Methods to extend Greek and Latin corpora with variants and conjectures
Mapping critical apparatuses onto reference text

Dr Federico Boschetti
Centro Interdip. Mente/Cervello
University of Trento
federico.boschetti@unitn.it

Abstract

The principal corpora currently available in classical literature, while quite thorough, are based on authoritative editions without critical apparatuses. However, philologists need to deal with textual variants attested by manuscripts and conjectures suggested by scholars through the centuries. This paper will explore some methods for information extraction applied to digitised apparatuses of critical editions and digital repertories of conjectures.

1 Overview

Literary corpora are usually collections of texts. From a philological point of view, this simple assertion opens 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 influenced by 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 former 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 formed to remove repetitions, besides 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 in its real extent. We can recover it only by an accurate comparison of information stored in critical apparatuses, where almost all variants and several conjectures are recorded.

The most complete collections of ancient Greek and Latin texts, such as the *Thesaurus Linguae Grecae* and the *Packard Humanities Institute*'s CD-ROMs, are based on authoritative modern editions, but they lack critical apparatuses. 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 even on printed editions, unfortunately is still valid. As we suggested 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.

But 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 apparatuses in digital editions, *cf.* Froger 1968, Bozzi *et al.* 1986, Buzzetti 1999, Mordenti 2001 and Bozzi 2004.

2 Motivation

Currently, there are two main approaches to add apparatuses 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[3]. 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 apparatuses 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 issues. Digital diplomatic editions have a practical, economical limit in the number of operators that can afford 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

1 Further information on http://www.chs.harvard.edu/publications.sec/homer_multitext.ssp

2 Further information on http://www.ricercaitaliana.it/prin/dettaglio_prin-2005105953.htm

3 http://www.tei-c.org

editions to the reference edition) we can reconstruct every diplomatic edition previously collated. This assumption is very useful even 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 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 several times 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 overlap, these last cases are similar to post-its 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. 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 apparatuses 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 for a team that must follow a common protocol for annotations. The automatic parsing of apparatuses 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 is limited to the correction phases. This third approach has a double goal: on one hand it aims to parse automatically existing critical apparatuses 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 apparatuses and repertories must be used to build new critical editions, indexes, concordances and systems for information retrieval based on variants.

3 Methodology

The first problem to afford 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 apparatuses and repertories.

Usually the reference edition is the currently most authoritative edition, by agreement of scholars. Anyway, when a new authoritative edition substitute 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 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, meanwhile the appendix of conjectures gathered by West 1990 and his own apparatus are mappable on West 1998.

One edition can differ from another one 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 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*, in smaller parts. Only the sequential position of words in the entire text provides the grid to switch from one edition to another, and even the colometry and verse numbering is based on this grid: *e.g.* βασιλεύς is on 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 apparatuses and repertories on a unified system.

Murray 1955 is the main reference edition: each word of its text has a progressive number, from the beginning to the end of each tragedy. The other reference editions, aligned on 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 fractional numbers (in case they differ for text addition: text that is between two consecutive positions in the Murray edition). Information contained in repertories are mapped on these grids.

Apparatuses and repertories, built along two centuries, differ in typographical conventions and in quantity of information, more or less accurate. Anyway, 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.

3.1 Typical structures

Apparatuses and repertories (as well as commentaries) are organised 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 extracted by apparatuses and repertories have been annotated by hand, adopting a format easily transformable by XSL in 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 ἦ$_{819}$ δ᾽$_{820}$ ἐσφάδαζε$_{821}$, καὶ$_{822}$ χεροῖν$_{823}$ ἔντη$_{824}$ δίφρου$_{825}$

197. αὐτὴ δίφρον Canter.

```
<itm>
    <vrs>197.</vrs>
        <rdng><g pos="824">αὐτὴ</g> <g pos="825">δίφρον</g></rdng>
    <resp>Canter</resp>.
</itm>
```

Simple surveys on the manual annotations confirmed that the simplest (and 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 postions were filled by blanks, or decimal numbers were used in case of insertions (*e.g.* `<itm><vrs>164.</vrs> <rdng><g pos="">κἀμὲ</g><g pos="594" val="595"/></rdng> <resp>Bothe</resp>. </itm>`, to map κἀμὲ on καὶ με or `<itm><vrs>213.</vrs> ... <rdng><g pos="917">δείμα </g> <g pos="917.001">τ᾽</g></rdng> <resp>Stanley</resp> ...</itm>` to map δείμα τ᾽ on δείματ᾽).

3.2 Reference to verses

Usually any line of the apparatus refers to one verse (*e.g.* 10.), but it might refer even to a range of verses, in particular to a couple (*e.g.* 10-11.), when the variant extends on both the verses. Rarely the

line refers to different verses (*e.g.* 800 et 820.), for instance when the same variant (conjecture) is repeated. The expressions *ante* and *post* are used if the variant (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 even in the middle (*e.g.* when a conjecture is conditioned by the suppression of another verse).

3.3 Typology of readings and sources

The simplest (and fortunately rather frequent) case is when the reading is an orthographic or morphological variant that substitute 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, an hyper/hyponym or an unrelated word, in apparatuses 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 the last words fits exactly the text of the reference edition.

The other textual operations are deletion, addition and transposition of text. Deletion usually is indicated by the word(s) to delete, 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, sometimes followed by an accurate bibliographical indication. Different apparatuses 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 the West's apparatus each manuscript is always identified by name meanwhile in the Wecklein's repertory usually manuscripts different by M (the *codex optimus*) are labelled just by recc. In the West's apparatus each modern edition is identified by the name of its author and one number (e.g. Bothe^{3}), meanwhile in the Wecklein's repertory previous editions are distinguished by the last one by the expression *olim* x (e.g. olim Bothe).

3.4 Complex cases

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

For instance (lines in smaller size are extracted from Wecklein 1885):

289 στυγναί γ᾽ Ἀθῆναι δάοις·

289. στυγναὶ δ᾽ Ἀθᾶναι recc. Δάοις Merkel, δαμόταις Oberdick.

In this case three chunks of information are easily separable in three couples reading-source.

Complex cases 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 ἕν, 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.

3.5 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, manuscript abridgement, 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 growing tables. The starting 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 recognised 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.

3.6 Alignment

About 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 apparatuses 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, in order to map the readings on the exact position of the verse in the reference edition.

In fact, we cannot limit 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.

Alignment algorithms are well known, for instance, in genomic studies, where strings of proteins must be compared and aligned. Optimised alignment algorithms with block moves, necessary to deal with transpositions, are discussed, among the others, in Tichy 1984 and in Comrode and Muthukrishnan 2007. Alignment algorithms, that evaluate the similarity of any string with another string or part of it, are based on the edit distance, *i.e.* the evaluation of costs to perform additions, subtractions, substitutions and transpositions of blocks in order to transform the first string in the second one or in a part of it. Following this principle, any chunk of text (the reading) is aligned with the portion of text (the part of the line in the reference edition) where the edit distance is lowest (*i.e.* the similarity is highest).

Optimised aligned algorithms, very efficient with huge amounts of data to compare, usually do not deal with well defined intermediate units between the characters and entire strings, like words; even moved blocks fit better the concept of stem than the idea of inflected form, sometimes generating issues in the exact anchoring of boundary words. Strings to align in our current work are relatively short, so we preferred to tune precision on a "brute force" combinatorial algorithm, with the purpose of

affording optimisation in a future step. Currently, time consumption due to complexity of the algorithm is acceptable.

An example should clarify how the algorithm works. We can consider *Pers.* 406 and the relative line in the Wecklein's repertory:

406. ἐλευθεροῦτε πατρίδ᾽, ἐλευθεροῦτε δέ

406. ἐλευθεροῦτε δή ALudvig

The algorithm reconstructs all the combinations of adjacent words in the reference text (capitalised and without spaces) and it compares them with the reading and its permutations. The best score is assigned applying the formula: 1 - edit_distance(str1, str2) / max(length(str1), length(str2))

```
ΕΛΕΥΘΕΡΟΥΤΕ
ΕΛΕΥΘΕΡΟΥΤΕΠΑΤΡΙΔ
ΕΛΕΥΘΕΡΟΥΤΕΠΑΤΡΙΔΕΛΕΥΘΕΡΟΥΤΕ
ΕΛΕΥΘΕΡΟΥΤΕΠΑΤΡΙΔΕΛΕΥΘΕΡΟΥΤΕΔΕ
           ΠΑΤΡΙΔ
           ΠΑΤΡΙΔΕΛΕΥΘΕΡΟΥΤΕ
           ΠΑΤΡΙΔΕΛΕΥΘΕΡΟΥΤΕΔΕ
                 ΕΛΕΥΘΕΡΟΥΤΕΔΕ     <--  ΕΛΕΥΘΕΡΟΥΤΕΔΗ/ΔΗΕΛΕΥΘΕΡΟΥΤΕ  (best score)
                           ΔΕ
```

All the permutations are checked only if the reading contains few (actually up to five) words, otherwise only a selected number of them are performed (up to ten items, permutations of words that are not on the left or right boundaries are excluded) or they are not performed at all, if they are too many.

3.7 Towards the processing of items containing Latin sentences

The method seen above is applied only on items constituted by Greek sequences, immediately followed by source. Anyway, in *c.* 13 percent of cases, the item to process contains an explanation (in Latin language) of the textual operation to perform (*e.g. addit*, *delet*, *transponit*), or a judgement (*damnat*, *spurium putat*).

Currently these items remain unprocessed, but sequences of Latin words contained in these chunks of information constitute a predicate that, in a future stage of the work, will be automatically processed. The typical structure (obj – pred – subj) is: (Greek_sequence) – Latin_sequence – source (*e.g.* 3. καὶ πολυχρύσων delet Bothe, or 13. post οἴχωκεν aliquot versiculos intercidisse putat Schuetz), where the Greek_sequence + Latin_sequence or the Latin_sequence alone is the reading and indicates a textual operation (in these examples, the deletion of two words or the presence of a probable *lacuna*).

In order to prepare this future second stage of information extraction from apparatuses and repertories, Latin words have been grouped, manually lemmatised and associated to morphological features, like in the sample below:

```
colloco<v> / collocat<vipa3s>
commemoro<v> / commemorati<vpt>
compono<v> / componit<vipa3s>
conicio<v> / coniciebat<viia3s>|conicit<vipa3s>|coniecit<vita3s>
coniectura<n> / coniectura<nnfs>
constituo<v> / constituit<vipa3s>
continuo<v> / continuat<vipa3s>
cum<p> / cum
do<v> / dabat<viia3s>|dat<vipa3s>|datos<vpt>
damno<v> / damnat<vipa3s>
```

Lexical variety (and semantic ambiguity) is very reduced: in the *Persians*' section of the Wecklein's

repertory only *c.* 200 headwords have been extracted. This words have been grouped in synsets and semantic relations of hyper/hyponymy, holo/meronymy have been established. For example, *antistrophe – epodus – mesodus – strophe – systema –* [*versiculus*, *versus*] are hyponyms of *metrica_divisio*, meanwhile *choreuta – chorus – choryphaeus – epodus – hemichorium – nuntius* are hyponyms of *dramatis persona*. This organisation will be used in order to build frames to perform automatically textual operations of addition, deletion, seclusion etc.

4 Results

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 evaluated by hand. Processed items (83 on 95: *c.* 87 percent) are formal substitutions (*i.e.* items containing Latin predicates are excluded). Correct processed items are 73: *c.* 77 percent on the total but a rather encouraging 88 percent on the processed items.

In the following table results are compared with methods adopted in previous stages.

	Mapping word by word	Mapping chunk by chunk w/o 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

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 necessarily shows bad performances with inserted and split words. 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 on ποιμα-| νόριον, 641 ἆρ᾽ on ἦ ῥ᾽, or, on the contrary, the mapping of one word onto two words, *e.g.* 636 δ᾽ἀμβαΰζω on διαβοάσω. Permutation of reading's elements allows the correct mapping for short transpositions: 330, *e.g.* πλεῖστον εἶς ἀνὴρ on εἶς ἀνὴρ πλεῖστον.

5 Conclusion

The system works because, probably, even the human act of mentally mapping readings on their contexts is largely based on an unconscious evaluation of edit distances. Anyway, there are even 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. Anyway, luckily, this cases are very rare.

The "natural language" of critical apparatuses and repertories of conjectures is very schematic and

deals with few textual objects and textual operations. It is possible in this way to identify recurrent frames in order to automatically extract information from them. Incorrect mappings must be checked by hand, but performances are encouraging. The output is compatible with the format used by the *Musisque Deoque* project, thus manual corrections can be performed using the editor created by the Venetian research group, that implemented even a system for building digital concordances on texts and variants.

Using that application for information retrieval, a search string like

(ANHP|ANΔP.*) EP.*? (inflected forms of ἀνήρ followed by inflected forms of ἔρος)

will produce the result

...

136. λέκτρα δ᾽ **ἀνδρῶν >>πόθῳ<<**

ἔρῳ Heimsoeth

...

according to the correct mapping of ἔρῳ on πόθῳ.

References

Bozzi, A., A. Nikolova, G. Cappelli, G. Giuliani (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 *XVI*, 155-179.

Bozzi, A. (2004) 'Verso una filologia computazionale'. *Euphrosyne n.s. XXXII*, 127-138.

Buzzetti, D. (1999) Rappresentazione digitale e modello del testo, in Carrà et al. (edd.) Il ruolo del modello nella scienza e nel sapere, pp. 127-161. Roma: Accademia Nazionale dei Lincei.

Cormode, G. and Muthukrishnan, S. (2007) 'The string edit distance matching problem with moves'. *ACM Transactions on Algorithms III*, 1, 1-19.

Dawe, R.D. (1963) The collation and investigation of the manuscripts of Aeschylus. Cambridge: Univ. Press.

Dawe, R.D. (1965) Repertory of conjectures on Aeschylus. Leiden: Brill.

Degani, E. (1992) 'Il mostro di Irvine'. *Eikasmos III*, 277-278.

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

Mordenti, R. (2001) Informatica e critica dei testi. Roma: Bulzoni editore.

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

Pickering, P.E. (2000) 'Repetitions and their removal by the copyists of Greek tragedy'. *Greek, Roman and Byzantine Studies XLI, 2,* 123-139.

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

Wecklein, N. (1885) Aeschyli fabulae. Berlin: S. Calvary.

Wecklein, N. (1893) Appendix propagata. Berlin: S. Calvary.

West, M.L. (1990) Studies in Aeschylus. Stuttgart: Teubner.

West, M.L. (1998) Aeschylus. Tragoediae cum incerti poetae Prometheo. Lipsiae: Teubner.