characters. The consequences were lack of compatibility, problems with alphabetically sorted lists and with retrieval functions. When most of the recent development projects started, Unicode fonts were not yet available. It was still the time of Apples System 7 and Microsoft's DOS and Win3.x. At the beginning Macintosh PCs were the only one able to display Greek without major problems. For this reason Pandora for Macintosh was one of the most used programs to work with the PHI and TLG texts in that times.

The software development for the PHI and TLG text editions was divided mainly into the following regions of interest:

1. Implementation of standard text processing features:
   1.1. Text visualization
   1.2. Text export
2. General needs within classical philology:
   2.1. Retrieval functionalities
   2.2. Concordance tools
   2.3. Morphology, dictionaries, and translation tools
3. Special research requests:
   3.1. Text to Speech (TTS) and Braille export for visually impaired persons
   3.2. Statistical tools

Due to different development goals the single software projects lead to different approaches and function sets. However, the basic features of most software packages are Greek and Latin text display, export capabilities, and retrieval functions, while there are significant differences in more advanced research functionalities. The possibility to perform cooccurrence analyses - which is a standard text mining method[4] - has been implemented only in LECTOR 2007 [5]. The development of the LECTOR text analysis software was started by the author in 1991. The first focus was to create an adequate text visualization of both Latin and Ancient Greek texts. A second important requirement was the capability to perform searches of words or word fragments within all Latin and Greek texts stored on the PHI and TLG CD-

[4] See e.g. C.Belica et al.: „Methoden der Korpusanalyse und –erschließung", http://www.ids-mannheim.de/kl/projekte/methoden/

ROMs. In the following years several other text analysis features were implemented. The current LECTOR version runs under MS Windows and contains advanced analysis features like creating concordances of a set of texts sorted by word beginning or word end and text comparison with user defined parameters.

Table 2. Feature list of current standard software packages to work with the TLG and PHI CD-ROM editions (X=full support, O=support to some extent, -=not supported)

	Musaios [9]	SNS Greek & Latin [7]	Diogenes [6]	Workplace Pack [8]	LECTOR 2007 [5]
Greek text display	X	X	X	X	X
Text export	X	X	X	X	X
Word Searches	X	X	X	X	X
Concordances	O	O	-	O	X
Morphology	-	-	X[1]	O[1]	O
Dictionary	-	-	X[1]	O[1]	O
Text comparison	-	-	-	-	X
Statistics	-	-	-	-	O
Text to Speech (TTS)	-	-	-	-	X

[1]The morphology and dictionary tools used in these projects are provided by the Perseus project at Tufts University [3].

To understand needs and benefit of various research functionalities in classical philology it is necessary to know who are the users of electronic editions of Latin and ancient Greek texts. The author experienced that most users are members of at least one of the following groups with slightly different research interests:

Table 3. Main user groups versus principal research interests

	text	retrieval	concordances	morphology	translation	comparison
Students	X	X	O	X	X	-
Teachers	X	O	O	O	O	-
Scholars	O	X	X	X	O	X

The basic text processing tools are interesting for students and teachers while more advanced functionalities like concordances, comparison, and morphology are used primarily by scholars.

4 Beyond PHI and TLG: Other Digitalization Projects

The digitalization projects of PHI and TLG did not contain morphological data. Those data have been created for Latin and Ancient Greek within the Perseus Digital Library

project started in 1985 at Tufts University. The toolset for morphological analysis is based on two different models:

- The Perseus Digital Library project:
  - Morpheus:
    The first model was implemented in the morphological analysis software "morpheus" and is based on word lists excerpted from a dictionary and inflection schemes which are used to generate Latin and Greek forms in order to analyze form lists.
  - Hopper:
    The second model is based on a virtually complete list of all Latin resp. ancient Greek forms against which input forms can be checked. The completeness and correctness of the original text corpus where the form list has been deduced from is extremely important to generate reliable results by this method.
    These morphological data are available online or for download from Perseus and can be used for text analysis and translation tools. The Perseus project offers also access to electronic versions of the Liddell-Scott Greek Lexicon and of the Lewis And Short's Latin-English Lexicon.
- The CAMENA project:
  - CAMENA offers online access to a huge collection of Latin texts from the early modern period [4]. Within this project a significant quantity of older lexical and encyclopedical works is available.

5 State of the Art

A short overview of the capabilities of some current software packages to work with the PHI and TLG CD-ROM editions has already been given. The most advanced research functionalities offered by current software tools shall be discussed in the following.

The standard requirements for research software to be used with the PHI and TLG text editions are:

- Correct display of Latin and Greek texts including diacritical signs and special characters: Within the PHI and TLG texts a series of special characters have been used which are currently not represented in Unicode.
- Text export capabilities compatible to standard text processors: Most software packages for PHI and TLG are designed to work with Microsoft Windows and support the two standard ways of text exchange between different applications, i.e. copying text portions to the clipboard and saving texts into a common format like Rich Text Format (RTF).
- Word searches: Standard word search functions offer the possibility to search up to n forms with wildcards and Boolean operators simultaneously.
- Concordances: Most software packages offer the possibility to create or use concordances for an entire author.

Some functions already implemented in advanced software packages like Diogenes and LECTOR 2007 will become more important in future developments:

- Advanced concordance tools:
  - Combined concordances: LECTOR 2007 enables the user to create combined concordances of a group of authors. It is therefore possible to create a complete concordance from all texts of the PHI or TLG CD-ROM. This is necessary to simplify the comparison of lexical differences between two or more authors.
  - Special sort options: LECTOR 2007 offers the possibility to create concordances sorted by the word end. These concordances are helpful to study the use of verb forms and similar morphological phenomena.
  - Word frequency lists: LECTOR 2007 is able to create word frequency lists for virtually every author or group of authors.
- Text comparison:
  - LECTOR 2007 offers the possibility to compare texts in order to detect cooccurrences. This cooccurrence analysis is based on concordances and uses a set of variable parameters to define the degree of correlation of text passages: the context length, the number of words and the minimum length of identical word beginnings.
- Morphology and translation tools:
  - Diogenes includes Perseus morphological data and uses the electronic versions of Liddell-Scott Greek Lexicon and Lewis and Short's Latin-English Lexicon for word translation.
- Stylistic searches:
  - Some experiments have already been undertaken to detect stylistic devices in Latin and Greek texts. The DOS version of LECTOR was able to find alliterations and sound accumulations.

The following examples show some aspects of the concordance and text comparison features of LECTOR 2007. To create concordances LECTOR generates reference tables for the text content of the PHI or TLG data ignoring accents and diacritical signs. In the first example the relation of the most frequent word forms in the writings of C. Iulius Caesar (De Bello Gallico, De Bello Civili and minor works) and of Aulus Hirtius (De Bello Gallico, book 8) is shown.

Table 4. The most frequent word forms in the concordances of Iulius Caesar and Aulus Hirtius.

rank	Caesar	Hirtius	rank	Caesar	Hirtius
1	in (2.25%)	in (2.52%)	9	quod (0.79%)	ut (0.63%)
2	et (2.06%)	cum (2.01%)	10	ab (0.68%)	quae (0.61%)
3	ad (1.59%)	ad (1.27%)	11	qui (0.68%)	se (0.58%)
4	cum (1.08%)	et (1.20%)	12	non (0.64%)	atque (0.58%)
5	ex (1.05%)	qui (0.70%)	13	a (0.53%)	quam (0.55%)
6	atque (1.01%)	non (0.69%)	14	neque (0.52%)	esse (0.54%)
7	ut (0.91%)	ex (0.67%)	15	caesar (0.49%)	caesar (0.52%)
8	se (0.86%)	quod (0.66%)	16	quae (0.48%)	aut (0.46%)

One should expect the vocabulary of Aulus Hirtius to be close to that of Caesar since both were writing on quite similar topics. As we can see from this table the nominative "Caesar" occupies exactly the same extraordinary 15th position in both authors, while the order and content of the other entries is similar but not identical.

In the next example the correlation of word length and word frequency is illustrated for prose and poetry in both languages, Latin and Ancient Greek.

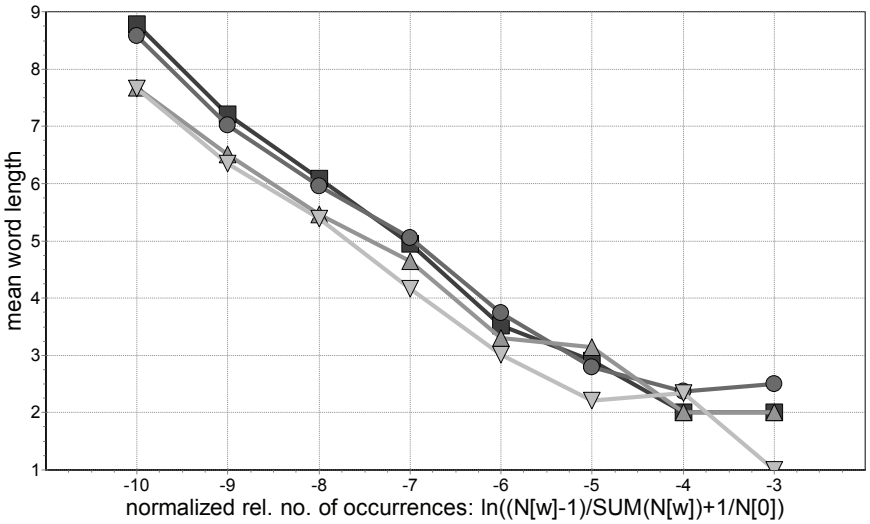

Fig. 1. Relation of mean word length and normalized relative number of occurrences in Tacitus (■, Latin prose), Herodot (●, Greek prose), Vergil (▲, Latin poetry), and Homer (▼, Greek poetry). N[w] is the number of words of a given length, SUM(N[w]) is the total number of words, N[0] is the number of words to which the distribution is normalized.

On the left side the less frequent forms are positioned while the most frequent words (like et or και) are positioned on the right. The word counts have been normalized in a way that a hapax legomenon will appear on position -10, independently from the total number of words. The distribution itself resembles Zipf's law [11]. We see from this chart that forms tend to be shorter in poetry than in prose in both languages, Latin and Greek.

The last example shows some results of a cooccurrence analysis performed with LECTOR 2007 again on the works of C. Iulius Caesar and Aulus Hirtius. To perform cooccurrence analyses LECTOR compares two concordances using three user defined parameters: (1) the context length, (2) the number of identical word beginnings within the context, and (3) the minimum length of those word beginnings. This type of cooccurrence analysis is quite flexible and able to generate results with different levels of correlation: from similar expressions to exact citations. Up to now the parameter setup and the evaluation of the results needs to be done by the user.

Table 5. Some results of a cooccurrence analysis of Caesar and Hirtius

Caesar	Hirtius
Caes, Gal, 2, 5, 6, 4: castra in altitudinem pedum XII vallo fossaque duodeviginti pedum muniri iubet.	Hirt, Gal, 8, 9, 3, 2: haec imperat vallo pedum duodecim muniri, loriculam pro hac ratione eius altitudinis inaedificari, ...
Caes, Gal, 3, 17, 4, 3: magnaque praeterea multitudo undique ex Gallia perditorum hominum latronumque convenerat, ...	Hirt, Gal, 8, 30, 1, 3: ..., qui, ut primum defecerat Gallia, collectis undique perditis hominibus, servis ad libertatem vocatis, ...
Caes, Civ, 1, 3, 4, 2: omnes amici consulum, necessarii Pompei atque ii, qui veteres inimicitias cum Caesare gerebant, in senatum coguntur.	Hirt, Gal, 8, 52, 5, 3: quod ne fieret, consules amicique Pompei evicerunt atque ita rem morando discusserunt.

In this example the context length was 5 words and at least 4 words had to start with at least 5 identical characters. It is evident that these examples are not real citations but just similar expressions. The method used in LECTOR delivers results independently from the word order. Therefore it is possible to detect not only direct citations but also juridical and military terms that are used by different authors or frequently by the same author.

6 Future Developments and Requirements

Up to now the text collections of PHI and TLG were used independently for reading the texts themselves, for searching word forms, for working with concordances, and for basic text comparison. In future it will be desirable and necessary to be able to perform more complex researches on the text corpora. In the field of classical philology there are still several needs for additional functionalities that can be offered best by using text mining methods:

- Integration of different text corpora: In future it will be important to be able to perform searches through different text corpora. The approach of the eAQUA project is to integrate those text corpora by means of text mining [10].
- Named Entity Recognition (NER): Significant effort has already been made within the CAMENA project for Latin texts. There is still no support within the current software for PHI and TLG texts.
- Cross language text comparison: The detection of translations and cooccurrences within texts written in different languages is a challenging task. A possible approach is to use synonym lists and translations to create an interface between Latin and Greek.

- Advanced statistical methods: There is still no support for extended statistical analysis within current software projects for the PHI and TLG texts. Using text mining methods would lead to a noticeable progress in this area.
- Morphology: Morphological tools are important to analyze word forms and their function and to compensate orthographical variance.
- Lexica and translation tools: It is desirable to connect the original texts with lexica and translation tools to enable the user to look up words from the original context and to add entries to a user generated lexicon.
- Advanced stylistic searches: Retrieval of metrical and stylistic elements is important for poetry analysis.
- Text to Speech: A basic Text to Speech (TTS) interface has already been developed for LECTOR 2007. A standard TTS interface is still a desirable feature for online resources.

7 Conclusion

Several software projects are already dealing with the Latin and ancient Greek text editions on CD-ROM provided by Packard Humanities Institute (PHI) and Thesaurus Linguae Graecae (TLG) and partly include basic text mining concepts. The next step of working with these text editions should be to use advanced text mining tools and methods in order to perform extended cooccurence and statistical analyses and to integrate enhanced morphological, lexical, and stylistic tools.

8 Acknowledgment

The author would like to thank R. Gruhl and M. Büchler for helpful discussions.

References

1. Packard Humanities Instiutute, http://www.packhum.org/phi/
2. Thesaurus Linguae Graecae, http://www.tlg.uci.edu/
3. The Perseus Digital Library, http://www.perseus.tufts.edu/
4. CAMENA, http://www.uni-mannheim.de/mateo/camenahtdocs/camena.html
5. LECTOR 2007, http://www.maierphil.de/lector/
6. Diogenes, http://www.dur.ac.uk/p.j.heslin/Software/Diogenes/
7. SNS-Greek & Latin, http://snsgreek.sns.it/sns_05.html
8. Workplace Pack, http://www.silvermnt.com/wpinfo.htm
9. Musaios, http://www.musaios.com/
10. eAQUA project, http://www.eaqua.net/
11. Li, Wentian: Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution. IEEE Transactions on Information Theory, Vol. 38, No. 6, 1842—1845 (1992)