A Corpus-based Approach to Philological Issues

Federico Boschetti

Advisor: Marco Baroni

Center for Mind / Brain Sciences

University of Trento, Italy

A thesis submitted for the degree of

PhilosophiæDoctor (PhD)

2009, December

1. Reviewer: Andrea Bozzi

2. Reviewer: Alessandro Lenci

3. Reviewer: Isabella Poggi

Day of the defense: January 18, 2010

Abstract

The aim of this work is the application of techniques developed in the domain of corpus linguistics to a collection of ancient Greek texts, taking into account not only the canonical text established by modern editors, but also the variant readings recorded in the critical apparatus or in the repertories of conjectures. The dissertation is divided in three parts consistently connected: construction, mapping and analysis of the corpus.

The first part is devoted to the corpus construction and it is focused on the techniques to improve the OCR accuracy on classical critical editions. This task is challenging because critical editions are multilingual, the set of characters to recognize is wide and the quality of last centuries paper is variable. Three OCR engines are applied to the same texts and a bayesian classifier, joint to a specific spell-checker, evaluates the most probable output. It is demonstrated that the improvement is significative.

The second part is devoted to the alignment of the contents extracted from critical apparatus and repertories of conjectures to the reference text. A parser has been developed to classify the chunks of information (verse number, Greek word sequences, textual operation, scholar that suggested the conjecture). Align algorithms used to find the precise position of the conjecture in its context are illustrated in detail.

The third part is devoted to the study of the semantic spaces of ancient Greek. The chapter is focused on the specificity of the corpus, that is morphologically complex, literary (both poetical and prosastic) and diachronical (from VIII century B.C. to XV century A.D.). The

word senses in documents belonging to different genres are explored, and the diachronical change of meaning is observed.

Finally, a group of meaningful conjectures extracted in the first part is analysed, evaluating the most interesting reciprocal relations in the semantic space.

Contents

1

Introduction

The aim of this work is the application of corpus-based techniques to the domain of classical philology, contributing to the development of computational philology. This work addresses three topics strictly chained: digitization of critical editions and secondary sources; parsing of critical apparatus and repertories of conjectures and, finally, exploration of semantic spaces of ancient Greek, as a support to the evaluation of variants and conjectures.

1.1 Specificity of computational philology

Philology is both the floor and the ceiling of classical and modern literary studies. On one hand, it addresses the reliability of the actual text, the object for all the further criticisms and literary studies. In the case of ancient works, philologists evaluate textual variants attested in manuscripts or conjectures suggested by scholars (*constitutio textus*). In the case of modern works, scholars observe the development of printed editions reconstructing the author's changes on the autograph, the editor's interventions, the improvements or regressions from former to latter editions (*genetic philology*). On the other hand, philology addresses the horizons of sense, the well-founded criteria, either formal or historical, contextual or pragmatic, that suggest if an interpretation is plausible, merely possible or highly rejectable.

In the early stage of computer-assisted literary studies, computational philology was often wrongly confused with computational linguistics (see Degani, 1992,

for a criticism of this approach), due to the application of the same methods and techniques, without specific adjustments. From the perspective of computational linguistics, texts are serial sequences of textual units, whereas from the perspective of computational philology, texts (with variant readings) are parallel sequences of textual units that insist on the same textual positions. Overlooking specific needs of computational philology, the creation of indexes and concordances was based on singular editions without critical apparatus. If there is no way to distinguish between words attested in the manuscripts and conjectural emendations made by latter scholars, it is impossible to study linguistic phenomena that could be shadowed by the normalizations due to wrong beliefs of modern scholars. A basic philological problem is that the attribution of a word, a collocation, a concept or an idea to an author is both a starting and an ending point.

Nowadays, computational philology is defining its identity as a necessary bridge between computer science, filtered by computational linguistics, and traditional philology. Conferences devoted to "digital philology", "e-philology" and "computational philology" are populating the scene of the last decade, as, among the others, Ciula and Stella (2007), Zurli and Mastandrea (2009), Bozzi (2004b) or Boschetti (2009) demonstrate. Even if in the scientific literature the terms are apparently used like synonyms, three areas can be delimited. "Digital philology" concerns the construction of digital libraries of philological works. "E-philology" involves the creation of the cyberinfrastructure (see Crane *et al.*, 2009) that allows the interoperability and promotes the communication among scholars. "Computational philology" pertains to the development of procedures to parse, process and analyze texts contained in digital corpora, and the current work is focused on these aspects.

Bozzi (2004a) points out one of the most important differences between old and new philology: the transition from the "hand-crafted" procedures of traditional philology to the "industrial" processes promoted by computational philology, where texts are the raw material, elaborated by teams of specialized operators, the scholars, in order to create the final products: dictionaries, indexes, concordances, etc. Every industrial restyling requires a trade-off between a loss of precision and attention for the details and a gain of scalability and objectivity.

Attention shifts from the final product (edition as a static book) to the process (edition as a dynamic aggregate of features that can be improved), which provides an open and flexible work.

1.2 Digital scribes and computational scholars

Digital Libraries can grow along two different dimensions: breadth and depth. In the first case, works of many authors extend the existing collections. In the second case, different editions of the same works and related studies populate a monothematic region of the library.

The most complete Greek and Latin corpora of texts, such as the Thesaurus Linguae Graecae (TLG) and the Packard Humanities Institute (PHI) Latin collection, are based on authoritative, most recent critical editions of each classical author. In these collections, only the text established by the editor is digitized, whereas the critical apparatus is omitted. Such approach to the ancient text, just about acceptable for literary and linguistic purposes, is unfeasable for philological studies. In fact, the philologist needs to identify manuscript variants and scholars' conjectures, in order to evaluate which is the most probable textual reading, accepting or rejecting the hypotheses of the previous editors. Furthermore, he or she needs to examine the commentaries, articles and monographs concerning specific parts of the text. Thus, the extension in breadth of the afore-mentioned collections needs to be integrated by the extension in depth, according to the paradigms of a new generation of digital libraries (see Crane *et al.*, 2006; Stewart *et al.*, 2007).

In order to go in depth, philological studies are necessarily focused on single authors, genres or periods, even if they need to find links and parallels in the entire Greek and Latin literature. For this reason, teams of specialists need to share a common infrastructure, as pointed out by Crane *et al.* (2009).

For instance, the Perseus Project[1] is building a cyberinfrastructure to interrelate different philological and archeological projects. The Musisque Deoque

[1] http://www.perseus.tufts.edu

Project[1], on the other hand, has created a large platform to manage textual variants of Latin texts. The Multitext Homer Project[2], even if it is focused on a single ancient author, has developed a suite of services that can be easily extended to other authors.

Alluding to Reynolds and Wilson (1991), we can divide the studies in two parts, postulating that textual transmission and textual criticism, even if empowered by computational methods and tools, do not break the secular tradition of classical philology.

1.2.1 Digital scribes

The last decades are protagonist of an epochal mutation in the transmission of texts: from the papyrus to the volumen, from the manuscript to the printed editions, from the book to the magnetic support, changing preservation techniques also implies rethinking textual structures.

Digitization of classics is following a sounding list of priorities: first step was the completion of corpora in Greek, Latin and other ancient languages, such as Persian and Coptic. Texts were extracted from canonical editions, dismissing the critical apparatus and all the paratextual information, such as preface, introduction and indexes. One of the most valuable products of this phase is the *Thesaurus Linguae Graecae*, manually digitized by a team of operators, without the necessity of specific skills in classical philology.

The second step is the digitization of variants. This phase is the most interesting one from a theoretical point of view, because it involves both the competence of the philologist to identify selectional criteria and the competence of the computer scientist to create suitable tools to manage variants. Bozzi (2002) illustrates the features of the philological workstation developed at the CNR of Pisa.

A critical apparatus, either printed or digital, usually is a selection of all the existing variants and conjectures. Textual variants are limited by the manuscripts checked and conjectures are limited by the relevance that editors accord to the

[1] http://www.mqdq.it

[2] http://chs.harvard.edu/chs/homer_multitext

suggested emendations. Mastandrea (2009) illustrates the *Musisque Deoque* project, aimed to provide a minimal apparatus to a large amount of texts and Mondin (2009), collaborator of the project, points out how critical apparatus can establish different selectional criteria according to different aims: for the study of the language, orthographical and morphological variations cannot be missed, but for the study of intertextuality, lexical and semantic variants must have the priority in a minimal apparatus.

The third step of digitization follows two lines: on one hand the digitization of printed critical editions and on the other hand the digitization of the secondary sources, such as commentaries, articles, dictionaries, encyclopediae. This phase scales up the quantity of texts that will be available to the digital philologist. Integration of OCR techniques, information extraction, textual mapping and linking are necessary to deal with the mass of information.

1.2.2 Computational scholars

Computational scholars are philologists skilled in both classical philology and computer science.

Computational philology deals with diachronical corpora. According to Hilpert and Gries (2009), which provides the most updated references on the topic, "the use of corpora that are divided into temporally ordered stages, so-called *diachronic corpora*, is becoming increasingly wide-spread in historical corpus linguistics, creating a natural bridge between corpus linguistics and computational philology".

A corpus-based approach has been applied by O'Donnell (2005), who focuses his attention on the Greek of the New Testament. In particular, the author tackles the topic of textual variants and corpus-based criteria to select them. A large part of his work is devoted to lexicographical analyses of the New Testament. The study of collocations shows, for example, how ἐγείρω (to wake up) and ἀνίστημι (to rise) are synonyms in the specific context of the Christian resurrection.

1.3 Overview

The next chapters illustrate the different aspects discussed above.

Chapter 2 is devoted to the improvement of OCR performances.

OCR can be applied to XIX and XX century critical editions, reaching up to 99% of accuracy on the text and more than 90% of accuracy on the critical apparatus. Somehow, it is important to point out that the critical apparatus is in average less than 10% of the page.

These performances are obtained by the alignment and merging of three different OCR outputs and the application of an automated system of spell-checking, supported by the evidence of the OCR outputs. After suitable training, three OCR engines are able to deal with polytonic Greek. Each OCR engine is more or less reliable for specific characters, but the merging system developed in this work computes the most probable character in each position and the result significatively overwhelms the performances of the single engines.

Chapter 3 is devoted to the alignment of variants and conjectures on the text of the reference edition.

Repertories of conjectures register not only the corrections to the ancient text suggested by the editors in their own editions, but also the proposals for emendation contained in commentaries and articles. The repertories of conjectures have a trivial structure: in fact, more than 90% of the items are constitued by the reference to the verse affected, the text of the conjecture and the name of the scholar that has made the proposal. A parser identifies these chunks of information and an aligment algorithm is applied to find the exact position in the verse where the conjectures are intended to be collocated.

Chapter 4 is devoted to the exploration of ancient Greek semantic spaces.

Peculiarities of the ancient Greek corpus are illustrated, such as the diachronical stratification and the decomposition in subcorpora per genres. Semantic relations are explored, in particular antonymy, hypernymy and meronymy with effective exemplifications. The changes of meaning due to scientific, philosophical and religious mutations are discussed. Finally, a concrete and an abstract term are investigated in different semantic spaces, generated by the epic, tragic and

philosophical subcorpora, in order to show how the well known characteristics of these genres emerge in the semantic associations generated by the seed words.

In conclusion, chapter 5 is devoted to exemplify how digital philology can support classical philology and it summarizes the results achieved.

2

Improving OCR Accuracy

This chapter[1] describes a work-flow designed to populate a digital library of ancient Greek critical editions with highly accurate OCR scanned text. While the most recently available OCR engines are now, after suitable training, capable of dealing with the polytonic Greek fonts used in 19th and 20th century editions, further improvements can also be achieved with postprocessing. In particular, the progressive multiple alignment method applied to different OCR outputs based on the same images is discussed in this chapter.

After the introduction and the exposition of related works (sections 2.1 and 2.2, section 2.3 illustrates the methodology to improve OCR performances, whose results are discussed in section 2.4. Section 2.5 illustrates how to model the manual corrections and section 2.6 discusses the remapping of text on images. Finally, section 2.7 summarizes the main results.

2.1 Introduction

The new generation of Greek and Latin corpora that has become increasingly available has shifted the focus from creating accurate digital texts to sophisticated digital editions. Previously prefaces, introductions, indexes, bibliographies, notes, critical apparatus (usually at the end of the page, in footnote size), and

[1]This work was supported by a grant from the Mellon Foundation. I would like to express my gratitude to G. Crane and all the Perseus Project's staff. See Boschetti *et al.* (2009).

textual variations of different editions have either been discarded or systematically ignored in the creation of early digital collections. The ancient text that we read in modern editions, however, is the product of editors' choices, where editors have evaluated the most probable variants attested in the manuscripts or the best conjectures provided by previous scholars. Humanists thus need both textual and paratextual information when they deal with ancient works.

Critical editions of classics are challenging for OCR systems in many ways. First, the layout is divided into several text flows with different font sizes: the author's text established by the editor, the critical apparatus where manuscript variants and scholars' conjectures are registered and, optionally, boxes for notes or side by side pages for the parallel translation. Second, ancient Greek utilizes a wide set of characters to represent the combinations of accents and breathing marks on the vowels, which are error prone for OCR systems. Third, critical editions are typically multilingual, because the critical apparatus is usually in Latin, names of cited scholars are spelled in English, German, French, Italian or other modern languages, and the prefaces, introductions, translations and indexes are also often in Latin or in modern languages. Finally, 19th century and early 20th century editions can have many damaged text pages that present great difficulties for conventional OCR.

2.2 Related Work

We can divide works related to the digitization of ancient texts into three groups: the first one concerns the analysis of manuscripts and early printed editions, the second group concerns the structure of digital critical editions (i.e. editions that register variants and conjectures to the established text) and the third group concerns OCR work performed on printed critical editions from the last two centuries.

The general approach for the first group is to provide methods and tools for computer assisted analysis and correction. Moalla *et al.* (2006) developed a method to classify medieval manuscripts by different scripts in order to assist paleographers. Jlaiel *et al.* (2007) suggested a strategy to discriminate Arabic and Latin modern scripts that can be applied also to ancient scripts. Leydier *et al.*

(2007), Leydier *et al.* (2005) and Bourgeois and Emptoz (2007) used a method of word-spotting to retrieve similar images related to hand written words contained in manuscripts. Edwards *et al.* (2004), on the other hand, developed a method based on a generalized Hidden Markov Model that improved accuracy on Latin manuscripts up to 75%.

The second group of studies explored recording variants and conjectures of modern authors, for instance Cervantes, such as Monroy *et al.* (2007) or of ancient texts, for instance in Sanskrit, such as Csernel and Patte (2007) or in Latin, such as Bozzi (2002) and Mastandrea (2009).

The third group of studies concerned improvements of OCR accuracy through post-processing techniques on the output of a single or multiple OCR engines. Ringlstetter *et al.* (2005) suggested a method to discriminate character confusions in multilingual texts. Cecotti and Belaïd (2005) and Lund and Ringger (2009) aligned multiple OCR outputs and illustrated strategies for selection. Namboodiri *et al.* (2007) and Zhuang and Zhu (2005) integrated multi-knowledge with the OCR output in post-processing, such as fixed poetical structures for Indian poetry or semantic lexicons for Chinese texts. Bozzi (2000) illustrates an integrated technique between OCR and spell-checking applied to damaged text documents.

Our work is focused to improve the achievements of the third group of studies. Limits established in the first group of studies are taken into account: an experiment on a Latin *incunabulum*, illustrated in 2.3.7, confirms the state of the art performances. The creation of digital critical editions, theme of the second group of studies, is the aim of our work, as illustrated in the next chapter.

This chapter further develops some guidelines first expressed in Stewart *et al.* (2007). In that previous research, the recognition of Greek accents in modern editions was not considered due to the technological limitations imposed by the OCR systems available.

2.3 Methodology

Our main interest in this research is to establish a work-flow for the massive digitization of Greek and Latin printed editions, with particular attention to the scalability of the process. The principal factors that determine the preparation of

different pre- and postprocessing procedures are book collection specificities and preservation status.

2.3.1 Texts

Our experiments have been performed on different typologies of samples, in order to combine the aforementioned factors. Three editions of Athenaeus' *Deipnosophistae* and one of Aeschylus' tragedies have been used, by randomly extracting five pages from each exemplar. All documents have been downloaded from Internet Archive[1]. Athenaeus' exemplars belong to different collections and they are distributed along two centuries: Meineke (1858) and Kaibel (1887) are in the Teubner classical collection, whereas Gulick (1951) is in the Loeb classical library. Teubner and Loeb editions sensibly differ for script fonts, so that two different training sets have been created. They differ also for content organization: Meineke has no critical apparatus, Kaibel has a rich apparatus and Gulick has a minimal critical apparatus, supplementary notes and an English translation side by side.

The posthumous edition of Aeschylus by Hermann (1852), published by Weidmann, has no critical apparatus and has a script very similar to the Teubner editions.

In this study, Greek text and critical apparatus have been separated manually, whereas English translation and notes have been disregarded. In a second stage of the work, simple heuristics will be applied to classify textual areas.

Finally, in order to evaluate if and how the system could be extended to very early printed editions, an experiment has been performed on the *incunabulum* of Augustinus' *De Civitate Dei*, Venetiis 1475. In this case, even if the quality of the image is good, the irregularity of the script and the use of ligatures and abbreviations is very challenging.

2.3.2 OCR engines suitable for Ancient Greek recognition

Three OCR engines have been employed: Ideatech Anagnostis 4.1, Abbyy FineReader 9.0 and OCRopus 0.3 in bundle with Tesseract 2.03.

[1] http://www.archive.org

Anagnostis[1] is the unique commercial OCR engine that is provided with built-in functionality for ancient Greek and it can also be trained with new fonts. Accents and breathing marks are processed separately from the character body, improving the precision of the recognition system. On the other hand, Anagnostis is not able to recognize sequences of polytonic Greek and Latin characters, such as are present in the critical apparatus. In this case, Latin characters are rendered with the Greek characters most similar in shape (for example, the Latin letter *v* is transformed into the Greek letter *ν*).

FineReader[2] is capable of complex layout analysis and multilingual recognition. Even if polytonic Greek is not implemented natively, it is possible to train FineReader with new scripts, associating the images of glyphs to their Unicode representations. For these reasons, FineReader is currently the most reliable engine to recognize texts where different character sets are mixed.

OCRopus[3] is an open source project hosted by Google Code, that can be used in bundle with Tesseract[4], illustrated by Smith (2007), which is one of the most accurate open source OCR engines currently available. OCRopus/Tesseract needs to be trained in order to recognize polytonic Greek (or other new scripts, except Latin scripts) and the recognition of mixed character sets is acceptable. The output format is plain text or xhtml enriched with a microformat to register positions of words (or optionally single characters) on the page image.

2.3.3 Training of single engines

The training process is divided into two phases. First, each OCR engine has been trained with five pages randomly selected from the editions used in the experiments, verifying that the training set had no overlappings with the test set. Anagnostis and FineReader have been trained with the same sets of pages, whereas OCRopus/Tesseract has been trained with a different set, in order to increase the possibility of capturing character samples ignored by the other engines. In fact, the major issue in training FineReader and OCRopus/Tesseract

[1] http://www.ideatech-online.com
[2] http://www.abbyy.com
[3] http://code.google.com/p/ocropus
[4] http://code.google.com/p/tesseract-ocr

with ancient Greek is caused by the high number of low frequency characters (as expected because of Zipf's law). Unicode represents polytonic Greek both by pre-combined characters and combining diacritics, but during the training process these engines seem to analyze glyphs only as whole characters, without separation between vowels and diacritics, as Anagnostis is able to do. The entire set of pre-combined characters for ancient Greek contains more than two hundred glyphs, but some of them are employed with a very low frequency. For example, in the Athenaeus' Kaibel edition, the letter ᾷ (alpha with circumflex accent, rough breathing mark and iota subscript) occurs only twice out of more than one million characters. Thus, the probability that these rare characters are sampled in the training sets is quite low. For the sake of scalability, training is based on collections and not on exemplars. For this reason, only one training set per engine has been created for the Teubner editions, mixing pages from both Kaibel's and Meineke's exemplars.

FineReader has a good built-in training set for modern (monotonic) Greek and it is possible to use the user defined training sets either alone or in bundle with the built-in trainings. Unfortunately, while this increases the accuracy for the recognition of non-accented characters it also decreases the accuracy for the recognition of vowels with accents and breathing marks. Thus, two training sets have been created for FineReader: with and without the addition of built-in training sets.

Second, the errors produced by each engine after the first stage have been compared with the ground truth, in order to calculate the error patterns that can be corrected by the cooperation of different OCR engines. The new training sets must be identical for all the engines. For Weidmann's edition, a new set of five pages, different from both the training set and the test set, has been extracted and the hand transcription has been used as ground truth. For the other editions, a k-fold cross validation method has been performed, using all the pages but the testing one for the training.

OCR output has been post-processed with a script that adjusts encoding and formatting errors, such as Latin characters inside Greek words with the same or very similar shape (e.g. Latin character *o* and Greek character ο, omicron), spaces followed by punctuation marks and other illegal sequences. A second script

adjusts a small set of very frequent errors by the application of regular expressions. For example, a space followed by an accented vowel and by a consonant, an illegal sequence in ancient Greek, is transformed into space, followed by a vowel with breathing mark and a consonant.

The adjusted OCR output has been aligned to the ground truth by a dynamic programming alignment algorithm, according to the methods explained in Feng and Manmathan (2006) and in van Beusekom *et al.* (2007). As usual, alignments are performed minimizing the costs to transform one string into the other, adding gap signs when it is necessary. In this way, n-gram alignments can be a couple of identical items (correct output), a couple of different items (error by substitution), an item aligned to a gap sign (error by insertion) or, finally, a gap sign aligned to an item (error by deletion). After the alignment, the average number of substitutions, insertions and deletions has been used to compute the average accuracy of each OCR engine. Navarro (2001) offers a survey on methods to calculate approximate string matchings.

Data concerning alignments of single characters, bigrams, trigrams and tetragrams are registered in the error pattern file. For the sake of efficiency, data related to correct alignments of n-grams are registered only if the n-gram occurs at least once in a misalignment. In fact, we are particularly interested in comparing the probability that one n-gram is wrong to the probabilty that it is correct, as we will see below. The error pattern file is a table with four columns: number of characters the n-gram is constituted by, n-gram in OCR output, aligned n-gram in ground truth and a probability value, illustrated by formula (2.1).

$$\frac{C(a \rightarrow b)}{C(b)} * \left(\frac{C(b)}{N}\right)^{1/3} \tag{2.1}$$

The first factor of this value expresses the probability that, given a character (or n-gram) a in the OCR output, it represents a character (or n-gram) b in the ground truth (a is equal to b, in case of correct recognition). It is represented by the number of occurrences of the current alignment, $C(a \rightarrow b)$, divided by the total number of occurrences of the b character (or n-gram) in the ground truth, $C(b)$. The second factor of this value is the cubic root of $C(b)$ divided by the total number of characters or n-grams, N. This factor is equal for every engine, because it is based only on ground truth. The cubic root of this value is provided, according to the formula (2.6), which will be explained below.

2.3.4 Tests and adjustments

Tests have been performed on each OCR engine and the output has been adjusted with the simple post-processing scripts used also for the training samples. First of all, the two FineReader outputs (with and without the built-in trainings) have been aligned with the same methodology explained below for the alignments among different engines and we have obtained a new, more accurate FineReader output to be aligned with the other engines.

2.3.5 Multiple Alignment and Naive Bayes Classifier

Alignment algorithms have their origin in the domain of bioinformatics to align DNA sequences, but they are commonly employed also in computational linguistics to evaluate the similarity of textual sequences (see, for instance, Kondrak, 2002). The principle is quite simple, and it will be explored more extensively in the next chapter. The basic idea is to assign a cost to each textual operation (substitution, insertion and deletion). For substitution, a similarity matrix determines lower costs for characters with more similar shapes. Insertions and deletions have a fixed cost, whereas the identity of characters has no cost. When two strings have to be aligned, a matrix is created that put the characters of the first string in row and the characters of the second string in column. The cumulative costs to transform one string in the other string are stored, moving from left to right and from top to bottom along the two strings, in each cell of the matrix. Backtracking the path from the last cell to the first cell, following the minimal cost path, it is possible to determine where gaps must be inserted, in order to align the two strings.

Outputs of the three engines have been aligned by a progressive multiple sequence alignment algorithm, as illustrated in Spencer and Howe (2003). The general principle of progressive alignment is that the most similar sequence pairs are aligned first, necessary gaps to align the sequences are fixed and supplementary gaps (with minimal costs) are progressively added to the previous aligned sequences, in order to perform the total alignment. In order to establish which pairs are more similar and then must be aligned first, a phylogenetic tree should be constructed, but for our triple alignment it is enough to rate each engine according to the average accuracy value established during the training process. In

our tests, FineReader has scored the highest, followed by OCRopus and Anagnostis. For this reason, FineReader and Anagnostis are aligned first. The resulting OCRopus string with gap signs is aligned to Anagnostis and the new gap signs are propagated to the previously aligned FineReader string. The triple alignment is shown in Figure 2.1 (further discussed below), where the gap sign is represented by underscore.

The alignment in itself is not enough to determine the most probable character: even if two engines are in agreement, but are poorly reliable for a specific character identification, the most probable character could be provided by the third engine in disagreement. Even if all the engines are in agreement, the most probable character could be another one, such as when three engines are only able to recognize Greek characters and the text is written in Latin. This situation, however, is not considered in the current study, which is limited to the selection among characters provided by at least one engine.

Formally, the probability that the current position in the original printed page e_0 contains the character x, given that the first engine e_1 provides the character c_1, the second engine e_2 provides the character c_2 and the third engine e_3 provides the character c_3, is expressed by the formula:

$$P(e_0 = x | e_1 = c_1, e_2 = c_2, e_3 = c_3) \tag{2.2}$$

where, in general, $P(E_0|E_1, E_2, E_3)$, denotes the posterior probability for the event E_0, given the conjunction of the events $E_1 \cap E_2 \cap E_3$.

For example, (2.2) expresses the probability that the character $\acute{\alpha}$ is in the current position on the printed page, knowing that the first engine has provided $\grave{\alpha}$, the second engine has provided $\ddot{\alpha}$ and the third engine has provided $\acute{\alpha}$. These probabilities are deduced by the error pattern data recorded during the training process.

To find the highest probability among the three items provided by the engines, we have implemented a naive Bayes classifier. In virtue of the Bayes' theorem, (2.2) equals:

$$[P(e_1 = c_1, e_2 = c_2, e_3 = c_3 | e_0 = x) * P(e_0 = x)]/P(e_1 = c_1, e_2 = c_2, e_3 = c_3) \tag{2.3}$$

Given that a naive Bayes classifier is based on the conditional independence assumption, the first factor in the numerator of (2.3) can be rewritten as

$$P(e_1 = c_1 | e_0 = x) * P(e_2 = c_2 | e_0 = x) * P(e_3 = c_3 | e_0 = x) \tag{2.4}$$

Considering that we are not interested in finding the value of the highest probability, but simply in finding the argument x_0 that provides the highest probability, we can omit the denominator of (2.3) and use the following formula:

$$x_0 = \arg\max_x P(e_1 = c_1|e_0 = x) * P(e_2 = c_2|e_0 = x) * P(e_3 = c_3|e_0 = x) * P(e_0 = x) \tag{2.5}$$

Generalizing, we can write the equation (2.5) as

$$x_0 = \arg\max_x \prod_{i=1}^{n} P(e_i = c_i|e_0 = x) * P(e_0 = x)^{1/n} \tag{2.6}$$

where n is the number of OCR engines, e_i is a specific engine, c_i is the character provided by that engine. This equation explains why we computed the cubic root of the ground truth character probability in the equation (2.1). For the sake of efficiency, in this way we do not need to search for this factor and multiply it for the other factors all the times that we compute the requested term.

In our implementation, a triple agreement is unprocessed and in case of probability equal to zero, the output of the first engine (FineReader, in this case) is selected. In Figure 2.1 the result of the selection performed by the system is shown. In blue and red are indicated the correct characters selected from OCRopus and Anagnostis, despite the character recognized by FineReader.

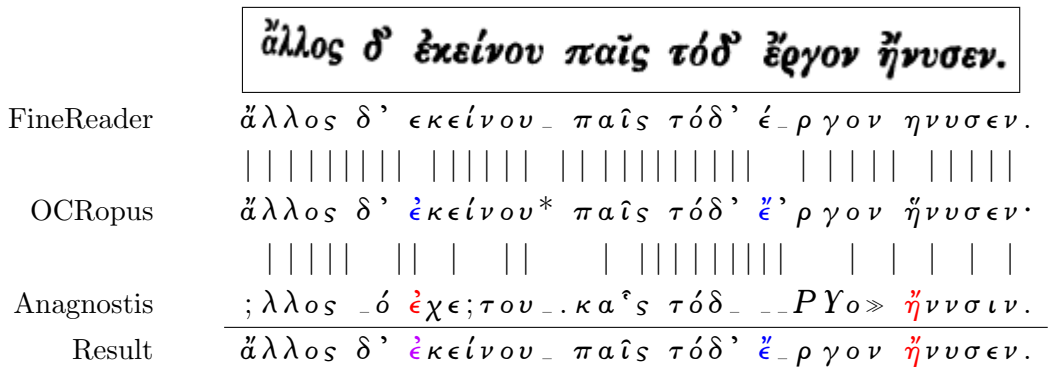

Figure 2.1: Multiple alignment of the three engines output

The high number of ancient Greek pre-combined characters reduces the probability that the training sets contain some error patterns present in the test sets. In this case, the probability for a correct item is zero, which should be avoided by Laplace (add-one) smoothing.

2.3.6 Spell-checking supported by multiple alignment evidence

OCR output can be corrected by spell-checkers but, as explained in Reynaert (2008b) and Stewart *et al.* (2007), the automatic spell-checking applied to misspelled words alone is often unreliable; the first suggestion provided by the spell-checker could be wrong or, as is often the case, the word list of the spell checker does not contain proper names and morphological variants, and it thus replaces a correct word with an error. In order to reduce these issues, we have adopted a spell-checking procedure supported by the engines output evidence, filtering only the spell-checker suggestions that match a regular expression based on the triple alignment.

In order to integrate the spell-checker in our system, we have used the Aspell API[1] and we have used the word list generated by Morpheus, the ancient Greek morphological analyzer (see Crane, 1991). The string generated by the naive Bayes classifier is analyzed by the spell-checker. When words are rejected by the spell-checker because they are not contained in the word list, a regular expression is generated from the aligned original outputs, according to these simple rules: a) characters in agreement are written just once; b) two or three characters in disagreement are written between brackets; c) gaps are tranformed into question marks (to indicate in the regular expression that the previous character or couple of characters between brackets are optional). For example, given the aligned outputs: a) ἥλασεν, b) ἤλαστν and c) ἤλασ_ν, the regular expression generated is /[ἥἤ]λασ[ετ]?ν/. All the suggestions provided by the spell-checker are matched with this regular expression, and only the first one that matches is selected, otherwise the mispelled word is left unchanged. Further examples are shown in Figure 2.2. The first example, ἐξερήμωσεν, and the last example, ευφρων, merit some further consideration. The first case illustrates a situation in which a correct morphological variant is not present in the spell-checker word list. No suggestion provided by the spell-checker matches the regular expression generated by aligned outputs, thus the word is correctly left unchanged. On the other hand, ευφρων is an incorrect ancient Greek word because it has neither accent nor breathing mark. In this case, none of the suggestions of the spell-checker are supported

[1] http://aspell.net

by the aligned outputs evidence, thus in this case the word is incorrectly left unchanged. While the first suggestion of the spell-checker is incorrect, the third one is correct.

FineReader output	RegEx matching all OCRs	Spell-checker suggestions	Result
ἐξερήμωσεν	ἐξερήἑ?[μι]ωσεν	ἐξερήμωσε, ἐξερήμωσέ, ἐξηρήμωσεν	ἐξερήμωσεν
ωπασεν	[ωοὦ]π[αο]σ[εό]ν	ὦπασεν, ὤπασέν, σπάσεν	ὦπασεν
εν’	[εἕ]ν’	ἐν, ἐν’ ... ἕν’ (34th item)	ἕν’
επάσης	ε?ἀ?πάσης	πάσης, πάσῃς ... ἀπάσης (11th item)	ἀπάσης
ἐὐθυντὴριον	[εἐ][ὐυ]θυντ[ὴή]ριον	εὐθυντήριον, εὐθυντήριόν, εὐθυντῆρι	εὐθυντήριον
πρώτος	πρ[ώῶ]τος	πρῶτος, πρῶτός, πρωτὸς	πρῶτος
Κύρος	[KXH][ύῦῖ]ρος	Κῦρος, Κῦρός, Κύπρος	Κῦρος
εθηκε	[εἔ]θηκε	ἔθηκε, ἔθεκέ, θῆκε	ἔθηκε
Δυδῶν	[ΔΛ]υδῶν	Δυῶν, Διδῶν ... Λυδῶν (6th item)	Λυδῶν
λάὸν	λ[αά][ὸό]ν	λαὸν, λαόν, Λάιόν	λαὸν
ἤλασεν	[ἢἤ]λασ[ετ]?ν	ἤλασεν, ἤλασέν, ἤασεν	ἤλασεν
ευφρων	ε?ι?[υὔ]φρωο?ν	ἐύφρων, Εὔφρων, εὔφρων (correct)	ευφρων

Figure 2.2: Spell-checking supported by OCR evidence

2.3.7 Test on a Latin *incunabulum*

The last test has been performed using a singular engine, OCRopus, on Augustinus’ *De Civitate Dei*, Venetiis 1475. We were interested in training OCRopus with Latin abbreviations and ligatures, encoded in Unicode according to the directions of the Medieval Unicode Font Initiative[1]. Images have been preprocessed with the OCRopus libraries for morphological operations, such as erosion and dilation (see Shih, 2009) to smooth the character image and improvements due to preprocessing have been compared to ground truth.

[1]http://www.mufi.info/fonts

2.4 Results

Results are evaluated comparing the accuracy of singular engines with the accuracy of the merged, spell-checked output. In order to compute the accuracy, the final output has been aligned with the ground truth. Following Reynaert (2008a), the accuracy has been calculated as:

$$\frac{matches}{matches + substitutions + insertions + deletions} \tag{2.7}$$

Accuracy is calculated as the ratio between the number of characters correctly recognized and the sum of correct, substitued, wrongly inserted and missed characters provided by the OCR output.

2.4.1 Accuracy of the single engines

The training sets created for each collection determine accuracy performances of single engines. Results are shown in Tab. 2.1. Both the most accurate OCR commercial application, Abbyy FineReader, and the most accurate OCR open source application, OCRopus/Tesseract are now provided with training sets that allow them to deal with polytonic Greek. In the case of Kaibel's exemplar, we have obtained better results with OCRopus/Tesseract than with Abbyy FineReader, suggesting that the open source software is currently mature enough to be applied to classical critical editions.

As said above, FineReader allows to apply the user defined training sets either alone or in bundle with the built-in trainings. As shown in the table, without built-in trainings performances, in average, are better than with built-in trainings.

Results on Kaibel's and Meineke's exemplars, both Teubner editions, have been obtained using a single training set. The similarity of these results suggest that the project is scalable with pre-processing data reusable on exemplars of the same collections.

2.4.2 Improvements due to alignment and spell-checking

Improvements due to alignment can be divided in two steps. In fact, the first gain is due to the alignment of the FineReader outputs, with and without the built-in training set, in cooperation with the user training set. In average, the

Edition	FR w/o built-in training	FR with built-in training	OCRopus	Anagnostis
Gulick (Loeb)	96.44%	94.35%	92.63%	93.15%
Kaibel (Teubner)	93.11%	93.15%	95.19%	92.97%
Meineke (Teubner)	94.54%	93.79%	92.88%	91.78%
Hermann (Weidmann)	97.41%	—	91.84%	78.64%

Table 2.1: Accuracy: single engines

improvement is +1.15% in relation to the best single engine, which is FineReader without the built-in training except in the case of Kaibel, as stated in the previous section.

The second step is the triple alignment and constrained spell-checking, which provides a gain, in average, of +2.49% in relation to the best single engine. A t-test for each exemplar demonstrates that improvements are always significant, with $p<0.05$. Analytical results are provided in Tab. 2.2. Alignment alone provides, in average, an improvement of 1%.

The best result, as expected, concerns the most recent Loeb edition, with an accuracy rate of 99.01%. If we consider only the case insensitive text (without punctuation marks, breathing marks and accents), the accuracy arises to 99.48%. This value is especially important if we are interested in evaluating the expected recall of a text retrieval system, where ancient Greek words can be searched in upper case.

Edition	Alignment and spell-checking	Aligned FR	Best engine
Gulick (Loeb)	99.01%	98.02%	96.44%
gain	+2.57%	+1.58%	0.00%
Kaibel (Teubner)	98.17%	95.45%	95.19%
gain	+2.98%	+0.26%	0.0%
Meineke (Teubner)	97.46%	96.15%	94.54%
gain	+2.92%	+1.61%	0.00%
Hermann (Weidmann)	98.91%	—	97.41%
gain	+1.50%	—	0.00%

Table 2.2: Accuracy: alignment and spell-checking

2.4.3 Accuracy on the critical apparatus

Tests on the critical apparatus of Gulick's and Kaibel's editions have been performed without a specific training for the footnote size, but with the same training sets applied to the rest of the page. Only the FineReader output with the built-in training set has been used, because the output created without it had a very low accuracy.

The average accuracy due to the triple alignment is 92.01%, with an average gain of +3.26% in relation to the best single engine, that is FineReader on Gulick's edition and OCRopus/Tesseract on Kaibel's edition. Analytical results are provided in Tab. 2.3. Also on the critical apparatus, t-test demonstrates that improvements are significant, with $p<0.05$.

It is important to point out that the critical apparatus, according to estimations computed in Stewart *et al.* (2007), is approximately, on average, 5% of the page in editions with minimal information (such as Loeb editions), and 14% of the page, on average, for more informative apparatus (such Teubner editions).

	Alignment and spell-checking	FR with b.-in	OCRopus	Anagnostis
Gulick	90.88%	87.99%	64.79%	59.08%
gain	+2.89%	0.0%	-23.20%	-28.91%
Kaibel	93.14%	87.68%	89.54%	57.11%
gain	+3.60%	-1.86%	0.0%	-32.43%

Table 2.3: Accuracy: critical apparatus

2.4.4 Accuracy on the *incunabulum*

The test performed with OCRopus on Augustinus' *De Civitate Dei* provides an accuracy of 81.05%, confirming results reached by Reddy and Crane (2006).

2.5 Modelling the manual correction process

We have shown above tests based on few pages, but we must shift attention to the scalability of the project, in order to build the digital library of classical editions.

In the production work-flow, the results achieved by automatic procedures are sent to an external data entry service (see Crane *et al.*, 2006). General purpose data entry firms work with multilingual and multiscript documents, on which they have no specific competence. Indeed, the recognition of errors performed by data entry employees is espected to be based on graphical comparison between the page image and the digital text. For this reason, texts that must be manually corrected are sent to the data entry firm as OpenOffice documents, using fonts that are most similar to the original page scripts, such as the Porson fornt for Loeb editions or the Teubner font for Weidmann and Teubner editions.

In order to speed up the correction process, the spell-checker integrated in OpenOffice 3.0, Hunspell, has been provided with the ancient Greek dictionary, using the word list produced by *Morpheus* illustrated above. Statistics on the first samples (20 pages) sent to the data entry firm have shown that, in this phase of assessment, the time spent by different members of the team for corrections is not correlated to the number of errors contained in each page. These data are useful to observe the individual learning curve of the correctors, which must deal with a large set of characters to recognize and with a complex keyboard layout to insert corrections that are not suggested by the spell-checker.

The standard accuracy expected on the output corrected by data entry professional services is 99.95 percent (see Stewart *et al.*, 2007).

2.6 Remapping the text on the page image

When FineReader and Ocropus perform OCR, they can map the recognized characters on the original image page. FineReader can save PDF files with text mapped on the page, whereas Ocropus provide character coordinates, which can be processed in DjVu files, but the original mapping must be adjusted after the corrections performed during post-processing.

Both PDF and DjVu formats are commonly used for OCR page images. The first one is most popular, but the second one is most optimized for page images, because it uses a compression algorithm capable to separate the page background (on which highly lossy compression can be applied) and the script foreground (on which lower compression can be applied).

Both formats support the addition of a text layer beneath the image, which allows the document to be searchable. For instance, Google Books[1] provides searchable .pdf files and Internet Archive[2] provides both searchable .pdf and .djvu files, associated to reach metadata encoded in .xml.

There are many advantages to map the text on the page image; in particular, the digital text is strongly coupled to its image source, the original layout is preserved, even if the searchable content is formatted in plain text and the user can easily reject the result of a query, comparing the wrong text retrieved with the correct word image.

The three OCR engines have different capabilities about the text mapping on the page images. In fact, Anagnostis does not export any information about the mapping, because it deals only with plain text or .rtf files. FineReader can export searchable .pdf files. Anyway, when a .pdf file or a .djvu file are imported, the mapped text is unchable and a new OCR performed on the images extracted from the .pdf discard the previous text. OCRopus/Tesseract, on the contrary, can export in a format easy to be modified and remapped on the image.

2.6.1 hocr microformat

In general, OCR engines isolate each character or sequence of characters that must be recognized in a box, defined by its upper left and bottom right positions. OCRopus scripts can access this information provided by the Tesseract engine, in order to enrich the textual output with the coordinates on the page. In fact, OCRopus deals with two output formats: plain text and html enriched with a specific microformat, called hocr. A microformat is a reasonable trade-off between pure html and more sofisticated annotation systems, because it extends html tags with attributes that can be processed by parsers or ignored by web browsers. (For a general introduction to microformats, see Allsopp, 2007)

We can consider the first line of the page image example in Fig. 2.3. The corresponding plain text recognized by the OCR engine is *θεσμῶ'ν ἐν Συρακούσαις φησὶ τοῖς παντελείοις το·ν*, with three errors: an addition in the case of *θεσμῶ'ν*

[1] http://books.google.com
[2] http://www.archive.org

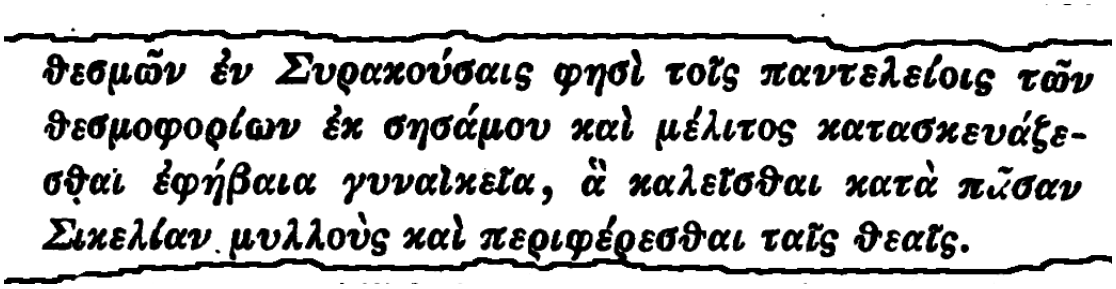

Figure 2.3: Page image example

instead of $\theta\epsilon\sigma\mu\hat{\omega}\nu$ and a substitution plus an addition in the case of $\tau o{\cdot}\nu$ instead of $\tau\hat{\omega}\nu$.

The hocr microformat stores the mapping information in the `title` attributes of `div`, `p` and `span` tags. The `class` attribute determine the nested division: `ocr_page`, `ocr_par` or `ocr_line`. In Fig. 2.4 it is possible to see the coordinates related to the first line of Fig. 2.3. The coordinates are provided character by character, in a long sequence of numbers related to the the entire line.

2.6.2 djvuxml format

The DjVuLibre Project[1] provides an open source DjVu library, a viewer and command line tools. In particular, `djvutoxml` extracts the searchable text from a .djvu file. The coordinates of the words on the page images, stored in the same .djvu file, are encoded in xml.

The .xml file contains metainformation about scanning parameters, such as dpi and gamma, and about the page image location. The hierarchy of the hidden text divisions is: page column, region, paragraph, line and word. The smallest unit that is mapped is the word, not the character.

The text extracted with `djvutoxml` can be manipulated with a text editor and, eventually, injected again in the .djvu file with `djvuxmlparser`. This tool reads the metainformations related to the page images and remaps the text on the original page.

[1] http://djvulibre.sourceforge.net

<html>
[...]
<body>

<span class="ocr_line" title="bbox 114 187 1099 234; bboxes 114 194 140 224, 140 203 154 224, 155 199 175 224, 175 201 198 233, 200 189 228 224, 222 188 227 197, 226 200 249 224, 271 187 293 223, 271 187 293 223, 289 202 312 223, 332 192 368 222, 332 192 368 222, 366 202 390 223, 390 201 411 233, 411 200 433 223, 433 200 454 223, 455 200 473 221, 473 187 498 222, 498 197 520 222, 519 200 543 221, 543 201 554 221, 556 200 574 228, 593 201 620 233, 593 201 620 233, 621 202 643 234, 643 199 664 224, 665 189 678 224, 697 201 717 223, 697 201 717 223, 717 202 736 223, 737 195 753 224, 751 202 767 229, 787 201 812 225, 787 201 812 225, 812 203 834 225, 835 202 878 224, 835 202 878 224, 879 202 897 225, 897 194 918 225, 920 203 937 226, 939 193 957 226, 953 204 972 227, 973 206 986 227, 987 206 1007 233, 1027 208 1049 229, 1027 208 1049 229, 1049 209 1067 230, 1063 198 1078 230, 1076 210 1099 231">θεσμῶ'ν ἐν Συρακούσαις φησὶ τοῖς παντελείοις το·ν<br/ >
[...]

[...]

[...]
</body>
</html>

Figure 2.4: hocr format example

With simple transformations, the .html files enriched with hocr microformat are translated in the djvuxml format. The structure of the two files are very similar, but the character by character coordinates, which must be reduced to the word by word coordinates, keeping only the upper left corner of the first word character and the bottom right corner of the last word character. The result of the mapping is shown in Fig. 2.5

2.6.3 Stand-off mark-up and realignment

As illustrated above, manual corrections are performed by the data entry firm on OpenOffice documents, that provide the ultimate correct plain text. In order to remap the correct text on the original image, it is necessary to align it with the OCRopus output, that contains recognition errors, but that features the page coordinates.

```
<DjVuXML>
<HEAD>file://localhost/ocr/ath-meineke/djvu/ath35.djvu</HEAD>
<BODY>
<OBJECT data="file://localhost/ocr/ath-meineke/djvu/ath35.djvu"
type="image/x.djvu" height="1967" width="1275" usemap="ath35-194.djvu">
<PARAM name="DPI" value="100"/>
<PARAM name="GAMMA" value="2.200000"/>
<PARAM name="PAGE" value="ath35-194.djvu"/>
<HIDDENTEXT><PAGECOLUMN><REGION><PARAGRAPH>
<LINE>
<WORD coords="114,194,249,233">θεσμῶ'ν</WORD>
<WORD coords="271,187,312,233">ἐν</WORD>
<WORD coords="332,187,574,233">Συρακούσαις</WORD>
<WORD coords="593,187,678,234">φησὶ</WORD>
<WORD coords="697,187,767,234">τοῖς</WORD>
<WORD coords="787,187,1007,234">Παντελείοις</WORD>
<WORD coords="1027,187,1078,234">το·ν</WORD>
</LINE>
[...]
</PARAGRAPH></REGION></PAGECOLUMN></HIDDENTEXT>
</OBJECT>
</BODY>
</DjVuXML>
```

Figure 2.5: DjVuXML format example

Alignment can be performed only between two sequences of plain text. For this reason the OCRopus .html output must be processed in order to strip out the mark-up. A technique of stand-off mark-up is applied, in order to split in two different files the plain text with the OCR content and the pointers that will allow us to recover the positions of the original tags, after the alignment. In Tab. 2.4 it is possible to see the alignment of the corrected text with the target character positions of the OCRopus output. Substituted characters inherit the same box coordinates; inserted characters share the same box coordinates with the neighbor on the left and deleted characters ignore the box coordinates.

The resulting .html file with the corrected text mapped on the original coordinates can be easily transformed in the djvuxml format, according to the procedure seen above. The result is illustrated in Fig. 2.6

Correction	θεσμῶ_ν ἐν Συρακούσαις φησὶ τοῖς παντελείοις τῶ_ν
	||||| ||||||||||||||||||||||||||||| |
Target	θεσμῶ'ν ἐν Συρακούσαις φησὶ τοῖς παντελείοις το·ν

Table 2.4: Alignment for remapping

[...]
<LINE>
<WORD coords="114,194,249,233">θεσμῶν</WORD>
<WORD coords="271,187,312,233">ἐν</WORD>
<WORD coords="332,187,574,233">Συρακούσαις</WORD>
<WORD coords="593,187,678,234">φησὶ</WORD>
<WORD coords="697,187,767,234">τοῖς</WORD>
<WORD coords="787,187,1007,234">Παντελείοις</WORD>
<WORD coords="1027,187,1078,234">τῶν</WORD>
</LINE>
[...]

Figure 2.6: Result of the realignment

When the .xml file is eventually injected in the original .djvu file with the page images, the file becomes fully searchable and the correct text is retrieved with a reference to the correct position of the words in the page, as shown in Fig. 2.7.

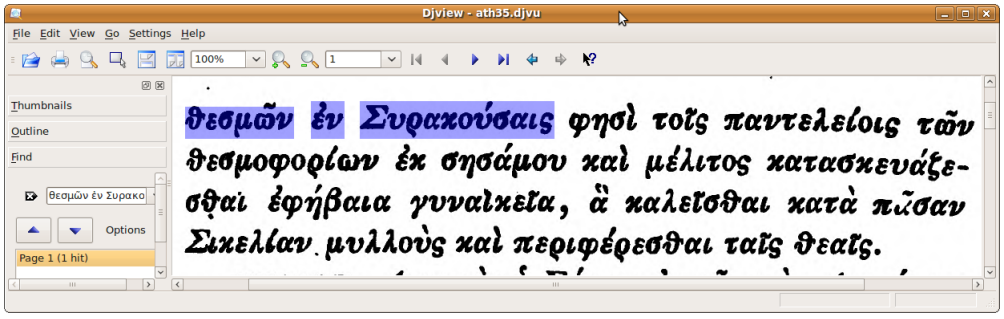

Figure 2.7: Text retrieval in Djview

2.7 Conclusion

As claimed in Crane *et al.* (2006), in order to go beyond digital incunabula it is necessary to build a digital library of classical critical editions, on which information extraction, natural language processing and corpus analysis techniques should be perfomed. A satisfactory OCR accuracy rate for the whole content of a critical edition (text and apparatus), that will allow us to lower the costs for post-corrections by hand, is one first necessary step to build the new generation of textual corpora.

In this chapter we have illustrated how multiple alignment and spell-checking supported by OCR evidence can improve OCR accuracy on classical editions and we have discussed the work-flow to scale the digitization process, from scanning to remapping of corrected text on the original page image.

3

Alignment of variant readings

The principal corpora currently available in classical literature, although quite thorough, are based on authoritative editions without critical apparatus. However, philologists need to deal with textual variants attested by manuscripts and conjectures suggested by scholars through the centuries and they need to annotate linguistic and metrical features not only on the reference editions but also on the collection of variants and conjectures. This chapter[1] is focused on methods for automated extraction methods applied to digitized apparatus of critical editions and digital repertories of conjectures.

After a general introduction (section 3.1), section 3.2 illustrates current approaches to add apparatus to digital critical editions. Section 3.3 lists the critical editions and the repertories used in the experiments.

Section 3.4 illustrates the textual operations involved in alignment: addition, deletion, substitution and transposition. Section 3.5 explains different types of alignment: sequence by sequence, word by word and character by character alignment. Section 3.6 illustrates the algorithms used in the current work and exemplifies the results, whereas 3.7 explains how lemmatization can improve the alignment. Section 3.8 provides data about the performances of the algorithms and section 3.9 discusses the results.

Section 3.10 addresses position annotation word distance issues. Section 3.11 summarizes the main achievments.

[1] See Boschetti (2008).

3.1 Introduction

Literary corpora are usually collections of texts. But from a philological point of view, this simple assertion raises non trivial questions. In fact, classical texts are the result of a complex process of corruptions and corrections. The editor must evaluate variants contained in manuscripts and conjectures suggested by scholars during the centuries, in order to reconstruct a textual hypothesis. Therefore, the text established is the result of a selective process that involves good knowledge of tradition, of the author's style and of linguistic and historical context. Choices are motivated, but subjective: a new edition is always different from the previous ones. The editor can remain close to the textual evidence given by manuscripts, can prefer sharp conjectures suggested by reputable scholars in last centuries or can suggest his own emendations. He is under the influence of his school, its tradition and its current hermeneutic paradigm.

From this perspective, we must be aware that when we use a literary corpus, we are dealing with authors' texts filtered by editors. The problem is that we cannot study a linguistic or stylistic phenomenon if that phenomenon is masked by the choices of the editor. A typical example is the study of repetitions: the earlier paradigm tended to consider many short-term repetitions as mistakes made by copyists, therefore the editors preferred to delete or to replace these repetitions by (arbitrary) conjectures. The new paradigm, instead, recovers this stylistic device as a genuine one: the unexpected result discovered by Pickering (2000) is that scribes were trained to remove repetitions, instead of introducing them. If we want to support this claim by stylistic analyses of digital corpora, we do not find many repetitions attested in manuscripts precisely because editors suppressed them, so concordances based on these editions do not allow the study of the phenomenon to its real extent. We can recover it only by an accurate comparison of information stored in critical apparatus, where almost all variants and several conjectures are recorded.

The most complete collections of ancient Greek and Latin texts, such as the Thesaurus Linguae Graecae and the Packard Humanities Institute's CD-ROMs of Latin literature, are based on authoritative modern editions, but they lack critical apparatus. Therefore, the digital texts usually do not contain information about textual variants attested in manuscripts or conjectures suggested by scholars. Philologists use digital corpora but they must verify results on printed editions, in order to evaluate if the text retrieved is attested in every manuscript, only in the *codex optimus*, in an error prone family of manuscripts, in a scholium, in the indirect tradition or if it is conjectured by a modern scholar. In short, the text of the reference edition has no scientific value

without the apparatus, and the criticism by Degani (1992), that the philologist must work always also on printed editions, unfortunately is still valid. As we pointed out above, the text of the reference edition is the result of the choices made by the editor, who subjectively evaluates different likelihoods of variants and conjectures, keeping the preferred one.

Yet even the critical apparatus is a selection. If the final text is subjective in its substitutions, the critical apparatus is subjective in its omissions. The critical apparatus records variants and conjectures with bibliographical references, but it can be considered an anthology and not an exhaustive repertory of them. Only repertories of collations and repertories of conjectures can claim completeness, even if the first one is limited by the number of manuscripts investigated and the second one by the number of printed editions, commentaries and articles reviewed.

By the motivations explained above, the interest to enrich literary corpora with variants and conjectures is growing and it focuses the attention of several research groups; among many others, the *Homer Multitext Project*[1] at Harvard University and the *Musisque Deoque Project*[2] at Università di Venezia, for Latin texts. For a theoretical background about the relation between texts and apparatus in digital editions, see Froger (1968), Bozzi *et al.* (1986), Buzzetti (1999), Mordenti (2001) and Bozzi (2004b).

Digital corpora of ancient languages can be extended not only with variants and conjectures but also with annotations about lemmatization, parts of speech, morphological and metrical features, etc. Extensions to the same corpus can be asynchronous and performed by independent groups and institutions. In these cases, problems of maintenance, compatibility, cross reference and inheritance of features arise.

It is difficult to determine the basic unit of variants and conjectures. If attention is focused on how they originate from a paleographic point of view, the single character seems the most suitable basic unit. But from a linguistic and stylistic point of view, the basic unit should be the word, which can be chained in superunits (for example the verse that contains the variant), or splittable in subunits (for example the single characters or all the partitions of the verse, encoded in *scriptio continua*, that can match attested forms).

This chapter illustrates a method to automatically extract information from critical apparatus and repertory of conjectures, aligning word by word the items of the variant readings and the words that the variant should substitute in the context of the verse(s).

[1] htt://www.chs.harvard.edu/publications.sec/homer_multitext.ssp
[2] http://www.mqdq.it

3.2 Background

Currently, there are two main approaches to add apparatus to digital critical editions. The first one is based on automatic collations of diplomatic editions. Digital diplomatic editions are complete transcriptions of single manuscripts, enriched by information about layout, position and function (comment, correction, etc.) of any portion of text in the page, etc. Usually they are encoded in XML, according to the T.E.I. directions[1]. They can be used for rendering the original witness in a typographical fashion, for mapping (and retrieving) the digital text on the image of the page or for automatic collations, that are exploited by techniques similar to concurrent version systems (CVS or Submission). By the mean of the mark-up language, it is possible to separate the actual text of the manuscript from its interpretations: corrections, normalisations, explanation of abridgements, etc. This method is particularly useful with a restricted number of manuscripts, in absence of large secondary literature (commentaries, articles, etc.). The second approach is based on the employment of forms filled manually by operators. It is useful if the aim is the acquisition of large amounts of apparatus' information, on many texts of different authors. This method, for instance, is currently applied by the *Musisque Deoque Project*, that aims to give, for the entire corpus of the poetical Latin literature, at least a minimal apparatus: the principle of this project is that it is better to have essential critical information for the entire corpus than extremely accurate apparatus for a very restricted group of texts. Forms have fixed fields, so the operators must adapt the actual information of the original apparatus to the digital grid. Usual fields are: text of the variant or conjecture, indication of manuscript or scholar's name and notes where less structured, unprocessed information can be stored.

Both methods have their limitations. Digital diplomatic editions have a practical, economical limit in the number of operators that can perform transcriptions. The theoretical limit is more insidious. Automatic collation is based on the idea that each document (transcription of a manuscript or OCR recognition of a printed edition) is a complete instance of the text to reconstruct, with variations. From the reference edition and the database of automatic collations (the complete set of all differences of diplomatic editions to the reference edition) we can reconstruct every diplomatic edition previously collated. This assumption is very useful for the reconstruction of the *stemma codicum* that shows the relations between manuscripts, but it is inapplicable in other situations. When we have a very large direct and indirect tradition and a rich secondary

[1] http://www.tei-c.org

literature, we cannot always reconstruct a context for the variant or conjecture as large as the entire text. A variant that we extract from a *scholium*, an ancient commentary, has an indefinite context, because we do not know which was exactly the entire text read by the ancient commentator. Conjectures often are suggested in a disjunctive way: a *vel* b *vel* c, and sometimes we do not know which was the edition used by the scholar that invented the conjecture. If diplomatic editions are similar to layers that we can overlay, these last cases are similar to stickies that we do not know on which layer we should stick. If n diplomatic editions can be distributed on n dimensions, these chunks with an indefinite context theoretically exist in more complex topologies, generating an explosion of combinations. In short: diplomatic editions' collation methodology cannot cover the entire process of mapping readings on the reference edition, but must be integrated by other techniques.

The forms-to-fill methodology has a limit in the subjectivity of operators. They must decide how to adapt the original information of printed apparatus to the fields of the forms, how to integrate lacking information, how to omit the irrelevant one. Furthermore, there is no mapping between the original apparatus and the new adapted information. T.E.I. gives directions for this type of mappings, but the actual procedure (manual mark-up) is very difficult for large amounts of texts. For authors like Aeschylus, with a very large tradition and many conjectures registered in commentaries and reviews, both approaches are very time expensive for a single operator, and error prone and money expensive for a team that must follow a common protocol for annotations. The automatic parsing of apparatus and repertories, in addition to the automatic collation for a group of relevant diplomatic transcriptions, should be an acceptable trade-off. Subjective choices by operators in this case are limited to the correction phases. This third approach has a double goal: on one hand it aims to parse automatically existing critical apparatus and repertories of conjectures of Aeschylus and on the other hand it aims to discover heuristics useful for any collection of variants and/or conjectures with a similar structure. The accurate mapping of information extracted by apparatus and repertories must be used to build new critical editions, indexes, concordances and systems for information retrieval based on variants.

3.3 Reference editions and repertories

The first problem to tackle is the reference edition, that is the text that constitutes the basis for indices and concordances, the reference for commentaries and secondary literature, the line numbering system for apparatus and repertories. Usually the ref-

erence edition is the currently most authoritative edition, by agreement of scholars. Neverthless, when a new authoritative edition substitutes the previous one, old and new philological instruments map on different texts. Specifically, the present work on Aeschylus uses three different reference editions, because the critical apparatus and the repertories of conjectures by Wecklein (1885) and Wecklein (1893) are based on his own text (Wecklein, 1885), the collations of manuscripts executed by Dawe (1963) and his repertory of conjectures (Dawe, 1965) are based on Murray (1955), whereas the appendix of conjectures gathered by West (1990) and his own apparatus are mappable on West (1998). One edition can differ from another not only for textual variations, but even for disposition of verses, differently distributed on the lines, according to the metric and colometric interpretations of the editor. In this way, the reference to the number of the verse is not an effective device to switch from a reference edition to another one, because it is too ambiguous: e.g. *Pers.* 857-8 (Wecklein, 1885) *πανταρκής, ἀκάκας, | ἄμαχος βασιλεύς* have not the same distribution on vv. 855-56 (Murray, 1955) *πανταρκὴς ἀκάκας ἄμαχος βασι-| λεύς* ... because of a different colometry, i.e. the division of verses in *cola*, smaller parts. Only the sequential position of words in the entire text provides the grid to switch from one edition to another, and also the colometry and verse numbering is based on this grid: e.g. *βασιλεύς* is in the 4429th textual position in both editions, but the new line is mapped on the last character of the word in Wecklein (1885) and on the fourth character in Murray (1955). Complete collations of the three reference editions are performed, in order to have the grids for mapping apparatus and repertories on a unified system.

3.3.1 Collation and alignment of the reference editions

Murray (1955) is the main reference edition, because it has been the source for the annotated corpus built by the C.I.P.L. of Liège[1] (used for this work). Lemma and part of speech are associated to each word. Morphological features about declination and conjugation and metrical structure of each word have been added to the text of *Persae*.[2] Each word of its text has a progressive number, from the beginning to the end of each tragedy. The fact that Murray (1955) constitutes the main reference edition for the current work means that each word of its text has a progressive integer number, starting from the beginning of each tragedy.

[1] http://www.cipl.ulg.ac.be

[2] The first attempt of morphosyntactic analysis of the *Persae* is illustrated in Boschetti (2005). F. Mambrini completed the treebank of Aeschylus' seven tragedies at *Perseus Project*, http://www.perseus.tufts.edu. First results are illustrated in Bamman *et al.* (2009)

The other reference editions, aligned to this one, can have empty positions (if they differ for suppression of text: text that is present only in the Murray edition) or positions marked by decimal numbers (in case they differ for text addition: text that is between two consecutive positions in the Murray edition). Information contained in repertories is mapped on these grids. Apparatus and repertories, built along two centuries, differ in typographical conventions and in quantity of information, more or less accurate. However, the basic assumption is that it is possible to identify a small number of widely repeated schemes and expressions, in order to mark-up automatically every chunk of parsed information.

For instance, in Wecklein's reportory, verse numbers are followed by punctuation mark and different editions by the same editor are indicated by the formula: $conjecture_1$ olim, postea $conjecture_2$ *editor*, exemplified in:

132. λέκτρα δ' ἀμῖν μάταν Enger. πόνῳ Pauw, σπάνει olim, postea ἔρῳ Heimsoeth, ὀδῷ Oberdick.

In Dawe's repertory, verse numbers are not followed by punctuation mark and citations, related to journals and monographs, are followed by the page number, as exemplified in:

133 λέκτρα δ' ἀντ' ἀνέρων πόθωι Hoernle p. 89

3.4 Textual operations and structure of apparatus and repertories

Textual operations registered in critical apparatus and repertories of conjectures can be reduced to insertions, deletions, substitutions and transpositions. In fact, insertions can assume the specific function of iterations and deletions can be registered as *lacunae* or omissions, but from a computational point of view, the basic operations allow any transformation from the source string to the target string. Transposizion can be reduced to a deletion followed by an insertion in another place of the same text.

According to statistics performed on a sample that constitutes five percent of the Aeschylus' *Persae*, in apparatus and repertories roughly 90 percent of variants and conjectures are expressed only by the number of the verse and sequences of Greek words, followed by lists of witnesses or scholars' names. In most cases the sequence of Greek words represents a simple textual substitution, but sometimes the information is constituted by placeholders (boundary words identical to some

words in the reference edition) that provide the correct position to anchor a reading that contains a short addition, deletion or transposition of text. For instance, given the verse:
370 *ναυσὶν κρυφαίως δρασμὸν εὑρόντες τινά,*
in Wecklein's repertory we find:
370. *δρασμὸν ἄραντες* Naber.
The other 10 percent of variants and conjectures is composed of more complex structures, with a Latin sentence that expresses the textual operation that should be performed (e.g. *delet*, *iterat*, *transponit*, etc.). For instance, given the verse:
3 *καὶ τῶν ἀφνεῶν καὶ πολυχρύσων ἑδράνων φύλακες*
in Wecklein's repertory we find:
3. *καὶ πολυχρύσων* delet Bothe.
In the current work only sequences of Greek text followed by the responsible editor(s) of variant and conjectures are processed.

3.4.1 Manual annotation of samples

Apparatus and repertories (as well as commentaries) are organized by the editors in lines linked by reference to the text. In the first stage of the work, in order to discover the typical structures and evaluate their complexity and frequency, some samples (see Boschetti, 2005, pp. 33–52, for further information) extracted by apparatus and repertories have been annotated by hand, adopting a format easily transformable by XSL into a T.E.I. compliant one. Manual mark-up classifies the elements of each item and maps word by word different readings on the reference edition. An example of manual mark-up is below:
197. *ἥ δ' ἐσφάδᾳζε καὶ χεροῖν ἔντη δίφρου*
197. *αὐτὴ δίφρον* Canter.

```
<item>
  <verse>197.</verse>
  <reading>
    <g pos="824">αὐτὴ</g>
    <g pos="825">δίφρον</g>
  <scholar>Canter</scholar>.
  </reading>
</item>
```

Tags <item>...</item> divide repertory lines, devoted to one or more verses, numbered by <verse>...</verse>. Each line can contain one or more readings

(<reading>...</reading>). Readings contain sequences of Greek words, surrounded by <g>...</g> tags, with the pos attribute, which indicates the exact position of the word in the textual sequence. The name of the scholar that suggested the conjecture is surrounded by <scholar>...</scholar> tags.

Surveys (based on five percent of the tragedy) on the manual annotations confirmed that the most frequent chunk of information is constituted by 1) number of verse, 2) reading (variant or conjecture) that substitute one or more words in the text, 3) manuscript(s) or scholar(s) that exposes it. When the correspondence between the reading and the reference edition cannot be performed word by word, empty positions were filled by blanks, or decimal numbers were used in case of insertions. See the examples below, with a blank in 595th position:
mapping of κἀμὲ on καὶ με is annotated as

```
<item>
  <verse>164.</verse>
  <reading>
    <g pos="594">κἀμὲ</g>
    <g pos="595" val=""/>
  <scholar>Bothe</scholar>.
  </reading>
</item>.
```

Mapping of δείμα τ' on δείματ', which requires an insertion after 917th position, is annotated as

```
<item>
  <verse>213.</verse> ...
  <reading>
    <g pos="917">δείμα </g>
    <g pos="917.001">τ'</g>
  <scholar>Stanley</scholar>
  </reading>
</item>.
```

3.4.2 Reference to verses

Usually any line of the apparatus refers to one verse (e.g. 10.), but it might refer also to a range of verses, in particular to a successive couple (e.g. 10-11.), when the variant extends to both the verses. Rarely, the line refers to non-adjacent verses (e.g. 800 et 820.), for instance when the same variant (or conjecture) is repeated. The expressions *ante* and *post* are used if the variant or the conjecture

(usually an entire verse) must be inserted before or after an existing verse of the reference edition. Seldom reference to verses is not only at the beginning of the line, but also in the middle (e.g. when a conjecture is conditioned by the suppression of another verse).

3.4.3 Typology of readings and sources

The simplest (and fortunately rather frequent) case is when the reading is an orthographic or morphological variant that substitutes a single word in the reference edition. On the contrary, sometimes the variant splits the word in two parts: e.g. ἐν τλήμονι instead of εὐτλήμονι. When the substitution is a gloss, a synonym, a hypernym, a hyponym or an unrelated word, in apparatus and repertories it can be indicated by the formula x : y or x *pro* y (e.g. κιόντων Wecklein: ἰόντων codd.). When the substitution is large and complex, containing possible deletions and additions of text, usually the first and last words fit exactly the text of the reference edition.

Deletion usually is indicated by the word(s) to be deleted, followed by the expression *delet* (e.g. καὶ πολυχρύσων delet Bothe). Insertion of word(s) usually is indicated by the formula [*ante*/*post* x] *addit* y, where x is a word of the reference edition (e.g. ante βαλλὴν addit ἰωὰ Dindorf). Transposition is the combination of deletion and addition of text. It can be a simple inversion of words or it can affect one or more verses (e.g. 94-102 post 116 transponit OMueller). The source is one or more manuscripts for variants or one or more scholars for conjectures, which can be followed by an accurate bibliographical indication. Different apparatus and repertories can deal with different abbreviations for the names of manuscripts and scholars. Names must match items of a table that contains the canonical form of the name, abbreviations, orthographical variants and possible declinations (e.g. Paley: dat. Paleio). Information about sources can have different degrees of precision. For example, in West's apparatus each manuscript is always identified by name whereas in Wecklein's repertory usually manuscripts different by **M** (the *codex optimus*) are labelled just by recc. (i.e. *recentiores*). In West's apparatus each modern edition is identified by the name of its author and one number (e.g. Bothe^3), whereas in Wecklein's repertory previous editions are distinguished to the last one by the expression *olim* x (e.g. olim Bothe).

3.4.4 Complex cases

As shown above, the typical item structure is constituted by one or more reading-source couples about a part of the verse, possibly followed by one or more reading-source couples about other parts of the verse: verse reference - $\text{reading}_{1,1}$ $\text{source}_{1,1}$; ... $\text{reading}_{1,m}$ $\text{source}_{1,m}$ ... $\text{reading}_{n,n}$ $\text{source}_{n,n}$. For instance:

289 *στυγναί γ' Ἀθῆναι δάϊος·*

289. *στυγναὶ δ' Ἀθᾶναι* recc. *Δάϊος* Merkel, δαμόταις Oberdick.

In this case three chunks of information are easily separable in three reading-source couples. But complex cases are present in the repertories, which are constituted:

1) by groups of readings for a single source, as below:

36 *Πηγαστάγων Αἰγυπτογενής*,

36. *πηγασταγὼν* vel *πηγᾶς ταγῶν* vel *πηγᾶς ταγὼν* recc.

2) by variants of conjectures, as below:

468. *Ξέρξης δ' ἀνώμωξεν κακῶν ὁρῶν βάθος·*

468. *ἀνώμωξ, ἐν* (vel *ἔν*, olim *εὖ*) Bothe

3) by readings that contain conditions, as below:

155-156 *βασίλεια δ' ἐμή, προσπίτνω· | καὶ προσφθόγγοις δὲ χρεὼν αὐτὴν*

156. *καὶ προσφθόγγοισι χρεὼν* (vel si *πρσπιτνῶ* 155 deleatur) *προσφθόγγοισιν δὲ χρεὼν* Blomf.

The study of apparatus and repertory structures, facilitated by manual annotation of samples, allows the classification of readings and sources, with the help of the identification of recurrent patterns and isolation of exceptional complex cases.

3.4.5 Heuristics

Recurrent patterns examinated above can be automatically identified by suitable heuristics.

Each item is separated by a new line and the first task is the tokenisation of items. Tokens are classified in these categories: verse number, Greek word, Greek punctuation mark, metrical sign, Latin word, Latin punctuation mark, scholar name, bibliographical reference (title and pages). Verse numbers (as well as metrical signs) are identified by regular expressions and Greek words by the unicode

set of their characters. Greek punctuation marks are punctuation marks among Greek words. Scholar names, manuscript abridgements and bibliographical references (titles of books and reviews) are compared with information stored in tables. The table of scholar names is built by this heuristics: a scholar name is a Latin character word whose initial letter is always a capital letter (e.g. Abresch is recognized as a scholar name, but *Addit/addit* is automatically excluded). Manual control is necessary, in particular for the correct association of abridgements and orthographical variants. Tokens are then aggregated according to syntactic rules, in order to identify verse reference, readings and sources, as seen above.

3.5 Alignment

As seen above, c. 90 percent of readings, at least formally, are substitutions, i.e. chunks of text that should replace a reference edition's portion of one or more lines, represented in apparatus and repertories by a sequence of Greek words without predicates expressed in Latin language. Sometimes the substitution is only apparent: it is constituted by milestones (boundary words identical to some words in the reference edition) that give us the right position where to anchor the reading and surround a short addition, deletion or transposition of text. All substitutions, even the atypical ones, are parsed by an alignment algorithm,which will be illustrated below, in order to map the readings on the exact position of the verse in the reference edition. As we have seen in the previous chapter, alignment algorithms are well known, for instance, in genomic studies, where strings of proteins must be compared and aligned.

In fact, it is not enough to know in which verse the substitution must be performed; we need the precise position inside the verse, if we want to use all the amount of information stored by the parsing processes in order to create automatic indices and concordances and not only new print-like critical editions, with alternative readings on footnotes. A concordance needs to reconstruct a local context, and information retrieval systems, when they perform multiword queries, need to know which words actually are, or have the possibility to be, adjacent to other words.

3.5.1 Types of alignment

In apparatus and repertories, variants and conjectures are located only by reference to the verse, not by the precise position inside the verse. This information is superfluous for philologists and scholars, but it is not trivial to recover by automatic procedures. Alignment algorithms, that evaluate the similarity of a string with another string (or substring) are based on the edit distance, that is the evaluation of costs to perform additions, subtractions and substitutions in order to transform the first string into the second one or into a part of it. Following this principle, any chunk of text (the reading) can be aligned with the portion of text (the part of the line in the reference edition) with the lowest edit distance (i.e. greatest similarity). The alignment of variants with regions of the reference edition can be performed with different degrees of granularity for different purposes.

3.5.2 Sequence by sequence alignment

A coarse grained alignment identifies the part of the verse(s) in the reference edition that should be replaced with the variant (conjecture) or, in some cases, the point of insertion of the variant or the sequence in the reference edition that should be deleted.

Reference ed.	*Νεῖλος ἔπεμψεν·*	*Σουσισκάνης,* |	*Πηγαστάγων*	*Αἰγυπτογενής*
Blomfield	*Νεῖλος ἔπεμψεν·*	*Σουσας, Κάνης,*	*Πήγας, Πελάγων*	*Αἰγυπτογενής*

Table 3.1: Sequence by sequence alignment (*Pers.* 35-36)

This type of alignment is suitable for non-annotated corpora, for corpora annotated with features applied to verses or larger units (e.g. the metrical type of the verse, without details about the metrical structure of the words) and also for annotated corpora, if the correspondence between subunits of the variant and subunits of the affected verse is not relevant. The sequence by sequence alignment is preferable in cases of linking performed by human operators, because only the starting point and the end point must be determined, reducing individual choices. Among others, this solution has been adopted by the *Musisque Deoque Project*.

The link between the variant and the exact position in the verse is manually performed by operators, using a facility to drag and drop information of the critical apparatus on the reference edition.

3.5.3 Word by word alignment

When corpora are enriched by variants and conjectures, it is appropriate that redundant or irrelevant information is ignored, dropping the words of the reading with the mere function of placeholders. Tab. 3.2 shows possible ways to map the conjecture οὐδαμ᾽ οὗσ᾽ ἐμαυτῆς L. Schmidt to *Pers.* 165 μῦθον οὐδαμῶς ἐμαυτῆς οὗσ᾽ ἀδείμαντος, φίλοι. The last word, ἐμαυτῆς, helps the reader to find the correct position of the conjecture: it anchors it in the context of the reference edition, but it is not a necessary component of the reading.

Sequence by sequence alignment

Reference ed.	μῦθον	οὐδαμῶς ἐμαυτῆς	οὗσ᾽ ἀδείμαντος, φίλοι·
L. Schmidt		οὐδαμ᾽ οὗσ᾽ ἐμαυτῆς	

Word by word alignment and removal of placeholder(s)

Reference ed.	μῦθον	οὐδαμῶς	–	ἐμαυτῆς οὗσ᾽ ἀδείμαντος, φίλοι·
L. Schmidt		οὐδαμ᾽	οὗσ᾽	ἐμαυτῆς

Table 3.2: Identification of placeholders (*Pers.* 165)

If the items of the reference edition are annotated with lexical, morphological, metrical or semantic features, even the readings extracted from the repertories should be annotated according to the same criteria. One or more components of variants and conjectures often share the part of the reference edition that they should substitute with the same headwords, the same part of speech, the same metrical structure or the same synset. In this case, the fine grained alignment allows the inheritance of features associated with the correlated items. When the annotators fill the slots for the items of the variant, the default values suggested by annotation tools can be retrieved from the aligned items of the reference edition, according to a threshold of probability. For example, it is highly probable that two items with a small edit distance and different suffixes share the same headword; two words aligned with the same suffix probably share the same

morphological features; words aligned with a compatible prosody probably share the same metrical structure. Annotators can accept, reject or integrate these suggestions.

Reference ed.	βάσκε	πάτερ	ἄκακε	–	Δαριάν	οἶ.|
FWNewman	βάσκε	πατὴρ	ἀκάκας	ὁ	Περσᾶν	
Common headword		πατὴρ	ἀκάκας			
Common part of speech		Noun	Adj		Noun	

Table 3.3: Shared features by aligned items (*Pers.* 668)

After manual corrections and integrations, word by word alignment is useful to classify relevant items of the variant readings, in order to identify orthographic (same headword, same morphological features), lexical (different headwords) and morphological (same headword, different morphological features) variants. Metrical variants must be verified by applying sequence by sequence alignment, even if metrical structures of single words can be aligned and compared.

3.5.4 Character by character alignment

Character by character alignment is suitable when it is possible to assign to each manuscript or to each modern edition an independent layer and it is particularly useful for the study of errors caused by *scriptio continua*, which are very difficult for common systems of text retrieval (indexed word by word) to manage. With this type of alignment it is also possible to extract statistics on the substitution of characters, for paleographic purposes. The classic algorithms for alignment only take into account substitutions, insertions and deletions, but modified versions exist, which even take into account transposition of adjacent segments, or compression and expansion, where two contiguous units of one string correspond to a single unit of the other string.[1]

[1]Kondrak (2002) explains the application of these algorithms for language reconstruction, and provides the code, which has been adapted in the current work.

<table>
<tr><td>Reference ed.</td><td>πλαγκτοῖς ἐν διπλάκεσσιν.</td></tr>
<tr><td></td><td>ΠΛΑΓΚΤΟΙΣΕΝ- - ΔΙΠΛΑ - ΚΕΣΣΙΝ</td></tr>
<tr><td></td><td>| | | | | | | | | | | | | | | |</td></tr>
<tr><td></td><td>ΠΛΑΓΚΤ- - - ΕΝΣΠ - Ι - ΛΑΔ - ΕΣΣΙΝ</td></tr>
<tr><td>Hartung</td><td>πλάγκτ᾽ ἐν σπιλάδεσσιν</td></tr>
</table>

Table 3.4: Character by character alignment (*Pers.* 280)

3.6 Algorithms used in the current work

In the current work the alignment is performed in two steps. The first algorithm identifies the boundaries of the conjecture in the context of its verse(s) and the second one aligns the items word by word.

In the first step, a combinatorial algorithm compares the permutations of the words contained in the variant reading with all the subsequences generated by the verse(s) of the reference edition.

In the second step, a global alignment between the words of the variant reading and the words of the textual subsequence identified in the previous step is performed.

3.6.1 Combinatorial algorithm

The context of a conjecture is usually constituted by one or two verses and rarely by larger regions of text. In these conditions, a “brute force” combinatorial algorithm can be applied without excessive time consumption, increasing precision when compared to other optimized algorithms for alignment. Optimized alignment algortihms with block moves, necessary to deal with transpositions, are discussed, e.g., in Tichy (1984) and in Cormode and Muthukrishna (2007). But these kinds of algorithms do not fit well with intermediate units between the characters and entire strings, like words. In fact, the unit represented by a moved block is comparable to the prefix, suffix or stem, not to the inflected form as a whole.

Both the words of the conjecture and the words of the context (constituted by one or more verses) are capitalized and punctuation marks or spaces are erased.

Comparisons to find the best alignment are carried out using two nested loops. The external one provides every combination of adjacent words of the verse(s) in the reference edition, chained in a string. The internal one compares this string with relevant permutations of the words contained in the variant reading. Permutations are performed in order to find possible transpositions. Because the normalized edit distance between the strings determines the lowest similarity, the best score is assigned by

1 - edit_distance(str1, str2) / max(length(str1), length(str2))

An example will illustrate how the algorithm works. From *Pers.* 138-139 and the corresponding line in the Wecklein's repertory:
138-139. ἀκροπεν-|θεῖς ἑκάστα πόθῳ φιλάνορι
139. δ' ὄθη Schuetz
the algorithm reconstruct the following substrings:

ΑΚΡΟΠΕΝΘΕΙΣΕΚΑΣΤΑΠΟΘΩΙΦΙΛΑΝΟΡΙ ↓
ΑΚΡΟΠΕΝΘΕΙΣΕΚΑΣΤΑΠΟΘΩΙ ↓
ΑΚΡΟΠΕΝΘΕΙΣΕΚΑΣΤΑ ↓
ΕΚΑΣΤΑΠΟΘΩΙΦΙΛΑΝΟΡΙ ↓
ΕΚΑΣΤΑΠΟΘΩΙ ↓
ΕΚΑΣΤΑ ↓
ΠΟΘΩΙΦΙΛΑΝΟΡΙ ↓
ΠΟΘΩΙ ΔΟΘΗΙ / ΟΘΗΙΔ (best score)
ΦΙΛΑΝΟΡΙ ↑

The best score is assigned to the substring with the smallest normalized edit distance between itself and the conjecture under examination or one of its permutations. Due to the increase of time consumption, if the conjecture contains up to five words (the most frequent case), all the permutations are tested; if the conjecture contains up to ten words, only the words on the left and right boundaries are permuted in any position; if the conjecture contains more than ten words (very rare), the permutations are not performed.

3.6.2 Global alignment algorithm

The second step is a global alignment between the items of the variant reading and the items of the subsequence of context identified in the previous step. Navarro and Raffinot (2002) and Crochemore *et al.* (2007) provide detailed explanations about global (Needlman-Wunsch) and local (Smith-Waterman) alignment. A global alignment fits better with similar strings of similar length, whereas a local alignment attempts to identify similar regions in dissimilar strings. Global alignment is suitable in this case, because the similarity between the variant reading and the affected region of the reference edition has been established in the previous step.

The global alignment algorithm evaluates the costs to transform one sequence into the other one, minimizing the costs of substitutions, insertions and deletions. Substitutions have different costs, according to the similarity of the items substituted. In our case, identical words have the highest degree of similarity, which decreases according to the normalized edit distance between words. In the current work, similarity values are rescaled from -1 to 1. According to Tab. 3.5, for example, τε and ἐφαάνθην are totally dissimilar (-1), ἄρ᾽ and ἆρ᾽ are identical and πατρῴᾳ, which is the intended result, even if different from πατρίᾳ, is evaluated as very similar (0.75). The evaluation is performed on the capitalized characters, i.e. excluding differences due to accents.

Even the cost of gaps can be tuned. In the current work insertions and deletions have a penalty of -1, that is the same penalty used for a substitution with a totally dissimilar word.

The weight matrix (Tab. 3.6) is filled by assigning to each cell the minimal cost among an insertion (cell[i-1,j]+gap_penalty), a deletion (cell[i,j-1]+gap_penalty) and a substitution (cell[i-1,j-1]+similarity_score). The reconstruction of the path that produced the result in the bottom right cell determines the sequence of substitutions (movement on the diagonal), insertions (movement towards left) or deletions (movement to the top).

	Reference ed.	γᾶ	τε	πατρῴᾳ	κακὸν	ἄρ᾽	ἐγενόμαν
Brunck		*ΓΑΙ*	*ΤΕ*	*ΠΑΤΡΩΙΑΙ*	*ΚΑΚΟΝ*	*ΑΡ*	*ΕΓΕΝΟΜΑΝ*
καὶ	*ΚΑΙ*	0.33	-1	-0.50	-0.20	-0.33	-0.75
γᾶ	*ΓΑΙ*	1	-1	-0.50	-0.60	-0.33	-0.50
πατρίᾳ	*ΠΑΤΡΙΑΙ*	-0.43	-0.71	0.75	-0.71	-0.43	-0.75
κακὸν	*ΚΑΚΟΝ*	-0.6	-1	-0.75	1	-0.60	-0.50
ἅρ᾽	*ΑΡ*	-0.33	-1	-0.50	-0.60	1	-0.75
ἐφαάνθην	*ΕΦΑΑΝΘΗΝ*	-0.78	-1	-0.78	-0.56	-0.78	-0.56

Table 3.5: Similarity matrix (*Pers.* 936-937)

	Ref. ed.	γᾶ	τε	πατρῴᾳ	κακὸν	ἄρ᾽	ἐγενόμαν
Brunck	0	-1 $^{-1}$	-2 $^{-1}$	-3 $^{-1}$	-4 $^{-1}$	-5 $^{-1}$	-6 $^{-1}$
καὶ	-1 $^{-1}$	-0.33$^{-0.33}$	-1.33 $^{-1}$	-2.33$^{-0.50}$	-3.20$^{-0.20}$	-4.20$^{-0.33}$	-5.20$^{-0.75}$
γᾶ	-2 $^{-1}$	0 ^{1}	-1 $^{-1}$	-1.83$^{-0.50}$	-2.83$^{-0.60}$	-3.53$^{-0.33}$	-4.53$^{-0.50}$
πατρίᾳ	-3 $^{-1}$	-1$^{-0.43}$	-0.71$^{-0.71}$	-0.25 $^{0.75}$	-1.25$^{-0.71}$	-2.25$^{-0.43}$	-3.25$^{-0.75}$
κακὸν	-4 $^{-1}$	-2$^{-0.60}$	-1.71 $^{-1}$	-1.25$^{-0.75}$	-0.75 ^{1}	-0.25$^{-0.60}$	-1.25$^{-0.50}$
ἅρ᾽	-5 $^{-1}$	-3$^{-0.33}$	-2.71 $^{-1}$	-2.21$^{-0.50}$	-0.25$^{-0.60}$	1.75 $^{-1}$	0.75 $^{-075}$
ἐφαάνθην	-6 $^{-1}$	-4$^{-0.78}$	-3.71 $^{-1}$	-3.21$^{-0.78}$	-1.25$^{-0.56}$	0.75$^{-0.78}$	1.19 $^{-056}$
– γᾶ τε πατρῴᾳ κακὸν ἄρ᾽ ἐγενόμαν καὶ γᾶ – πατρίᾳ κακὸν ἅρ᾽ ἐφαάνθην							

Table 3.6: Weight matrix

3.7 Lemmatization

In order to improve alignment performance, the similarity of words with a high probability of having the same lemma is scored 1 by using a method that, to the best of our knowledge, has been applied for the first time. In fact, it is appropriate that forms of the same paradigm are aligned independently by their edit distance (for example, different forms of **φέρω** can have a very low edit distance, if compared with each other). In order to fulfill the lemmatization, every word of the reference edition is associated with its lemma retrieved in the

C.I.P.L. annotated corpus. Because the C.I.P.L. corpus was manually annotated, the accuracy is close to 100 percent. The probable lemmata of the inflected form present in the variant readings are retrieved by searching for the form in the annotated corpus. If the result is null, the form is parsed by the morphological analyzer *Morpheus*. Each element of the array of lemmata retrieved with this method is compared with the lemma associated with each word of the context verse(s) from the reference edition. If lemmata match, the similarity score is 1 and it is inserted into the similarity table.

3.8 Alignment performance

Performances are calculated on 56 verses of the Wecklein's repertory on Persae (about five percent of the entire tragedy constituted by 1076 verses). Correct mapping of conjectures on the reference text have been performed by hand. Each processed item is constituted by a singular conjecture. Processed items (83 on 95: c. 87 percent) are limited to formal substitutions (i.e. items containing Latin predicates are excluded). Correct processed items are 73: c. 77 percent onto the total but a rather encouraging 88 percent on the processed items. In Tab. 3.7 results are compared with methods adopted in previous stages.

	Mapping word by word	Mapping chunk by chunk without permutations	Mapping chunk by chunk with permutations
Absolute percentage of correct mappings	69	74	77
Percentage of correct mappings only on processed items	79	85	88

Table 3.7: Performances

Mapping word by word was performed by the evaluation of edit distance between any word of the reading and each word of the line in the reference edition. The algorithm shows bad performances with inserted and split words, as expected. Match without permutations is less efficient than match with permutations, even if permutations can produce errors avoided by the former algorithm. A short explanation about the performance of the final algorithm: correct mapping is driven by same beginnings and/or endings, e.g. 10 ὀρσοπολεῖται mapped on ὀρσολοπεῖται and διακλονεῖται even mapped on ὀρσολοπεῖται, or by the aid of

milestones, e.g. 166 μέγας στρατὸς on μέγας πλοῦτος, 365 οὐδὲ δαιμόνων on οὐδὲ τὸν θεῶν. The catenation of words in unique strings to check, as seen above, allows different segmentations, e.g. 165 οὖσα δείματος on οὐσ' ἀδείμαντος; the mapping of two words onto one word, e.g. 36 πηγᾶς ταγῶν on Πηγαστάγων, 75sq ποίμναν ἀνέρων ον ποιμα-| νόριον, 641 ἆρ' on ἦ ῥ', or, on the contrary, the mapping of one word onto two words, e.g. 636 δ' ἀμβαύζω on διαβοάσω. Permutation of reading's elements allows the correct mapping for short transpositions: 330, e.g. πλεῖστον εἷς ἀνὴρ on εἷς ἀνὴρ πλεῖστον.

3.9 Discussion on alignment results

The alignment performed in this work is a trade-off between the alignment of the most similar items and the prevention of unnecessary gaps. For this reason, sometimes the words aligned have only two or three letters in common (e.g. *Pers.* 199. ἄνευ aligned to δεσμοὺς, because they share ε and υ, even if they are morphologically unrelated, considering that neither lemmata nor affixes are shared). Anyway, the trade-off is generally satisfying because aligned words belong to the same paradigm, or have the same suffix or prefix, or have many contiguous characters in common. A lower gap penalty could increase the number of insertions and deletions. The upper limit is the search for the longest common sequence, where only equal items are aligned, thus preventing substitutions.

The evaluation of edit distances fits many cases of mapping readings on their contexts. But there are also errors unrecoverable by optimisation of edit distance techniques. The philologist usually is helped by the editor with milestones. On the contrary sometimes the editor knows that syntactic, semantic or metric knowledge is enough to place the *varia lectio* in its context, but this metric and syntactic knowledge currently is unsupported by our alignment algorithm. E.g. *Pers.* 210 θοοῖς is correctly mapped on δρόμῳ by the human philologist because both words are in dative, information not managed by the current algorithm. Fortunately, these cases are very rare.

3.10 Annotation of positions and word distance issues

Alignment provides mapping of digital variants onto reference edition positions, allowing text retrival systems to deal not only with canonical texts, but also with variant readings. For this reason, textual position annotation should be functional to enhance performance of text retrieval systems.

In order to perform text retrieval operations on annotated corpora, it is necessary to establish distance functions to evaluate the contiguity between words, the precise number of words interposed between the searched for items or the membership of words in the same superunit (e.g. same section, same tragedy, etc.). Common systems for text retrieval use the position (i.e. the progressive number) of each word inside the superunit to accomplish this task and both words and positions are indexed for efficiency reasons.[1] Corpora enriched with variants and conjectures are challenged by the computation of word distance, in particular if insertions and deletions have been performed.

The solution adopted in the present work aims to examine the following issues: a) maintenance: repertory reference editions and variant readings are mapped onto the main reference edition without altering the structure of the annotated corpus used to produce it; b) ordering simplicity: positions are expressed by decimal numbers to easily reorder textual sequences in the presence of insertions and deletions; c) efficiency: insertions and deletions are associated with offsets, and can be used to extend text retrieval systems without significant decrease in performance.

3.10.1 Context of the variant reading and position of the items

In critical editions, the context of variants registered in critical apparatus is the text established by the editor. In the present work the scenario is more complex, because there is a main reference edition (Murray) and other reference editions (Wecklein and West) for some repertories. Furthermore, as seen above, reperto-

[1] I am grateful to Luigi Tessarolo for a draft about the technical details fo the search engine used by the *Musisque Deoque Project*.

ries often register conjectures based on previous conjectures. In these cases the ultimate context of the reading must be reconstructed step by step along a chain of edits. An example from the repertory of Wecklein should illustrate the problem, even if it is an exceptional case, selected for its complexity that, at present, the automatic parser is not yet able to manage.

119sqq. ὀᾶ ὀᾶ (sic etiam 125), *Περσικοῦ* (*βαρβάρου* malit Schiller) *στενάγματος τοῦδε μὴ πόλις πύθηται* (vel potius *μέλος* vel *βοὰν τίθηται*) olim, postea ὀᾶ ὀᾶ *Περσικοῦ στρατεύματος, τούσδε μὴ στόνους πύθηται* Weil.

Wecklein's text (ὀᾶ| *Περσικοῦ στρατεύματος*| *τοῦδε μὴ πόλις πύθη-*|*ται*) provides the context for the two main conjectures of Weil: a) ὀᾶ ὀᾶ, *Περσικοῦ στενάγματος τοῦδε μὴ πόλις πύθηται* and b) ὀᾶ ὀᾶ *Περσικοῦ στρατεύματος, τούσδε μὴ στόνους πύθηται*. But the first conjecture of Weil constitutes the context for the conjecture of Schiller, that should be read: c) ὀᾶ ὀᾶ, *βαρβάρου στενάγματος τοῦδε μὴ πόλις πύθηται* and for his own minor conjectures: d) ὀᾶ ὀᾶ, *Περσικοῦ στενάγματος τοῦδε μὴ μέλος τίθηται* and e) ὀᾶ ὀᾶ, *Περσικοῦ στενάγματος τοῦδε μὴ βοὰν τίθηται*, that is expressed in the context of d). Considering that, fortunately, the cascading contexts are very rare[1], the best solution is to reconstruct the minimal variant context for each conjecture, ignoring the left and right placeholders, as in Tab. 3.8.

Position	427	427.1	428	429	430	431	432	433
Reference ed.	ὀᾶ	–	*Περσικοῦ*	*στρατεύματος*	*τοῦδε*	*μὴ*	*πόλις*	*πύθηται*
Weil[1]		ὀᾶ,	*Περσικοῦ*	*στενάγματος*				
Weil[2]		ὀᾶ	*Περσικοῦ*	*στρατεύματος,*	*τούσδε*			
Schiller		ὀᾶ,	*βαρβάρου*	*στενάγματος*				
Weil[1]		ὀᾶ,	*Περσικοῦ*	*στενάγματος*	*τοῦδε*	*μὴ*	*μέλος*	*τίθηται*
Weil[1]		ὀᾶ,	*Περσικοῦ*	*στενάγματος*	*τοῦδε*	*μὴ*	*βοὰν*	*τίθηται*

Table 3.8: Conjectures in the context of other conjectures

[1]In Wecklein's repertory on *Persae* they are only 25 out of 1077 verses and c. 2000 conjectures.

Positions are determined according to the alignment: substitutions and deletions receive the same positional number as the aligned items in the reference edition. In case of insertion, suitable decimal numbers are generated.

3.10.2 Offset and unique identifiers for variant readings

Because of insertions and deletions, positional numbers can only be used to order items in the context, not to compute word distances. Each variant reading, constituted by one or more items, is associated with a unique identifier and to a triplet of integer numbers: left and right boundaries and global offset produced by the reading. Boundaries are respectively the first integer positional number of the main reference edition before the variant reading and the first integer positional number after it. The global offset is the difference between the sum of insertions and the sum of deletions or, expressed in another way, the difference between the number of words contained in the variant reading and the number of words contained in the reference edition. Each item of the variant (if it is not a deletion) is associated with the offset from the left bound, as shown in Tab. 3.9.

main ref. ed. (Murr.)	γᾶς	ἀπ᾽	Ἀσίδος	ἦλθεν᾽	–	–	–	αἰαῖ	δᾴαν	Ἑλλάδα	χώραν
ref. ed. (Weck.)				ἦλθ᾽	–	–	ἐπ᾽	αἶαν			
M. Schmidt				ἐλθεῖν	βαιὰν	Ἑλλάδ᾽	ἐπ᾽	αἶαν	–	–	
position	1305	1306	1307	1308	1308.01	1308.02	1308.1	1309	1310	1311	1312
offset				1	2	3	4	5	–	–	
global offset	5-(1312-1307-1)=1										

Table 3.9: Offset (*Pers.* 273-274)

3.10.3 Computation of word distance

Given the position p associated with any word, its offsets, os, the left and right bounds of the variant under examination, l and r, and the global offset of the variant, g, the computation of the rescaled position of p is determined by the formula:

$$rp = \begin{cases} \text{p} & \text{if} \quad \text{p} \leq \text{l} \\ \text{l+os} & \text{if} \quad \text{p} > \text{l and p} < \text{r} \\ \text{p+g} & \text{if} \quad \text{p} \geq \text{r} \end{cases}$$

In fact, if p≤l, the word occurs before the variant and its position is the same as the position in the main reference edition. If the word occurs between the boundaries (p>l and p<r), the position is determined by the sum of the left boundary and the offset of the word. If p≥r, the word occurs after the variant and it is necessary to add its global offset. Finally, computation of word distance with a single contiguous variant in the context of the main reference edition is easily reduced to rp_2-rp_1, an operation that can be performed by systems for text retrieval with minimal computational costs. In the example seen above, *ἤλθετ' αἰαῖ* are contiguous in the main reference edition. The word distance for the related aligned words *ἐλθεῖν* and *αἶαν* in Schmidt's conjecture *ἐλθεῖν βαιὰν Ἑλλάδ' ἐπ' αἶαν*, with p_1=1308, p_2=1309, os_1=1, os_2=5, l=1307, r=1312, g=1, is given by rp_2-rp_1=4, where rp_1=1307+1=1308, rp_2=1307+5=1312, and rp_2-rp_1=1312-1308=4.

3.10.4 Computation of word distance for discontinuous variants

A discontinuous variant is usually signaled in apparatus and repertories by the presence of dots, for instance: 43sq. *οἵ τ'... κατέχουσιν ἔθνος, Μιτραγαθὴς* Schuetz.

	v. 43					v. 44				
ref. ed.	*ὄχλος, οἵτ'* –			*ἐπίπαν ἠπειρογενὲς*		*κατέχουσιν ἔθνος, τοὺς Μητρογαθὴς*| ...				
Schuetz		*οἵ*	*τ'*				–	*Μιτραγαθὴς*		
position	172	173	173.1	174	175	176	177	178	179	180
offset		1	2				–	1		
global offset	2-(174-172-1)=1					1-(180-177-1)=-1				

Table 3.10: Discontinuous conjecture (*Pers.* 43-44)

The parts of a discontinuous variant are referenced by the same identifier, but they are associated with different triplets (in the example above, the first

part is associated with the triplet [172, 174, 1] and the second part with the triplet [177, 180, -1]). The evaluation of the word distance must take in account the accumulated offsets produced by any part interposed between the positions under examination. For example, the word distance between ὄχλος and Μιτραγαθὴς is 9 and not 8 (according to the previous formula) because the first part of the conjecture, interposed between ὄχλος and Μιτραγαθὴς, provides a global offset of 1. Even in this case the computational cost for text retrieval is minimal, because the discontinuous variant is reconstructed by the unique identifier associated with its parts, that are ordered by the left boundary (discontinuous variants, by definition, never overlap). Global offsets are accumulated according to the relative position of words under examination and boundaries of the parts of the discontinuous variant.

3.11 Conclusion

In this chapter we have illustrated methods to map variants and conjectures onto reference editions. A preliminary study, facilitated by manual annotation of textual samples, has been performed in order to identify recurrent structures of critical apparatus and repertories of conjectures.

The application of suitable heuristics allows the classification of chunks of information related to variant readings, such as Greek sequence, textual operations (addition, deletion, substitution or transposition of text) and scholar that suggested the conjecture.

Alignment algorithms, mutuated by bioinformatics in the domain of computational linguistics, have been applied to digitized apparatus and repertories, in order to map, word by word, variant readings onto reference editions. Performance is promising.

An annotation system for textual positions that facilitates text retrieval operations has been illustrated, demonstrating how word distance computation can be easily and efficiently performed also for variant readings in their contexts.

4

Semantic Spaces of Ancient Greek

4.1 Introduction

This chapter studies the semantic spaces based on the TLG ancient Greek corpus[1] and some relevant subcorpora.

After an overview on the semantic space structure in section 4.2, the method to build the vector spaces that express the semantic relations is illustrated in section 4.3 and semantic distance, which allows the evaluation of the similarity between terms, is discussed in section 4.4.

A corpus based on classical literary texts has some peculiarities that are the topic of the following three sections. First, TLG corpus is a small-medium size corpus (less than 500 megabytes), despite the fact that methods used in this work are usually applied to very large corpora (more than one terabyte). The problem if there are categories of low frequency words that can provide relevant semantic associations is addressed in section 4.5. Second, a corpus distributed along 23 centuries promotes the study of the semantic change, due to cultural, scientific and spiritual mutations. In section 4.6 some key-words are tested in different temporal segments, in order to observe consistent changes of meaning.

[1]The ancient Greek corpus, named *Thesaurus Linguae Graecae* (http://www.tlg.uci.edu), is a collection of texts from Homer (eighth century B.C.) to the fall of Costantinople (fifteenth century A.D.)

Third, no native speakers have the competence to evaluate the multiple senses of ancient words, which can be reconstructed only in relation to the context and by conjecture. Polysemy is the topic of section 4.7, in which we discuss a method to study polysemic words in contexts belonging to different domains.

Semantic relations explored by word clustering are the topic of section 4.8 and in particular of subsection 4.8.1, devoted to antonymy and subsection 4.8.2, related to taxonomies. Some examples are discussed in detail for both the categories.

Finally, a couple of words, one concrete and one abstract, are used as keywords to highlight the main differences among relevant partitions (subcorpora) of the *Thesaurus Linguae Graecae*. Ancient Greek literature is structured in genres consistent both for lexical and semantic choices. Section 4.9 is divided in three subsections, devoted to Homeric poems (4.9.1), Tragedy (4.9.2) and Philosophy (4.9.3).

Section 4.10 summarizes the main achievments.

4.2 Semantic Space Structure

A Semantic Space is a computational model based on the word distribution of a corpus, in order to represent semantic similarity by spatial proximity (Lenci, 2008; Sahlgren, 2006). The multidimensional space where the terms are placed is typically built using the data of a co-occurrence table (Tab. 4.1), which crosses the m lemmatized word of the corpus with the n most frequent lemmatized terms that appear in the context window of the focused word, which has a fixed width, w. The value of each cell expresses the number of times the term in the row co-occurs in the context window with the term in the column. For instance, the cell that crosses ἁβρόγοος (wailing womanishly) with ἁβροχίτων (with soft coverings) has value 1 because the two words co-occur only once in the narrow window of twenty words (respectively at the verses Aesch. *Pers.* 541 and 543). On the contrary, ἁβροδίαιτος (living delicately) never appears close to the other terms in the narrow window. The width of the context window, w, can be enlarged or narrowed. Common choices are ten, twenty or one hundred words on the left and on the right side of the focused word (see Sahlgren, 2006, p. 68). A narrow window emphasize the syntagmatic relations, whereas a wide window emphasize

the paradigmatic relations. The hyperspace constructed with the data of the co-occurence table has n dimensions and it contains m points. The cells contain the coordinates of each point in the multidimensional space. In this way, close points in the hyperspace represent terms that belong to similar contexts. In fact, the semantic space theory depends by the distributional hypothesis: "Words that are similar in meaning occur in similar contexts" (Rubenstein and Goodenough, 1965).

	...	ἁβρόγοος	ἁβροδίαιτος	...	ἁβροχίτων	...
...	...	...	...	...	...	...
ἁβρόγοος	...	0	0	...	1	...
ἁβροδίαιτος	...	0	0	...	0	...
...	...	...	...	...	...	...
ἁβροχίτων	...	1	0	...	0	...
...	...	...	...	...	...	...

Table 4.1: Co-occurrence Table

The application of the Singular Value Decomposition (Wall *et al.*, 2003) allows the reduction of the original semantic space dimensions, reducing the data noise. Due to the dimensional reduction, terms that never co-occur in the same context window but that occur with terms that are in turn semantically similar become close in the new lower dimensional space. According to Grossman and Frieder (2004), "The key to similarity is not that two terms happen to occur in the same document [or window]: it is that two terms appear in the same *context*, - that is they have very similar neighboring terms". Here the term *context* is used in the broad sense either of first order or higher order co-occurrence. In fact, if AB co-occur in the context x, BC in the context y and CD in the context z (first order co-occurrences), AC and BD are co-occurrences of the second order and AD is a co-occurrence of the third order (Kontostathis and Pottenger, 2003). The Singular Value Decomposition makes allowance for capturing these relations. The new dimensions of the reduced hyperspace do not correspond anymore to single words of the context window: the higher order co-occurrences emerge by the grouping of information spreaded on the original dimensions.

Compared to other computational models used to determine the semantic similarity among terms (see for instance Fellbaum, 1998), in this model "the space is constructed with no human intervention, and with no a priori knowledge or constraints about meaning similarities" (Sahlgren, 2006, p. 21).

4.3 Method

Methods applied in this study are used by linguists on wide corpora of non literary documents, in English or other modern languages, reaching many millions of processed words, in particular if the corpus is based on documents retrieved on Internet. As it will be explained below, the number of occurrences influence dramatically the individuation of relevant semantic relations.

Very large corpora in English can be analysed without lemmatization. On the contrary, given the morphological complexity of ancient Greek and the relatively small size of the corpus contained in the TLG (approximately 76 million words, but c. 4 million words have been excluded in the present study, because they belong to fragmentary works), in our case the lemmatization is necessary, because it reduces sparseness of data. The inflected forms have been lemmatized by *Morpheus* (Crane, 1991), that is currently the most accurate lemmatizer for ancient Greek. A comparison of the automatic lemmatization on the TLG texts with the manual lemmatization made by the C.I.P.L. of Liége (Rigo, 1999) demonstrates that the precision can reach 80%. In fact, 38,474 inflected words out of 47,283 (81%) have been correctly processed, by assigning to each form a single lemma or the first of a list of suggestions. The recall, considering the entire list of suggestions, is close to 93%, but in this study only the first item of the list has been used. Even if the precision is not very high, it is sufficient for this study, because many wrong head words are etymologically related to the inflected form analyzed, and then the semantic relation is correctly preserved.

Instead, the main problem is due to compounds, because they are lemmatized without any kind of analysis of the components, by loosing the relation among them, which determines the semantic relation. For instance, the occurrences of ἁβρο- (with the idea of splendour, luxury, sweetness, delicacy, charm) are not recorded in *Pers.* 1073 ἁβροβάτης (delicately stepping), *Pers.* 541 ἁβρόγοος

(wailing womanishly), *Pers.* 135 ἁβροπενθής (delicately suffering) and *Pers.* 543 ἁβροχίτων (with soft coverings).

In this study four semantic spaces have been generated, based on the entire collection of the TLG, the Homeric Poems, the corpus of the Greek tragedy with related *scholia* and the corpus of the philosophers. The selection of the texts was based on the literary genres indicated in Squitier (1990). Words have been lemmatized by the aforementioned procedure.

The collection of words in textual order have been processed by Infomap[1], a specific application for the study of semantic spaces, which generate the table of the co-occurrences, performs the Singular Value Decomposition and compute the coordinates of the terms in the resulting semantic space. Parameters have been setted to 120,000 rows, 2,000 columns, 300 singular values, 300 singular value decomposition iterations, 100 left context words, 100 right context words.

4.4 Semantic Distance

For each term it is possible to evaluate the list of the closest points representing terms in the semantic space and thus, according to the distributional hypothesis, semantically most similar to the focused word. For instance, given the term κυβερνήτης, skipper, and the TLG semantic space, the top ten associations, ranked by similarity, are: ναύτης (seeman, sailor) 0.68, κυβερνάω (steer) 0.61, σκάφος (hull of a ship) 0.57, πλοῖον (floating vessel) 0.55, πρύμνα (stern, poop) 0.53, πηδάλιον (steering-paddle) 0.52, ἄγκυρα (anchor) 0.51, κῦμα (wave) 0.50, κυβερνητικός (good at steering) 0.50, πλωτήρ (navigable) 0.50. Cosine similarity, calculated by Infomap, is the dot product of the normalized frequency vectors. Given one or more words, Infomap's similarity tool, called `associate`, provides a list of associated terms, scored by cosine similarity.

Semantic relations are distributed both on the syntagmatic and the paradigmatic axis: synonymy, such as θάλασσα (sea) – πέλαγος (sea, but especially open sea), hypernymy, hyponymy, co-hyponymy, such as ὄρνις (bird) – ἱέραξ (hawk, falcon) – ἀετός (eagle), holonimy, meronymy, co-meronymy, such as ναῦς (ship) – πρύμνα (stern, poop) – ἱστίον (sail), antonymy, such as μέγας (big) – μικρός

[1] http://infomap-nlp.sourceforge.net

(small) or θάλασσα (sea) – ἤπειρος (*terra firma*, land) and membership to the same frame, according to the Fillmore (1997) definition, such as ἵππος (horse) – ἅρμα (chariot) or θάλασσα (sea) – ναύτης (sailor man).

4.5 Semantic Space and word frequency

Word frequency has a strong influence on the semantic associations. In fact, the focused word must appear in a suitable number of contexts. The experimental results emerged in this study suggest that it is necessary to have at least four occurrences of the term, otherwise the system finds inconsistent associations. However, due to the Zipf's law (Evert and Baroni, 2007; Zipf, 1949), in a relatively small corpus of text, such as the TLG corpus, more than half of the words appear less than four times. As a consequence, the results are satisfying only for a limited part of the corpus. For instance, considering a high frequency word such as λίθος, stone, (14,830 occurrences), all the top ten associations are consistent: the hyponym ὑάκινθος, aquamarine and σμάραγδος, emerald; the co-hyponyms σίδηρον, iron, and ξύλον, wood, with the adjective χάλκεος, of bronze, derived by a co-hyponym; the meronyms τοῖχος, wall, ὄροφος, roof[1], κίων, pillar, and ἄγαλμα, statue.

As said above, the semantic associations related to low frequency terms are unreliable. In fact, these terms (especially words with a single occurrence) provide associations with other low frequency terms in the same contexts, without semantic consistency. For example, ἀργυροφεγγής, silver-shining, occurs four times in the TLG corpus. Even the first four associations are unrelated: ἀρχοντιάω, wish to be ruler, μεθυσφαλέω, to be reeling-drunk, ὁμόφθογγος, sounding together, περισσόνοος, eminent for understanding.

But in particular conditions the associations can be effective. For example when the term occurs once in the original context and a few times in a lexicographic or scholiastic work. For instance, the term ἁβρόγοος, wailing womanishly and delicately, occurs only once in Tragedy (*Pers.* 541ἁβρόγοοι), but three times in *scholia*. In the semantic space of *scholia*, the term produces the following relevant top associations: ὑγιοζυγία, healthy union, ἀρτιζυγία, recent union (in

[1] ὄροφος means reed used for thatching houses and, metonymically, roof (ὀροφή).

context, cause of wailing), κατερείκω, tear,ἁβροχίτων, with soft coverings, πρωτόμορος, dying or dead first, κατασχίζω, tear, ἁπαλόχροος, soft skinned. Four terms are low frequency words that appear in the same context window: *Pers.* 542 ἀρτιζυγίαν (ἀρτιζυγία), 538 κατερεικόμεναι (κατερείκω), 568 πρωτομόροιο (πρωτόμορος) and 543 ἁβροχίτωνας (ἁβροχίτων). Two terms are *glossae*: ὑγιοζυγία, explains ἀρτιζυγία and κατασχίζω explains κατερείκω. Finally, the first component of the compound ἁπαλόχροος, ἁπαλός, delicate, is a synonym of ἁβρός, the first component of ἁβρόγοος.

4.6 Observing diachronical changes of meaning

The diachronical nature of the ancient Greek corpus shows its influence on the associations, especially for the abstract terms. TLG corpus has been divided in two subcorpora: the first one contains B.C. texts and the second one A.D. texts. Two high frequency terms have been arbitrarily selected: θάλασσα, sea, and θάνατος, death, in order to study a concrete term, such as *sea*, and an abstract term, such as *death*.

θάλασσα	sea	θάλασσα	sea
πέλαγος	high sea	πέλαγος	high sea
ἤπειρος	land	ἤπειρος	land
ποταμός	river	λίμνη	marshy lake
νῆσος	island	κῦμα	wake
πότιμος	drinkable	ποταμός	river
λίμνη	marshy lake	πότιμος	drinkable
ἁλμυρός	salt	ἐπικλύζω	to overflow
πόντος	open sea	ἁλμυρός	salt
ὄρος	mountain	ἀπειρόω	to multiply to infinity
κῦμα	wave	Ὠκεανός	Ocean

Table 4.2: θάλασσα: B.C. and A.D. associations

As Tab. 4.2 shows, θάλασσα generates associations with geographic terms or related adjectives and verbs, which have a very similar ranking both in the first and in the second subcorpus. On the contrary, as Tab. 4.3 shows, θάνατος

provides two different lists that demonstrate the change of cultural and religious connotations. Indeed, the second list (A.D.) contains terms with a strong Christian valence: ᾅδης, hell (in the vulgar era: the Christian hell); νεκρόω, to mortify; ἀνάστασις, rising from the dead; ἀθανασία, immortality; Θάνατος, the personification of death and, finally, the most explicit Christian term: συσταυρόομαι, to be crucified together with.

θάνατος	death	θάνατος	death
ἀποθνήσκω	to die	ἀποθνήσκω	to die
ἀποκτείνω	to kill	νεκρός	corpse
τιμωρία	vengeance	ᾅδης	hell
φόνος	murderess	νεκρόω	to mortify
φεύγω	to flee	ἀνάστασις	rising from the dead
θνήσκω	to die	ἀθανασία	immortality
συμφορά	misfortune	θνήσκω	to die
ἀνδροφόνος	homicide	Θάνατος	Death
τραῦμα	wound	συσταυρόομαι	to be crucified together with
τελευτάω	to accomplish	ζωή	life

Table 4.3: θάνατος: B.C. and A.D. associations

4.7 Polysemy

In case of terms with literal and figurative meaning, the most commonly encountered in the corpus use shadows the secondary use. For example, ναύτης, sailor, occurs 1,161 times in the TLG, mainly in its literal sense. In consequence, the top associations are: κυβερνήτης, skipper, πλοῖον, ship, σκάφος, hull of a ship, ναῦς, ship, πρύμνα, poop, κῦμα, wave, ἄγκυρα, anchor, πέλαγος, open sea, ἔμπορος, ship passenger, ἐμπλέω, to sail in. However, in the semantic space built on the TLG subcorpus of the philosophical works, ναύτης is used only 79 times (17 times in Plato's works), mainly as a metaphor in the domain of the political activity. In this case among the top associations we find: ἄρχω, to lead, to govern, ὑπήκοος, obeying ally, στρατηγέω, to be general, μονοπώλιον, right of monopoly, παραβοηθέω, come to aid allies.

While the entire corpus shadows the threads of meaning for the polysemic terms, the subcorpora are able to isolate the specific domains of meaning. For example, the polysemic word πούς, foot, in the TLG semantic space produces the list ποδόω, to be furnished with feet, ποδίς, shoe, δάκτυλος, finger, σκοῦτα, shield, σκέλος, leg, παρασκελής, with unequal legs, ὄργυια, the length of the outstrecthed arms and βησαλικόν, brick-work, missing different domains of meaning. On the contrary, in the semantic space of medicine the same term produces the list of associations: ποδίς, shoe, πτέρνα, heel-bone, σκέλος, leg, χείρ, hand, ποδόω, to be furnished with feet, κνήμη, part between knee and ankle, γόνυ, knee, πεδάω, bind, δάκτυλος, finger and κάμπτω, bend, most of them related to the domain of the body parts. In the semantic space of the *scholia* the same term produces the list: ἴαμβος, iambus, τροχαικός, trochaic, δάκτυλος, dactyl, ὑπερκατάληκτος, hypercatalectic, τροχαῖος, trochee, ἀδιάφορος, metrically indifferent, βραχυκατάληκτος, ending in a short syllable, ἀντίσπαστος, antispast, and παιών, paean, all related to the domain of the metrics.

Infomap's similarity tool can receive as parameter a single term or lists of terms. In case of polysemic words, one or more terms can determine a specific domain for the investigation, even if the semantic space is built on the entire TLG, which is not domain specific. For example, ποῦς, foot, and χείρ, hand, provide a series of associations related to the parts of body, whereas ποῦς, foot, and ἴαμβος, iambus, provide associations in the domain of metrical analysis.

4.8 Clustering

The dimension reduction of the original semantic space allows the representation of the distances among terms in a bidimensional chart. Dimension reduction is performed in two steps: in the first phase Infomap reduces the original dimensions, more than one thousand, to three hundred by singular value decomposition. In the second phase, multidimensional scaling is performed by an R script.

In order to represent data in a bidimensional graph, the reduced dimensions returned by Infomap (in this work, three hundred) are further reduced to two dimensions by the multidimensional scale function (Baayen, 2008, p. 146-148) and a hard partitioning is obtained by the computation of the centroids of each group (see Feng and Manmathan, 2006), by the k-means method. In fact, k-

means aims to partition n items into k clusters in which each item belongs to the cluster with the nearest mean.

This method emphasizes in particular antonomies and taxonomies. Members of antonymic couples, sharing very similar contexts, tend to be closely clustered with each other and clearly separated by different couples. Members of taxonomies tend to be disposed according to the taxonomic hierarchy, forming clusters for each category, possibly with meaningful sub-clusters.

4.8.1 Antonymy

Antonyms establish both syntagmatic and paradigmatic relations. On the syntagmatic axis, couples of antonyms frequently appear in the same contexts. Especially in philosophical texts and metalinguistic works, such as grammars, lexica, commentaries and scholia, couples of antonyms are object of study. Remarks on philosophical texts are presented in section 4.9.3.

On the paradigmatic axis, antonyms appear in similar contexts, because they express opposite properties of the same terms, in particular when the couple of antonyms is constituted by adjectives that can occur with a restricted number of names.

For these reasons, the components of antonymic couples are expected to be very close in the reduced semantic space. Tight relations are observed for domain-specific terms, such as ἔμψυχος (animate) and ἄψυχος (inanimate), commonly used in philosophical and medical texts, even if the metaphorical sense can be related to a limited number of other domains, such as rhetorics (e.g. vivid discourse or lifeless style). Loose relations are observed for couples of antonyms that can be used in any context, such as μέγας and μικρός. For example, the similarity score of ἄψυχος compared to ἔμψυχος is 0.8 and the similarity score of μικρός is 0.6. Similarity scores are provided by Infomap tools.

Finally, it is worth noting that polysemic words can have different antonyms, related to the different senses (see Fellbaum, 1998, p. 51). For example, βαρύς, grave, has ὀξύς, acute, as an antonym and βαρύς, heavy, has κοῦφος, light, as an antonym.

The example presented in Fig. 4.1 is based on nine couples of antonyms arbitrarily selected and listed in Tab. 4.4. All the words are high frequency

ἀνήρ	man	γυνή	woman
ἄρσην	masculine	θῆλυς	feminine
γλυκύς	sweet	πικρός	bitter
ἔμψυχος	animate	ἄψυχος	inanimate
ἥμερος	day	νύξ	night
θερμός	hot	ψυχρός	cold
λογικός	rational	ἄλογος	irrational
ὀξύς	acute	βαρύς	grave
φάος	light	σκότος	darkness

Table 4.4: Antonyms

terms: the most frequent is ἀνήρ (man) with 5,1845 occurrences and the least frequent is ἄψυχος (inanimate) with 2,712 occurrences. As shown in the figure, all the couples are correctly grouped using the k-means clustering method.

The disposition of the couples suggests that these couples obey to a hierarchical structure. The dendrogram in Fig. 4.2, which is just a different plot of the same data (based on two dimensions as well), better evidentiates this structure. The dendrogram plots the result of the hierarchical cluster analysis on the set of euclidean dissimilarities computed on the bidimensional scaling of the original data matrix.

A simple interpretation of these data is that ἄρσην, male, and θῆλυς, female, are grouped with ἀνήρ, man, and γυνή, woman, as gender determinations of the hidden hypernym ἄνθρωπος, human being. λογικός, rational, and ἄλογος, irrational, are entailed by ἔμψυχος, the positve term of the antonymic couple ἔμψυχος, animate, and ἄψυχος, inanimate, whereas λογικός, rational, is the specific difference, according to Aristotle, of ἄνθρωπος, human being. On the second half of the graph, φάος, light, and σκότος, darkness, are grouped with ἥμερος, day, and νύξ, night, due to the strong association among the terms. γλυκύς, sweet, and πικρός, bitter, θερμός, hot, and ψυχρός, cold, and eventually ὀξύς, acute, and βαρύς, grave, are grouped together as perceptual determinations (tasteful, tactile and auditive). Finally, this set is associated with the previous group of words related to visual perceptions.

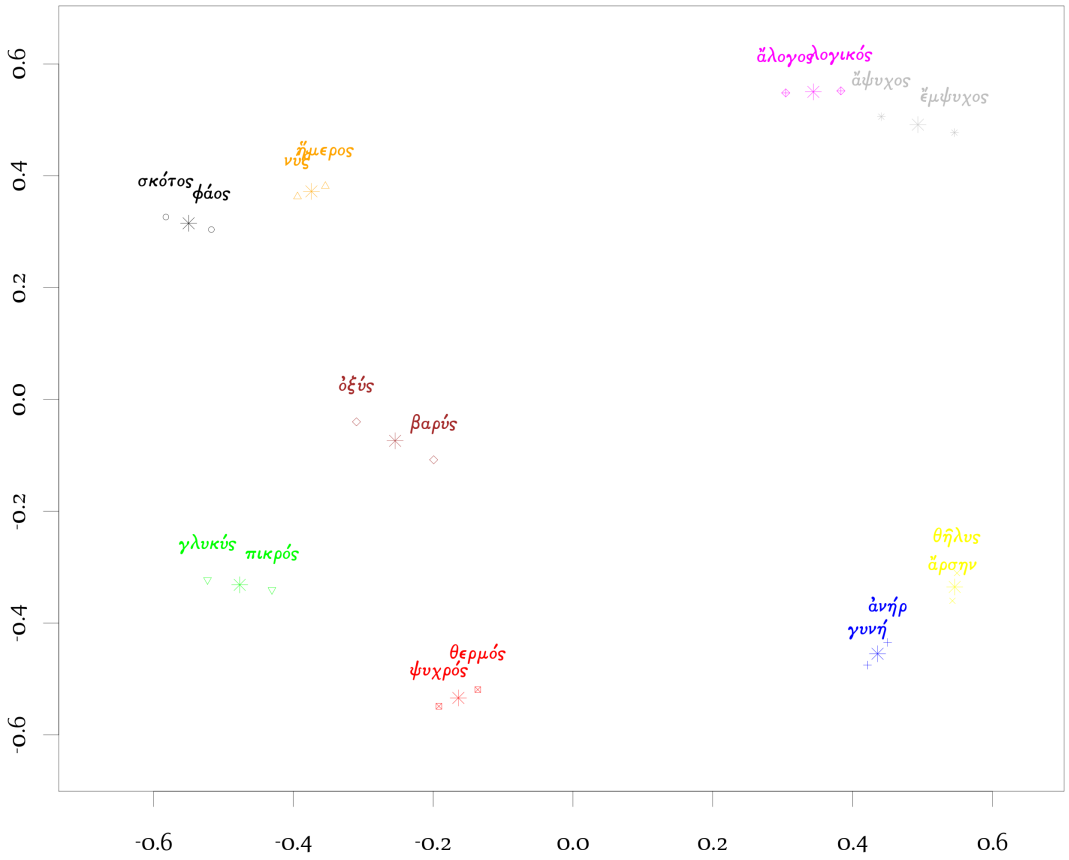

Figure 4.1: Clusters of antonyms

4.8.2 Taxonomies

Even if the TLG corpus size is not large, relevant taxonomies can be easily identified and critical items, apparently misclassified, sometimes can be explained with secondary senses of the terms, as it will be exemplified below.

The results are satisfactory for both concrete and abstract terms, if they belong to well-established paradigms, such as body parts, components of ships or parts of buildings in the case of concrete terms and family relations, emotions or virtues in the case of abstract terms.

In Fig. 4.3, based on data of Tab. 4.5, it is shown how the terms of the animal taxonomy are disposed in the bidimensional graph. The traditional partition in earth (*χερσαῖος*, terrestrial or *πεζός*, on foot), water (*ἔνυδρος*, aquatic) and air

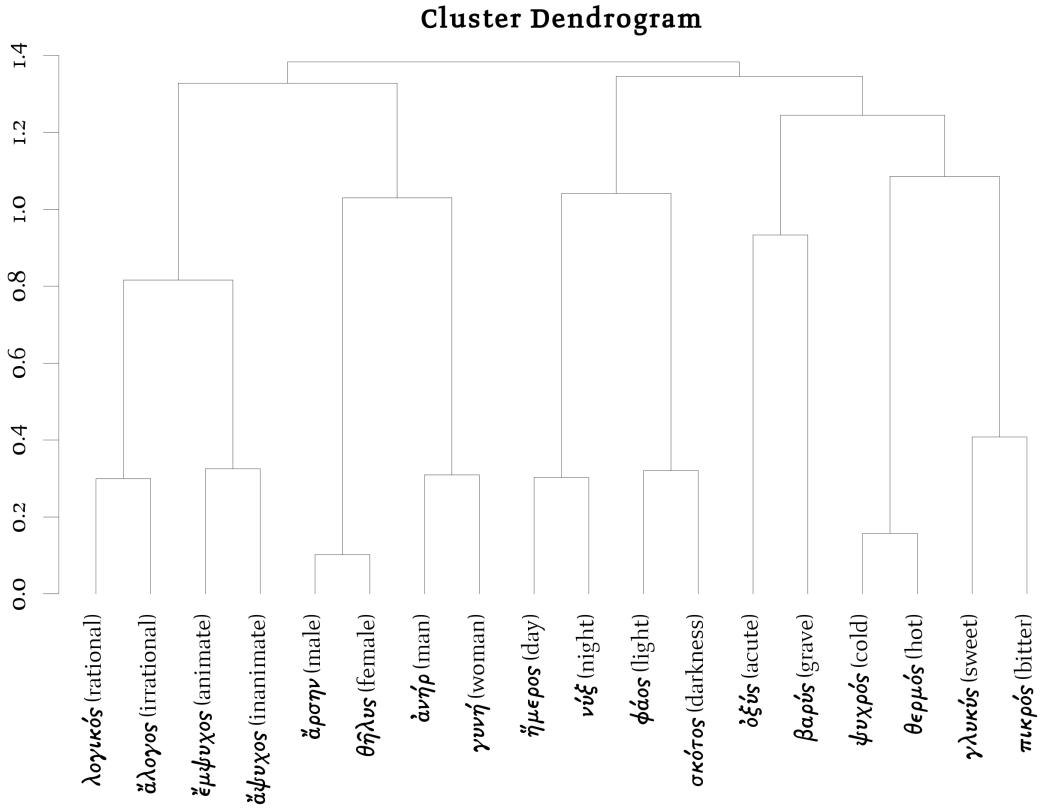

Figure 4.2: Hierarchy of the antonymies

(πτηνός, flier) animals is preserved, even if some cases need to be discussed. The main meaning of λαγῶς is hare, a terrestrial animal, but the item is classified as a bird. The misclassification is only apparent because λαγῶς, as indicated in the Liddell-Scott dictionary, is also "a bird with rough feathered feet, mentioned with the swallow". Furthermore, it is also "a kind of sea-slug, Lepus marinus". In a similar way, στρουθός, sparrow, is a bird but also "a flat fish, flounder, Pleuronectes flesus". As shown in the chart, στρουθός is in the middle between birds and fishes. The hypernyms θήρ, beast, ὄρνις, bird, and ἰχθύς, fish, are in the center of the graph. The flying insect τέττυξ, cicada, is classified as a wing animal, like birds.

θήρ	beast	ὄρνις	bird	ἰχθύς	fish
κριός	ram	γύψ	vulture	τρίγλη	red mullet
κύων	dog	χελιδών	swallow	σκάρος	parrot-wrasse
βοῦς	ox	ἀετός	eagle	θύννος	tuna
ὗς	boar	πέρδιξ	partridge	λάβραξ	bass
λέων	lion	ἀλεκτρυών	cock	κέφαλος	mullet
ἔλαφος	deer	κύκνος	swan	ἔγχελυς	eel
ταῦρος	bull	τέττιξ	cicada		
ὄνος	donkey	ἱέραξ	hawk		
ἵππος	horse	στρουθός	sparrow		
ἡμίονος	mule				
λαγῶς	hare				

Table 4.5: Animals

In the previous example, the items have been arbitrarily selected. The addition, subtraction or substitution of words can modify the reciprocal relations of the words, producing missclassifications. In the current example, modification of items can produce, in particular, overlappings between beasts and birds, for frequent sharing of features (e.g. swiftness or smartness), whereas the class of the fishes tends to avoid overlappings.

In order to reduce the arbitrariness in list creation, seed words for each expected cluster can be used. In this way, manual control is delegated to the manual rezooming into certain areas of the semantic space, in order to find few seed words able to generate well separated clusters.

Infomap's similarity tool finds semantic similarities related to single words but also to sets of words. As discussed in section 4.7, a couple of words is enough to generate a list of consistent associations. If the couples of words used as seeds are sufficiently contrastive, the lists generated by Infomap's similarity tool are expected to be clustered without overlappings. The rule of thumb to choose valuable seed words is to plot lists of arbitrary words, as seen above, and then to select a couple of items for each cluster that are on the margins of the cluster,

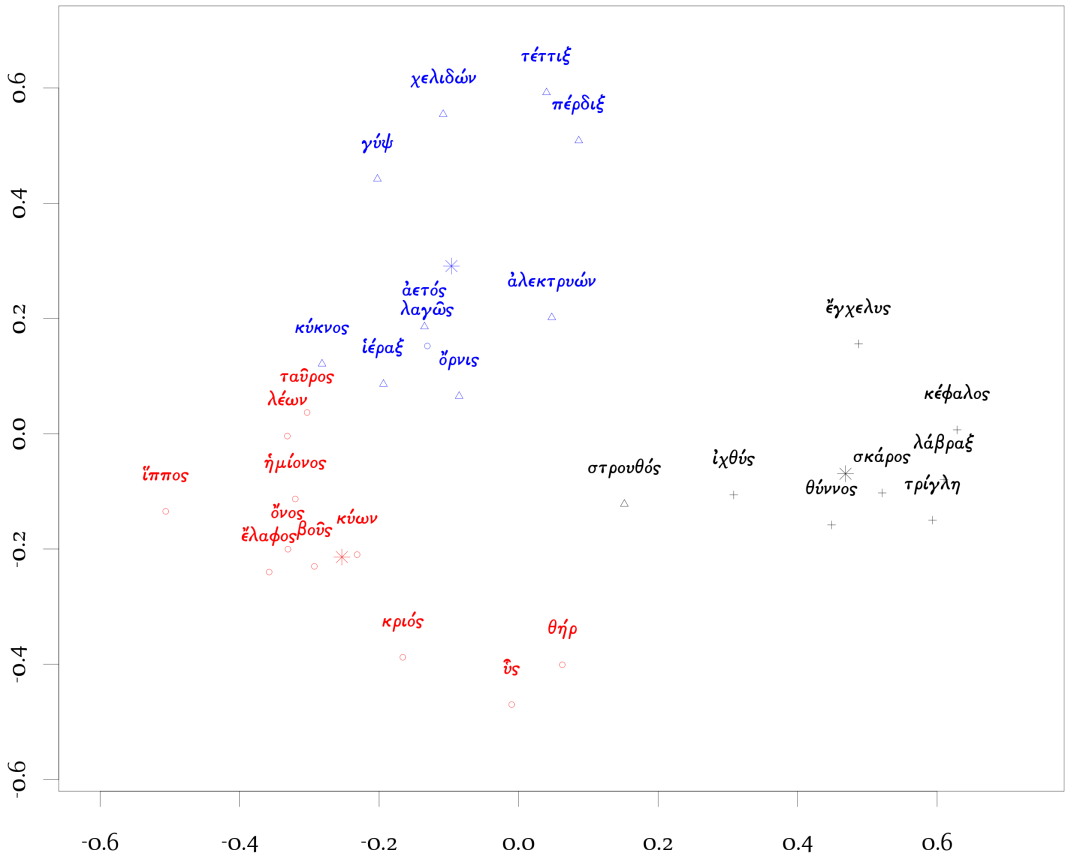

Figure 4.3: Clusters of animals

paying attention that at least one of the items is far from the overlapping area among clusters.

An example of the application of this method is shown in Fig. 4.4. In this case, the couples of words used as seed have been selected from the previous plot, generated with arbitrary items (Fig. 4.3). *ἵππος*, horse, and *κύων*, dog, have been selected for earth animals, *ἔγχελυς*, eel, and *θύννος*, tuna, for water animals and, finally, *ἱέραξ*, hawk, and *πέρδιξ*, partridge, for air animals. The rusults provided by Infomap's similarity tool for each couple of seed words are shown in Tab. 4.6, ordered by similarity.

The first column provides a series of cohyponyms of the seeds. *θήρα*, hunting of wild beasts, and *κυνηγός*, hound-leader, huntsman, belong to the frame of the

κύων	dog	ἱέραξ	hawk	ἔγχυλος	eel
λέων	lion	πέρδιξ	partridge	θύννος	tuna
λύκος	wolf	ὄρνις	bird	κωβιός	gudgeon
ἔλαφος	deer	χελιδών	swallow	γαλεός	dog-fish
πάρδαλις	leopard	ἀλώπηξ	fox	ἰχθύδιον	little fish
ἀλώπηξ	fox	πτηνός	winged animal	εὔχυλος	succulent
θήρα	hunting	ἀλεκτρυών	cock	κέφαλος	mullet
ἵππος	horse	ἀετός	eagle	θυννίς	female tuna
σκύμνος	(lion's) whelp	λαγῶς	hare	τίφη	a kind of boat
κυνηγός	huntsman	ἰχθύδιον	little fish	κίχλη	sea-fish

Table 4.6: List of animals generated by seeds

hunt. The word shares also some inflected forms with the hypernym θήρ, beast: θῆρ' and θηρῶν. *Morpheus*, the morphological analyzer, assigns both the forms to θήρ, losing for that occurrencies the association to θήρα.

The second column provides not only cohyponyms but also the hypernym ὄρνις, bird, and its hypernym πτηνός, winged animal, which includes birds and flying insects. λαγῶς, as seen above, means the hare but also a kind of bird whereas ἀλώπηξ, fox, (present also in the list of beasts) means also a large bat. ἰχθύδιον, little fish, nurrishment of sea-birds, is present in both the second and third list, but it is attracted into the cluster of fishes.

Finally, the third column provides cohyponyms but also εὔχυλος, succulent, an attribute applicable to cooked fish and τίφη, a kind of beetle, but also a kind of boat.

The dendrogram generated by these data does not provide the consistency observed in the dendrogram generated by the arbitrarily selected items, but some groupments show non trivial patterns. In particular, the couples ἀετός (eagle) – λαγῶς (hare) and ἔλαφος (deer) – πάρδαλις (leopard) show a strong association between predator and victim. The triplet σκύμνος (puppy) – θήρα (hunt) – κυνηγός (hunter) confirm the suggestion that, inside the categorical partitions shown in the scatter plot (Fig. 4.4), the syntagmatic relations expressed in frames and typical

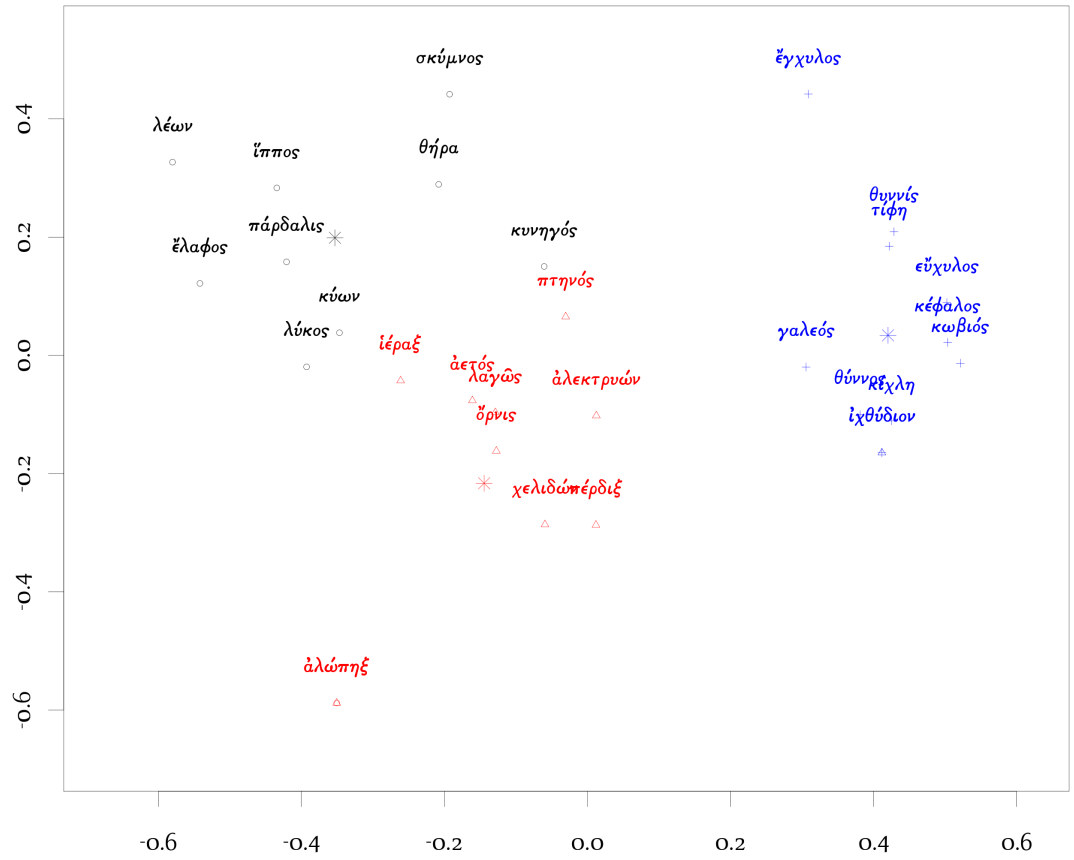

Figure 4.4: Clusters of animals generated by seeds

scenes (e.g. predation or hunting) can be captured by a different visualization of the same data.

Seed words are particularly effective to generate lists of comeronyms. Among the most productive domains, corresponding to technical and scientific knowledge of ancient Greeks, the following have been successfully tested: botanic (using the seeds *φύλλον*, leaf and *κλάδος*, branch, parts of the plant); medicine (using *λάρυγξ*, larynx and *στόμα*, mouth, body parts); nautical technique (using *πρύμνα*, poop and *πρῷρα*, prow, parts of the ship); military technique (using *κνημίς*, greave and *ἀσπίς*, shield, part of the armour) and building technique (using *κίων*, pillar and *αὐλή*, courtyard, part of the building). Most of the fifty words generated by these seeds are meronyms and they are grouped without relevant overlappings.

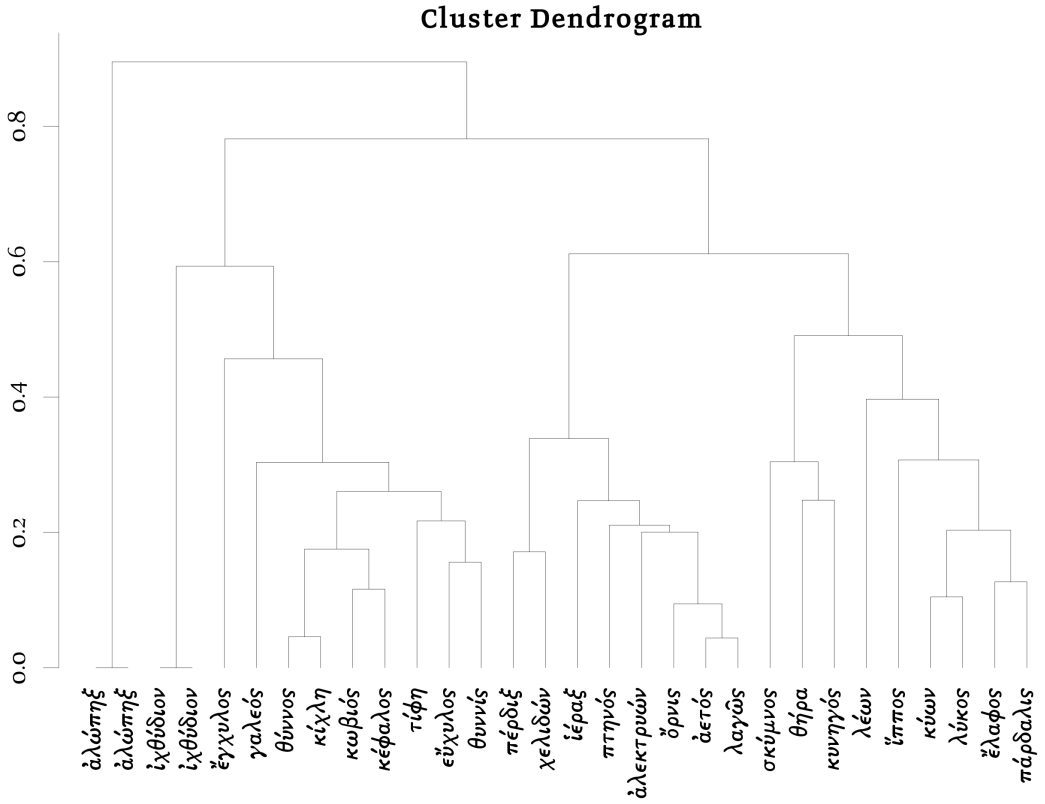

Figure 4.5: Dendrogram of animals generated by seeds

As anticipated in the introduction of this section, abstract terms that belong to traditional "tables", such as, for instance, virtues, feelings or emotions, can be correctly classified. Tab. 4.7 lists corporal, psychological and spiritual virtues, object of philosophical and theological speculation along the centuries. The list of virtues to clusterize has been manually generated. In fact, we were interested to explore groups of virtues that constitute traditional paradigms, such as faith, hope and charity, among other, less paradigmatic, such as soundness.

As shown in Fig. 4.6, clusters are clearly separated. On the right side of the plot are concentrated some physical virtues, arbitrarily selected among variable lists of virtues provided by the ancient philosophers and writers. On the left side of the plot are distributed the seven virtues of the Christian tradition, clearly

distincted in two groups: the theological virtues (charity, hope and faith) on the top of the chart and the cardinal virtues (manliness, justice, temperance and prudence) on the bottom part of the chart.

ἀρτιότης	soundness	ἀνδρεία	manliness	ἀγάπη	charity
εὐαισθησία	quick sensibility	δικαιοσύνη	justice	ἐλπίς	hope
εὐεξία	good habit of body	σωφροσύνη	temperance	πίστις	faith
ὑγίεια	health	φρόνησις	prudence		

Table 4.7: Virtues

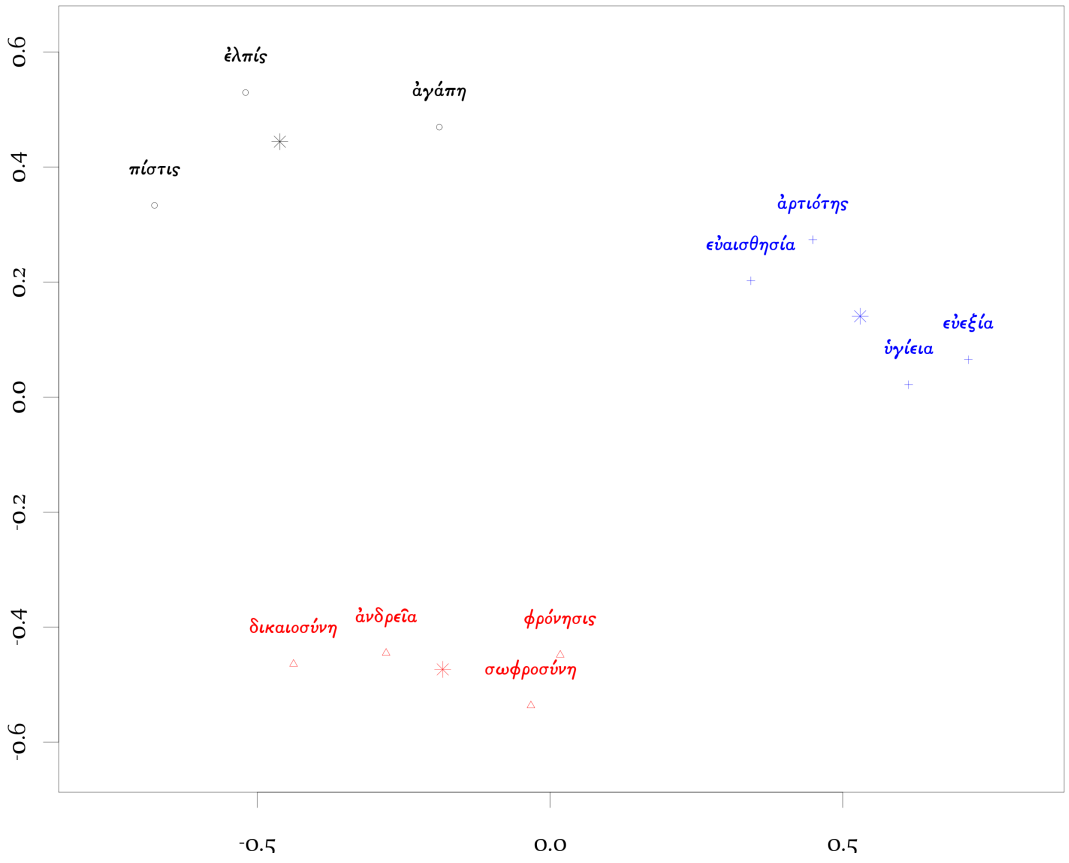

Figure 4.6: Clusters of virtues

Other abstract categories successfully tested are the emotions, using as seed the traditional primary emotions φόβος, fear, ἐλπίς, hope, χαρά, happiness and λύπη, pain.

Also abstract term clusters, like concrete term clusters seen above, can show their hierarchical structure, expressed by meaningful subsets. For example, in many experiments with different groups of virtues, the civilian virtues ἐλευθερία, freedom, and αὐτονομία, autonomy, are grouped in a common subset, due to the entailment of αὐτονομία and ἐλευθερία.

4.9 Subcorpora

Relevant subcorpora have been explored in order to show peculiar associations typical of three genres: epics, tragedy and philosophy. In fact, the study of semantic spaces built on TLG subcorpora allows the individuation of semantic associations typical of a genre or of an author. In order to show the idiosyncrasies of the subcorpora under examination, two high frequency terms have been selected: the concrete θάλασσα, sea, and the abstract θάνατος, death.

4.9.1 Homeric Poems

The formulaic diction of the Homeric Poems facilitates consistent associations in the semantic space, showing both the syntagmatic relations expressed by formulae (such as πόδας ὠκὺς Ἀχιλλεύς, swift-footed Achilles) and the paradigmatic relations expressed by formulaic systems (such as a fixed proper name followed by different epithets). The semantic consistency is valuable both for the concrete term and for the abstract one. In the case of θάλασσα (27 occurrences), the top associations are: κῦμα wave, κλύζω to wash, dash over (of the sea), κῆτος any sea-monster, θεμόω to drove the ship ashore, ἔκβασις way out of the sea, ἁλμυρός salt, ἐπαγλαίζω to honour, παλιρρόθιος refluent, ἀελπής unhoped, περιμήκης very long, high or large. For θάνατος (119 occurrences) the top associations are: τέλος, doom, μοῖρα, destiny, πότμος, evil destiny, death, ψυχή, life, soul, κήρ heart, ἐκπτύω spit out (last breath), ὑπόβρυχα under water, μόρος fate, destiny, θνήσκω die, φείδομαι spare persons and things. Both in the case of concrete terms and in

the case of abstract terms associations are meaningful, for paradigmatic reasons, such as synonymity, or for syntagmatic reasons, such as typical scenes or frames.

4.9.2 Tragedy

Due to the low frequency of the terms that characterize the tragical diction, the semantic spaces built on the Tragedy subcorpus provides not very consistent associations.

Tragedy subcorpus is constituted only by thirty-three tragedies and the tragical diction is based on daring metaphors and neologisms. The semantic space create on this subcorpus provides consistent associations only for terms that are key-words in Tragedy, such as γόος, wailing, or ἄτη, ruin, doom.

Following associations are provided by our test words. θάλασσα (34 occurrences) is associated to ῥοιβδέω, to move with a whistling sound, σύνεγγυς, near,ἐρυμνός, fortified, Ἐρυθραί, Erythrae, συγκλύζω, wash over, Ὕψισται, Thebes, Θόας, swift (of ships), πόντος, open sea, στέλεχος, trunk, τεράζω, to interpret portents. Better is the list of the semantic associations with the abstract term θάνατος (111 occurrences): μέλεος, miserable, δόλιος, treacherous, ἀνδρολέτειρα, murderess, πάθος passion, misfortune, παναίτιος, to whom all the guilt belongs, θεόκραντος, accomplished by the gods, ἰσόψυχος, of like soul, ἰχώρ, blood, βαρύμηνις, heavy in wrath, φίλιος, friendly.

Whereas θάλασσα is associated to a small group of consistent terms, θάνατος, key-word in Tragedy, provides more effective associations.

4.9.3 Philosophy

The semantic space of the ancient Greek Philosophy shows the analytical nature of the associations: concrete terms are associated to their hypernyms, hyponyms, co-hyponyms, holonyms, meronyms and co-meronyms and also to terms that express their attributes and qualities. Abstract terms are mainly associated to antonyms and synonyms or words belonging to the frame of death, such as γῆρας, old age. Both terms are associated also to semantically related verbs. θάλασσα (1,884 occurrences) produces the list ποταμός, river, ἅλμυρός, salt, πότιμος, drinkable, λίμνη, lake, ἁλμυρότης, saltness, πηγάς, earth hardened after rain, πέλαγος, open sea, στάσιμος, stagnant, εἰσβάλλω, drive to the sea, ῥέω, to flow. θάνατος

(2,107 occurrences) produces the list ἀποθνήσκω, to die, ζωή, life, θνήσκω, to die, συμφορά, misfortune, ζάω, to live, διάλυσις, dissolution, χωρισμός, departure, γῆρας, old age, τελευτάω, to end life, νεκρός, corpse.

πέρας	limit	ἄπειρον	infinite
περιττόν	odd	ἄρτιον	even
ἕν	one	πλῆθος	multitude
δεξιόν	right	ἀριστερόν	left
ἄρρην	male	θῆλυ	female
ἠρεμοῦν	resting	κινοῦμενον	moving
εὐθύ	straight	καμπύλον	curved
φῶς	light	σκότος	darkness
ἀγαθόν	good	κακόν	evil
τετράγωνον	square	ἑτερόμηκες	rectangle

Table 4.8: Pythagorean Antonomies

Philosophical texts tend to structure the knowledge in fixed tables, quoted and discussed following an established order and giving to common terms a technical meaning. Tab. 4.8 lists the Pythagorean antonymies, in order to show a clear example of the different distribution of special antonymic couples in the semantic space of the entire *Thesaurus Linguae Graecae* and in the subcorpus of the philosophical texts.

In Fig. 4.7 highlights with dotted circles the three misclassifications in the semantic space of the entire corpus. The correct clusters cross each other; καμπύλον, curved, is very close to πέρας, limit; ἑτερόμηκες, rectange, is heavily attracted by the couple περιττόν, odd and ἄρτιον, even. Also, the couples that are correctly classified do not show a strong association.

On the contrary, the philosophical semantic space shows strong associations of the Pythagorean antonymies and even the wrong associations are less serious than the wrong associations in the entire semantic space. In the philosophical semantic space mathematical terms περιττόν, odd, and ἄρτιον, even, are clearly separated from the geometrical terms τετράγωνον, square, and ἑτερόμηκες, rectangle, whereas in the total space of the TLG they overlap. In the philosophical

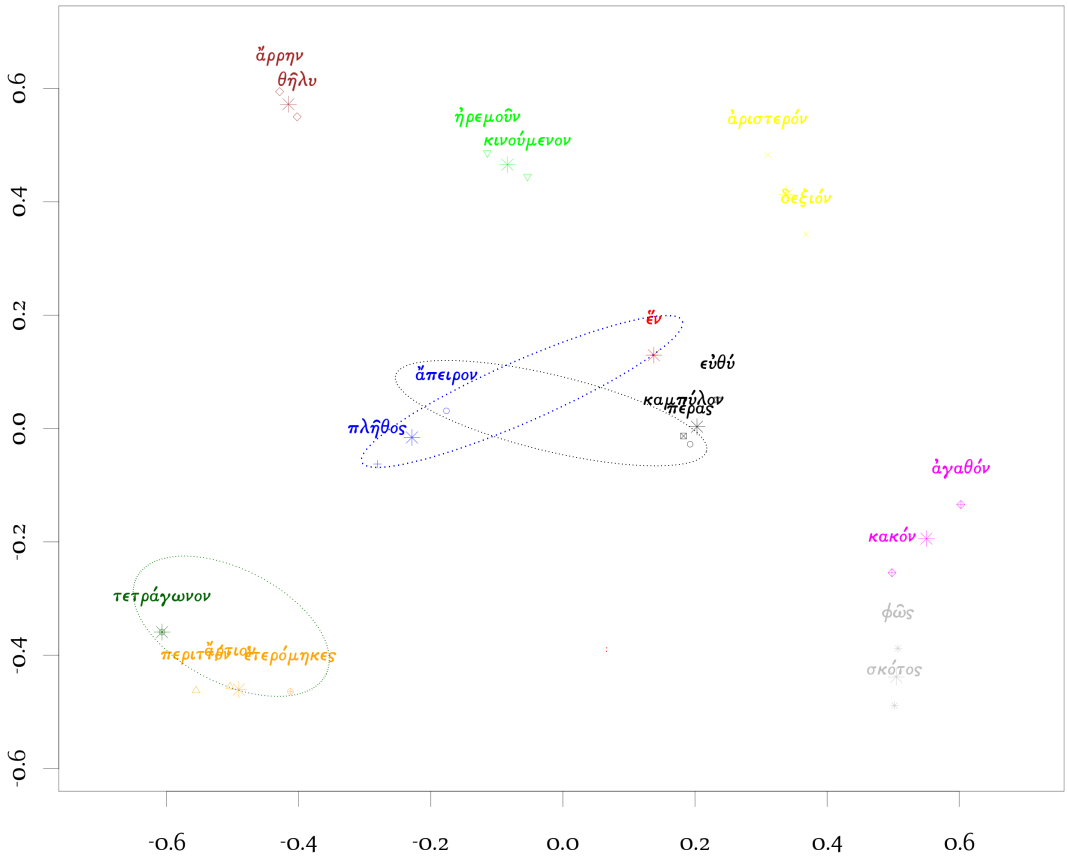

Figure 4.7: Antonymic couples in TLG

space, *ἕν*, one, and *πλῆθος*, multitude, are correctly isolated. The wrong clusters related to *εὐθύ*, straight, *καμπύλον*, curved, *πέρας*, limit, and *ἄπειρον*, infinite, do not cross. In the TLG space, the six terms are missclassified and they are close each to other in the central part of the plot.

4.10 Conclusion

On one hand, the study of Semantic Spaces usually is applied to corpora of modern languages, constitued by millions and sometimes billions of words, in particular when it is applied to the web as a corpus. On the other hand, the limited corpus of ancient Greek is diachronical, and constituted by prose and

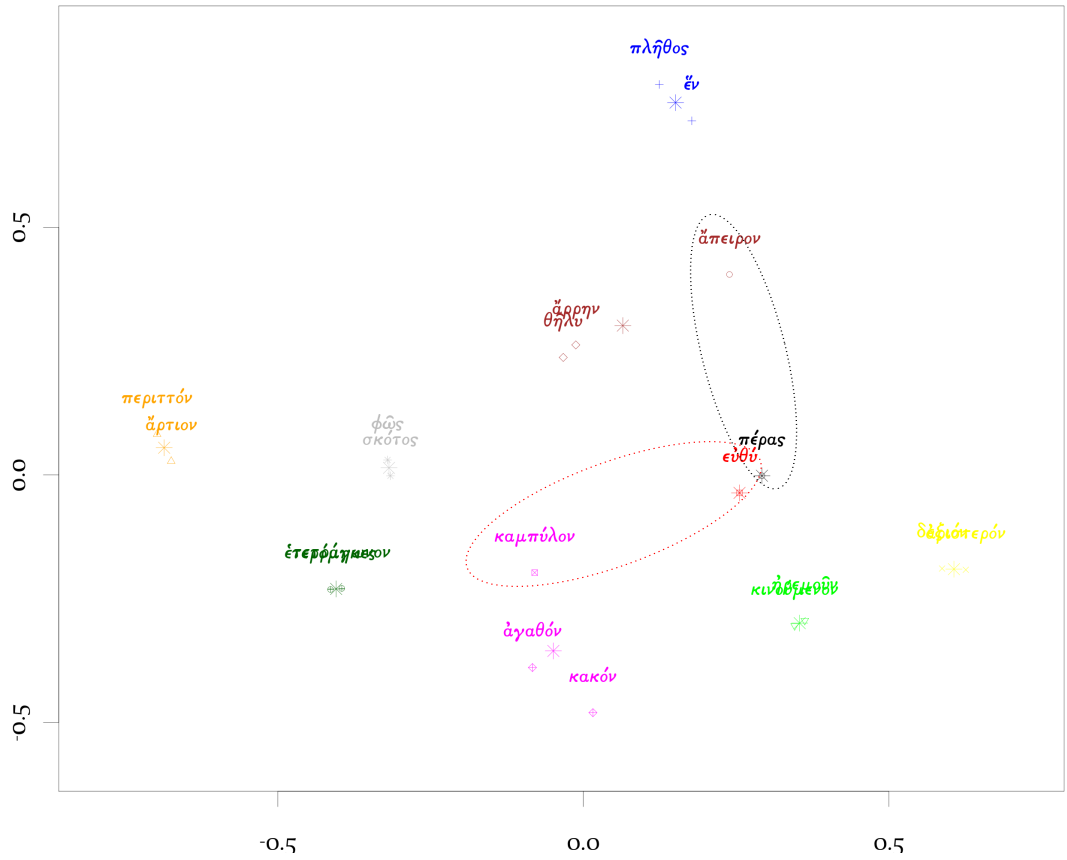

Figure 4.8: Structure of the antonymies in the philosophical corpus

poetry of different literary genres. In this chapter we have tried to explore the peculiarities of this corpus.

Semantic spaces based on chronological subcorpora allow us to study terms that have preserved or changed their meanings.

Polysemic words generate different associations in different subcorpora, consistently with specific domains that characterize the subcorpora (body parts, metrical analysis, etc.).

Antonymies and taxonomies can be studied by cluster analysis, showing how traditional antonymies, such as the Pythagorean one, are clearly grouped in a plot and how hierarchies not only of concrete but also of abstract terms can be identified.

Finally, the characteristic of semantic spaces based on Homeric poems, Tragedy and Philology have been explored by examples.

5

Conclusions

In this final chapter, we will try to illustrate how the traditional *modus operandi* of classical philology can be integrated with and empowered by digital and computational philology.

Section 5.1 points out how digital philology is leaving the age of innocence, in which variants, conjectures and secondary literature were ignored, and it is entering the maturity, from where the first "born digital" philological studies grow.

Section 5.2 and section 5.3 are devoted to some conjectures with relevant semantic implications. Evaluations are supported by the study of semantic spaces, as an example of their application in the domain of classical philology.

Section 5.4 summarizes the main achievments of this work about OCR applied to classical editions, mapping of variants and conjectures in the context of the reference edition and study of ancient Greek semantic spaces.

5.1 Putting all together

There is a weak and a strong sense of digital philology, which are strictly connected. In the first phase, the aim is to provide the philologist with digital copies of all the primary and secondary sources necessary to study classical texts, avoiding to visit physical libraries. General purpose initiatives, such as the *Million Book Project*, can satisfy this goal, with acceptable image quality of book pages, sufficiently precise metadata and decent OCR accuracy only on Latin characters.

In the second phase, that is in the strong sense of digital philology, full text must be not only fully readable (by scholars) but also fully searchable and actionable (by machines). For this purpose, both high OCR accuracy on classical editions and acceptable mapping accuracy of variant readings on reference texts are necessary. Indeed, digital philology must provide massive machine readable data that will be processed according to the methods developed by computational philology on limited case studies.

To this end, below we will illustrate the methods that can support the philologists to classify and evaluate the conjectures provided by scholars during the last centuries for the emendation of ancient Greek texts. In the following examples we have used deliberately only documents available in digital form (both digital images and texts).

5.2 Evaluating semantic similarity of conjectures

Usually philologists evaluate conjectures in comparison with *loci similes* (i.e. parallel places) of the same author, genre or period, using indexes and concordances. Text retrieval applications, such as SNS Greek or Diogenes, have facilitated the research of these parallelisms, but they have not modified the traditional methodology, based on the systematic individuation of lexical, instead of semantic similarities. In fact, lexical parallelisms can be determined by an exact word by word matching of the inflected or lemmatized forms. Semantic similarities, on the contrary, escape a universal agreement.

Semantic relations among conjectures can be explored through semantic spaces. The main purpose is to provide philologists with tools to suggest a quick and rough classification of conjectures, which should be manually adjusted. See Bozzi (2002) for the classification used in the philological workstation created at the CNR of Pisa.

For instance, we can consider *Pers.* 133-139 (West's edition):

λέκτρα δ' ἀνδρῶν πόθῳ
πίμπλαται δακρύμασιν·
Περσίδες δ' ἀβροπενθεῖς ἑκάστα πόθῳ φιλάνορι
τὸν αἰχμάεντα θοῦρον εὐνατῆρ' ἀποπεμψαμένα
λείπεται μονόζυξ.

And marriage-beds are filled with tears through longing for husbands; each Persian woman has sent to the field her warlike and fiery consort, and now in grief and longing for her beloved lord, is left forsaken by her mate. (transl. by H.W. Smyth)

The repetition of πόθῳ, longing, desire, at the distance of few verses was suspected mistaken by the 19th century scholars. In fact, we can see a plethora of conjectures in Wecklein's repertory, in order to avoid the repetition:

λέκτρα δ' ἀμῖν μάταν Enger. πόνῳ Pauw, σπάνει olim, postea ἔρῳ Heimsoeth, ὁδῷ Oberdick.

vain beds for us Enger. *toil* Pauw, *want* earlier, later *desire* Heimsoeth, *way* Oberdick

If, on one hand, the semantic classification of variants is important to study how interlinear *glossae* entered in the text or how the scribe had trivialized the original reading, on the other hand, the classification of conjectures can help the philologist to study the *modus operandi* of previous scholars, trends to correct by similarity or by contrast.

After the automatic alignment of the conjecture words with the text words, as shown in Tab. 5.1, it is simple to evaluate which words match some items in the list of semantic associations. The list of the closest words to πόθος, longing, contains ποθέω, to desire, φίλτρον, filter (of love), ἔρως, love, νυμφίδιος bridal, and ἔρος, desire. It is then verified by an automated procedure which reveals that only the later Heimsoeth's conjecture is a near synonym.

5.3 Conjecturing antonyms

From a semantic point of view, conjectures that amend a contradiction by an antonym are worthy of note. For instance, we can consider *Sept.* 705-708 (West's edition):

λέκτρα	δ'	ἀνδρῶν	πόθῳ	
		ἀμῖν	μάταν	Enger
			πόνῳ	Pauw
			σπάνει	Heimsoeth
			ἔρῳ	Heimsoeth
			ὁδῷ	Oberdick

Table 5.1: Alignment of Conjectures

μίμν' ὅτε σοι παρέστακεν, ἐπεὶ δαίμων
λήματος ἂν τροπαίαι χρονίαι μεταλ-
λακτὸς ἴσως ἂν ἔλθοι θελεμωτέρωι
πνεύματι· νῦν δ' ἔτι ζεῖ.

It is only at this moment (translating *νῦν* of the *paradosis* instead of West's *μίμν'*) that death stands close by you, for the divine spirit may change its purpose even after a long time and come on a gentler wind. But now it still seethes. (Translation by Smyth)

From West's apparatus:

705 *μίμν'* West[7]: *νῦν Ω*

707 *θελεμωτέρῳ* Conington[4]: *θαλλωτέρω* M[a]2: *θαλερωτέρω* cett.

we see that the paradosis is quite different, with *νῦν*, now, instead of *μίμν'*, stay, which avoids the repetition of *νῦν* at 708 and *θελεμωτέρῳ*, calmer, more quiet, instead of *θαλερωτέρῳ*, more luxuriant, more rapid, the term that will be discussed below (*θαλλωτερω* in the manuscript **M** before the correction, as visible in Fig. 5.1, is *vox nihili*).

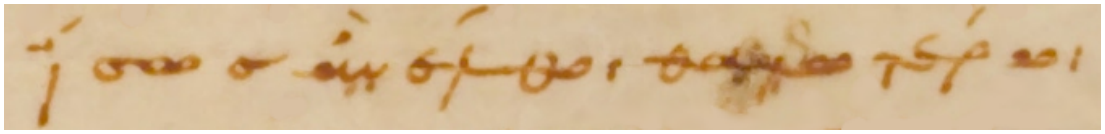

Figure 5.1: *Sept.* 707 in the manuscript **M**

The history of the conjectures starts with Hermann's adnotation:

Θαλερωτέτῳ in G explicat glossa *ἰσχυρωτέρῳ*, in Vit. *χλωροτέρῳ*. Sed hoc alienum est ab horum versuum sententia, quae id postulat, quod prior scholiastes per *ἀσθενεστέρῳ καὶ*

ἀναπεπτωκότι πνεύματι, alter per μαλακωτέρῳ expressit. Verum non potest dici θαλερώτερον πνεῦμα. Aliena sunt quae Blomfieldius in glossario attulit.

Θαλερωτέτῳ, more luxuriant, in G is explained by ἰσχυρωτέρῳ, stronger, in Vit. by χλωροτέρῳ, more green. But the sense is not compatible with the content of these verses, which requires what the first *scholiastes* has expressed by ἀσθενεστέρῳ καὶ ἀναπεπτωκότι πνεύματι, weaker and despondingly wind, and what the second *scholiastes* has expressed by μαλακωτέρῳ, more weak, milder, more tender. But we cannot say θαλερώτερον πνεῦμα, more luxuriant wind. Blomfield's justifications expressed in his glossary are inapt.

Hermann points out that the latter *glossae* in G and Vit. (in particular G) are incompatible with the earlier *scholia*. In fact, ἰσχυρός, strong, is the antonym of ἀσθενής, weak, and near antonym of μαλακός, tender. ἀσθενής is the first association with ἰσχυρός in the semantic space: as demonstrated in the previous chapter, usually the first association for adjectives is the direct antonym. μαλακός is the twenty-fourth association. Hermann's implicit conclusion is that the earlier readers read a different term than θαλερός, of opposite sense. His claim was followed by a plethora of conjectures in this direction: χαλαρωτέρῳ, more languid, by Hermann, θελεμωτέρῳ, gentler, by Conington, καθαρωτέρῳ, more pure, or μαλακωτέρῳ, more tender, by Heimsoeth and γαλερωτέρῳ, more cheerful, by Scheer.

But Blomfield's documentation[1], rejected by Hermann's sensibility in this context, shows the ambiguity of θαλερός, which can effectively oscillate from ἰσχυρός, strong, to ἁπαλός, delicate.

With explorative purposes, we have plotted πνεῦμα, the certain noun, surrounded by θαλερός, the word of the *paradosis* and the scholars' suggestions in the semantic space of the TLG. This plot suggests that θαλερός is neither incompatible with the noun πνεῦμα nor with the sense of μαλακός attested byone of the *scholia vetera*. This conclusion is also confirmed by other plots with single couples of words not shown here.

[1]The integral note of Blomfield says: Θαλερός. *Mollis.* Proprie dictum de θάλλοις, i. e. *surculis* plantarum, mollitiei notionem facile contraxit. Etymol. M. p. 441, 52. Θαλερός. ὁ ἐν ἀκμῇ ὢν τοῦ θάλλειν, ὁ ἀκμάζων νέος. καὶ Θαλερὸν δάκρυ, (Iliad. B. 366. ubi Schol. Venet. τὸ ἔνικμον, ἀπὸ τῆς τῶν φυτῶν μεταφορᾶς,) τὸ ἁπαλόν. Lex. Rhetor. MS. Bibl. Coslin. p. 500. Θαλερός. ἀκμαῖος, ἰσχυρός. νέος. ἢ ὁ ἁπαλός. καὶ θαλερὸν δάκρυ. τὰ γὰρ θάλλοντα φύλλα ἁπαλά.

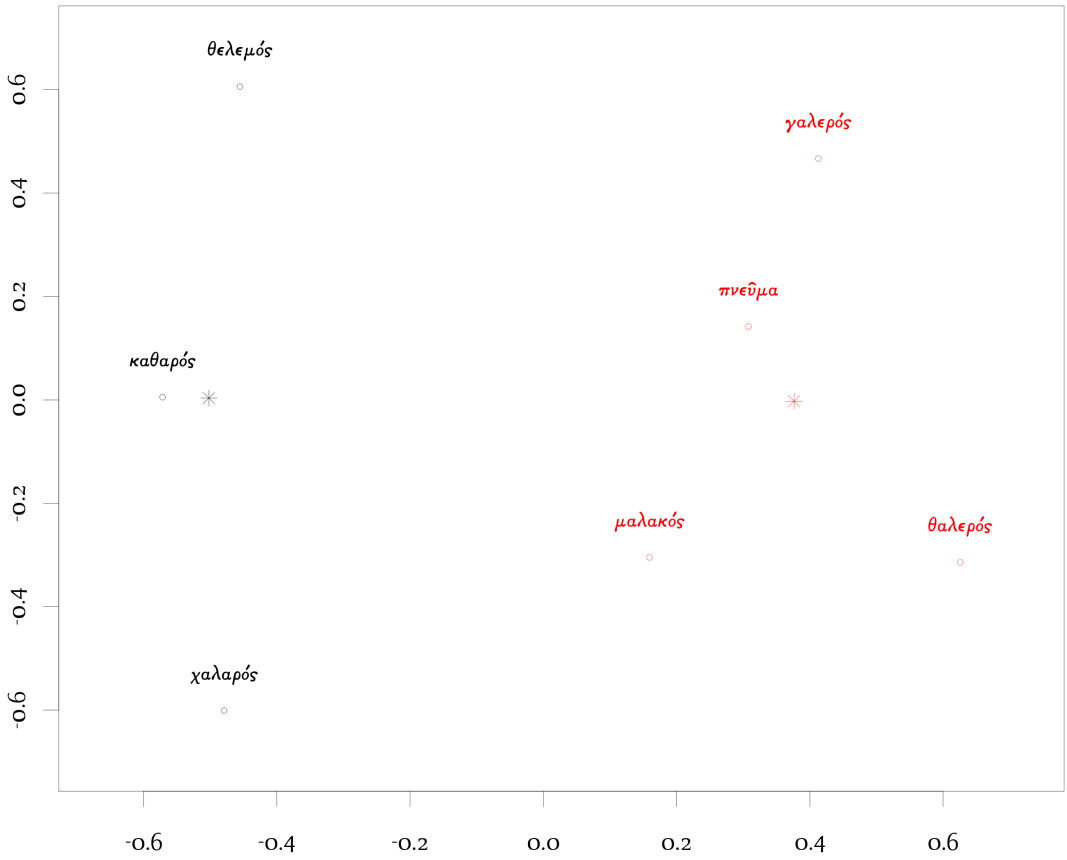

Figure 5.2: Conjectured words in the semantic space of the TLG

5.4 Conclusion

The main achievments of this work are an improvement of OCR accuracy for classical editions, the creation of tools to align variants and conjectures with the reference edition and an exploratory method to evaluate semantic relations among ancient Greek words.

Digitization of classical texts is the bottleneck that delays the development of digital philology and, consequently, of computational philology. Indeed, improving OCR accuracy is the necessary step to scale digital projects in classics.

Also the automated alignment of variants and conjectures reduces both the costs and the arbitrary choices of experts that edit digital critical apparatus.

Finally, the exploration of ancient Greek semantic spaces can suggest the philologist new lines of investigation to evaluate textual variants and conjectures of previous scholars.

Bibliography

Allsopp, J. (2007). *Microformats: Empowering Your Markup for Web 2.0*. New York, NY. 25

Baayen, H. (2008). *Analyzing Linguistic Data - A Practical Introduction to Statistics Using R*. Cambridge, MA. 65

Bamman, D., Mambrini, F., and Crane, G. (2009). An Ownership Model of Annotation: The Ancient Greek Dependency Treebank. In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories*, Milan. 36

Boschetti, F. (2005). *Saggio di analisi linguistiche e stilistiche condotte con l'ausilio dell'elaboratore elettronico sui* Persiani *di Eschilo*. Ph.D. thesis, Università degli Studi di Trento. 36, 38

Boschetti, F. (2008). Alignment of Variant Readings for Linkage of Multiple Annotations. In P. Zemanek, editor, *Proceedings of the ECAL 2007: Electronic Corpora of Ancient Languages*, pages 11–24, Praha. 31

Boschetti, F. (2009). Trends in computational philology. *Lexis*, **27**, 1–4. 2

Boschetti, F., Romanello, M., Babeu, A., Bamman, D., and Crane, G. (2009). Improving OCR Accuracy for Classical Critical Editions. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou and G. Tsakonas, editors, *Research and Advanced Technology for Digital Libraries – 13th European Conference, ECDL 2009, Corfu, Greece, Sept./Oct. 2009, Proceedings*, pages 156–167. 9

BIBLIOGRAPHY

Bourgeois, F. L. and Emptoz, H. (2007). DEBORA: Digital AccEss to BOoks of the RenAissance. *International Journal on Document Analysis and Recognition*, **9**, 192–221. 11

Bozzi, A. (2000). Character Recognition and the Linguistic Spelling Checker: an integrated technique. In A. Bozzi, editor, *Computer-aided recovery and analysis of damaged text documents*, pages 161–186. Bologna. 11

Bozzi, A. (2002). New trends in philology: a computational application for textual criticism. *Euphrosyne*, **n.s. 30**, 267–285. 4, 11, 84

Bozzi, A. (2004a). Postfazione a Zampolli Antonio, Filologia e informatica: le origini della filologia computazionale. *Euphrosyne*, **n.s. 32**, 21–24. 2

Bozzi, A. (2004b). Verso una filologia computazionale: la prima Euroconferenza della European Science Foundation. *Euphrosyne*, **n.s. 32**, 127–138. 2, 33

Bozzi, A., Nikolova, A., Cappelli, G., and Giuliani, G. (1986). Il trattamento delle varianti nello spoglio elettronico di un testo. Una prova sui Carmina di Claudiano. *Materiali e Discussioni per l'Analisi dei Testi Classici*, **16**, 155–179. 33

Buzzetti, D. (1999). Rappresentazione digitale e modello del testo. In C. et al., editor, *Il ruolo del modello nella scienza e nel sapere*, pages 127–161, Roma. Accademia Nazionale dei Lincei. 33

Cecotti, H. and Belaïd, A. (2005). Hybrid OCR combination approach complemented by a specialized ICR applied on ancient documents. In *8th International Conference on Document Analysis and Recognition*, pages 1045–1049. 11

Ciula, A. and Stella, F., editors (2007). *Digital Philology and Medieval Texts*. Ospedaletto–Pisa. 2

Cormode, G. and Muthukrishna, S. (2007). The string edit distance matching problem with moves. *ACM Transactions on Algorithms*, **3**(1), 1–19. 46

Crane, G. (1991). Generating and Parsing Classical Greek. *Literary and Linguistic Computing*, **6**(6), 243–245. 19, 60

Crane, G., Bamman, D., Cerrato, L., Jones, A., Mimno, D., Packel, A., Sculley, D., and Weaver, G. (2006). Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries. In *10th European Conference on Research and Advanced Technology for Digital Libraries, volume 4172 of Lecture Notes in Computer Science*, pages 353–366. Springer. 3, 24, 30

Crane, G., Seales, B., and Terras, M. (2009). Cyberinfrastructure for Classical Philology. *Digital Humanities Quarterly*, **3**(1), 1–27. 2, 3

Crochemore, M., Hancart, C., and Lecroq, T. (2007). *Algorithms on Strings*. Cambridge University Press. 48

Csernel, M. and Patte, F. (2007). Critical Edition of Sanskrit Texts. In *1st International Sanskrit Computational Linguistics Symposium*, pages 95–113. 11

Dawe, R. (1963). *The collation and investigation of the manuscripts of Aeschylus*. Cambridge University Press, Cambridge, U.K. 36

Dawe, R. (1965). *Repertory of conjectures on Aeschylus*. Brill, Leiden. 36

Degani, E. (1992). Il mostro di Irvine. *Eikasmos*, **3**, 277–278. 1, 33

Edwards, J., Teh, Y., Forsyth, D., Bock, R., Maire, M., and Vesom, G. (2004). Making Latin Manuscripts Searchable using gHMM's. *Advances in Neural Information Processing Systems*, **17**, 385–392. 11

Evert, S. and Baroni, M. (2007). zipfR: Word frequency distribution in R. In *Proceedings of the Demo and Poster Sessions of ACL 2007*, pages 29–32, East Stroudsburg, PA. 62

Fellbaum, C., editor (1998). *Wordnet - An Electronic Lexical Database*. Cambridge, MA. 60, 66

Feng, S. and Manmathan, R. (2006). A Hierarchical, HMM-based Automatic Evaluation of OCR Accuracy for a Digital Library of Books. In *Proceedings of the Joint Conference in Digital Libraries*, pages 109–118. 15, 65

BIBLIOGRAPHY

Fillmore, C. (1997). The need for a frame semantics in linguistics. *Statistical Methods in Linguistics*, **12**, 5–29. 62

Froger, D. (1968). *La critique des textes et son automatisation*. Dunod, Paris. 33

Grossman, D. A. and Frieder, O. (2004). *Information Retrieval: Algorigthms and Heuristics*. New York. 59

Gulick, C., editor (1951). *The Deipnosophists of Athenaeus*. London – New York, NY, 2nd edition. 12

Hermann, G., editor (1852). *Aeschyli tragoediae*. Lipsiae. 12

Hilpert, M. and Gries, S. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, **24**(4), 385–401. 5

Jlaiel, M. B., Kanoun, S., Alimi, A., and Mullot, R. (2007). Three decision levels strategy for Arabic and Latin texts differentiation in printed and handwritten natures. In *9th International Conference on Document Analysis and Recognition*, pages 1103–1107. 10

Kaibel, G., editor (1887). *Deipnosophistarum Libri XV*. Lipsiae. 12

Kondrak, G. (2002). *Algorithms for Language Reconstruction*. Ph.D. thesis, University of Toronto, Toronto. 16, 45

Kontostathis, A. and Pottenger, W. (2003). *A Framework for Understanding Latent Semantic Indexing (LSI) Performance* – Paper online. http://webpages.ursinus.edu/akontostathis/KontostathisIP&M.pdf. 59

Lenci, A. (2008). Distributional semantics in linguistc and cognitive research. A foreword. *Rivista di Linguistica*, **1**, 1–30. 58

Leydier, Y., Bourgeois, F. L., and Emptoz, H. (2005). Textual Indexation of Ancient Documents. In *2005 ACM symposium on Document engineering*, pages 111–117. 11

Leydier, Y., Lebourgeois, F., and Emptoz, H. (2007). Text search for medieval manuscript images. *Pattern Recognition*, **40**(12), 3552–3567. 10

Lund, W. and Ringger, E. (2009). Improving Optical Character Recognition through Efficient Multiple System Alignment. In *Joint Conference for Digital Libraries*. 11

Mastandrea, P. (2009). Gli archivi elettronici di 'Musisque deoque'. Ricerca intertestuale e cernita fra varianti antiche. In Zurli and Mastandrea (2009), pages 41–72. 5, 11

Meineke, A., editor (1858). *Athenaei Deipnosophistae*. Lipsiae. 12

Moalla, I., Lebourgeois, F., Emptoz, H., and Alimi, A. (2006). Image Analysis for Paleography Inspection. *Document Analysis Systems*, **7**, 25–37. 10

Mondin, L. (2009). Appunti per una critica (inter)testuale della poesia latina. In Zurli and Mastandrea (2009), pages 73–106. 5

Monroy, C., Kochumman, R., Furuta, R., Urbina, E., Melgoza, E., and Goenka, A. (2007). Visualization of Variants in Textual Collations to Analyze the Evolution of Literary Works in The Cervantes Project. In *6th European Conference on Research and Advanced Technology for Digital Libraries*, pages 638–653. 11

Mordenti, R. (2001). *Informatica e critica dei testi*. Bulzoni editore. 33

Murray, G. (1955). *Aeschyli septem quae supersunt tragoediae*. Clarendon Press, Oxford. 36

Namboodiri, A., Narayanan, P., and Jawahar, C. (2007). On Using Classical Poetry Structure for Indian Language Post-Processing. In *9th International Conference on Document Analysis and Recognition*, volume 2, pages 1238–1242. IEEE Computer Society. 11

Navarro, G. (2001). A Guided Tour to Approximate String Matching. *ACM Computing Surveys*, **33**(1), 31–88. 15

Navarro, G. and Raffinot, M. (2002). *Flexible Pattern Matching in Strings*. Cambridge University Press. 48

O'Donnell, M. B. (2005). *Corpus Linguistics and the Greek of the New Testament*. Sheffield, U.K. 5

Pickering, P. (2000). Repetitions and their removal by the copyists of Greek tragedy. *Greek, Roman and Byzantine Studies*, **41**, 123–139. 32

Reddy, S. and Crane, G. (2006). A Document Recognition System for Early Modern Latin. In *Chicago Colloquium on Digital Humanities and Computer Science: What Do You Do With A Million Books*, Chicago, IL. 23

Reynaert, M. (2008a). All, and only, the Errors: more Complete and Consistent Spelling and OCR-Error Correction Evaluation. In *6th International Conference on Language Resources and Evaluation 2008*, pages 1867–1872. 21

Reynaert, M. (2008b). Non-interactive OCR Post-correction for Giga-Scale Digitization Projects. In *CICLing 2008*, LNCS 4919, pages 617–630. 19

Reynolds, L. and Wilson, N. (1991). *Scribes and Scholars – A Guide to the Transmission of Greek and Latin Literature*. Oxford, U.K., 3d edition. 4

Rigo, G. (1999). *Eschyle. Opera et fragmenta omnia – Index verborum – Listes de fréquence*. Liège. 60

Ringlstetter, C., Schulz, K., Mihov, S., and Louka, K. (2005). The same is not the same - postcorrection of alphabet confusion errors in mixed-alphabet OCR recognition. In *8th International Conference on Document Analysis and Recognition*, volume 1, pages 406–410. 11

Rubenstein, H. and Goodenough, J. (1965). Contextual correlates of synonymy. *Comm. ACM*, **8**(10). 59

Sahlgren, M. (2006). *The Word-Space Model*. Ph.D. thesis, University of Stockholm, Stockholm. 58, 60

Shih, F. (2009). *Image Processing and Mathematical Morphology: Fundamentals and Applications*. Boca Raton, FL. 20

Smith, R. (2007). An Overview of the Tesseract OCR Engine. In *9th International Conference on Document Analysis and Recognition*, volume 2, pages 629–633. IEEE Computer Society. 13

Spencer, M. and Howe, C. (2003). Collating texts using progressive multiple alignment. *Computer and the Humanities*, **37**(1), 97–109. 16

Squitier, L. B. K. A. (1990). *Thesaurus Linguae Graecae Canon of Greek Authors and Works*. Oxford, 3 edition. 61

Stewart, G., Crane, G., and Babeu, A. (2007). A New Generation of Textual Corpora. In *Joint Conference of Digital Libraries 2007*, pages 356–365. 3, 11, 19, 23, 24

Tichy, W. (1984). The string-to-string correction problem with block moves. *ACM Transactions on Computer Systems*, **2**(4), 309–321. 46

van Beusekom, J., Shafait, F., and Breul, T. (2007). Automated OCR Ground Truth Generation. In *9th International Conference on Document Analysis and Recognition*. 15

Wall, M. E., Rechtsteiner, A., and Rocha, L. (2003). *Singular Value Decomposition and Principal Component Analysis*. Norwell, MA. 59

Wecklein, N., editor (1885). *Aeschyli fabulae cum lectionibus et scholiis codicis Medicei et in Agamemnonem codicis Florentini ab Hieronimo Vitelli denuo collatis*. 36

Wecklein, N., editor (1893). *Appendix Propagata*. S. Calvary. 36

West, M. (1990). *Studies in Aeschylus*. Teubner, Stuttgart. 36

West, M., editor (1998). *Aeschylus. Tragoediae cum incerti poetae Prometheo*. Teubner, Lipsiae. 36

Zhuang, L. and Zhu, X. (2005). An OCR Post-processing Approach Based on Multi-knowledge. In *9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*, pages 346–352. 11

Zipf, G. (1949). *Human behavior and the principle of least effort*. Cambridge. 62

Zurli, L. and Mastandrea, P., editors (2009). *Poesia Latina, Nuova E-Filologia – Opportunità per l'editore e per l'interprete*. Roma. 2, 95