Alignment of Variant Readings for Linkage of Multiple Annotations

Federico Boschetti

University of Trento

1. Introduction

Digital corpora of ancient languages can be extended in two directions: with variants and conjectures[1] and 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. Furthermore, even the items of variant readings need to be lemmatized, associated to the parts of speech, etc.
It is difficult to determine the basic unit of variants and conjectures. If attention is focused on their origin 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) 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 article illustrates a method for automatical extraction of information from the 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).

2. Reference editions, critical apparatus and repertories

The current study[2] is based on the text of Aeschylus. The main reference edition used is Murray 1955, because it had been the source for the annotated corpus built by the C.I.P.L. of Liège,[3] used for this work. Lemma and part of speech are associated with each word. With regard to the text of *Persians*, morphological features of declination and conjugation and metrical structure of each word have been added.[4]
The processed apparatus and repertories are based on three different reference editions: the critical apparatus and the repertories of conjectures edited by Wecklein

1 See Bozzi et al. 1986, Bozzi 2004 and Mordenti 2001.
2 For further details, see Boschetti 2007.
3 See `http://www.cipl.ulg.ac.be` (20.01.2008).
4 See Boschetti 2005.

1885 and Wecklein 1893 are based on the text established in his critical edition, Wecklein 1885; the collations of manuscripts edited by Dawe 1963 and his repertory of conjectures, Dawe 1965, are based on Murray 1955; the appendix of conjectures edited by West 1990 and his apparatus are based on West 1998.

2.1 Collation and alignment of reference editions

Murray 1955 constitutes the main reference edition for the current work: operatively, it means that each word of its text has a progressive integer number, starting from the beginning of each tragedy. The other two reference editions, aligned to Murray, can have empty positions (in case of deletion) or positions marked by decimal numbers (in case of addition). Information contained in the repertories is mapped to their reference editions, according to the array of positions.

3. 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. Sometimes insertions assume the specific function of iterations, deletions are registered as *lacunae* or omissions, etc., but from a computational point of view the basic operations[5] allow any transformation from the source string to the target string.

In apparatuses 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 represented 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. The other 10 percent of variants and conjectures are composed of more complex structures, with a Latin sentence that expresses the textual operation that should be performed (e.g. *delet, iterat, transponit,* etc.).

In the current work only sequences of Greek text followed by the responsible(s) of variant and conjectures are processed, because more complex structures require techniques of natural language processing that will be addressed in the second stage of the work.

4. Kinds of alignment

In apparatus and repertories, variants and conjectures are located only by the reference to the verse, not by the precise position inside the verse. In fact, this informa-

5 Transposition can be reduced to a deletion and insertion in another place of the same text.

tion is superfluous for philologists and scholars, but it is not trivial to be recovered by automatic procedures.
Alignment algorithms, evaluating the similarity of a string with another string or part of it, are based on the edit distance, i.e. 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. highest similarity).
The alignment of variants with regions of the reference edition can be performed with different degrees of granularity for different purposes.

4.1 Sequence-by-sequence alignment

A coarse grained alignment identifies the part of the verse(s) in the reference edition that should be substituted by 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 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 even 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 case of linkage performed by human operators,[6] because only the starting point and the end point must be determined, reducing individual choices.

4.2 Word-by-word alignment

When corpora are enriched by variants and conjectures, it is suitable that redundant or irrelevant information is ignored, discarding the words of the reading with the mere function of placeholders.
Table 2. shows possible ways to map the conjecture *οὐδαμ᾽ οὖσ᾽ ἐμαυτῆς* L. Schmidt to Pers. 165 *μῦθον οὐδαμῶς ἐμαυτῆς οὖσ᾽ ἀδείμαντος, φίλοι*. The last word, *ἐμαυτῆς*,

6 Among others, this solution has been adopted by the Musisque Deoque Project (`http://www.mqdq.net` [20.01.2008]), aimed to provide a minimal digital critical apparatus to a large number of Latin poetical texts. 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.

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 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 same headwords, the same part of speech, the same metrical structure or the same synset with the portion of the reference edition that they should substitute.[7]
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 the hints.

Reference ed.	βάσκε	πάτερ ἄχαχε -- Δαριάν	οἷ.|
FWNewman	~~βάσκε~~	πατὴρ ἀχάχας ὁ Περσᾶν	
Common headword Common part of speech		πατὴρ ἀχάχας Noun Adj Noun	

Table 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

7 In the *Wordnet* terminology, a synset is a set of synonyms. (`http://wordnet.princeton.edu` [20.01.2008])

(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.

4.3 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 powerful for the study of errors caused by scriptio continua, which are very difficult to be managed by common systems of text retrieval, indexed word-by-word.
With this type of alignment it is even possible to perform statistics about the substitution of characters, for paleographic purposes.
The classic algorithms for alignment only take into account substitutions, insertions and deletions, but modified versions exist, that 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.[8]

Reference ed.	πλαγκτοῖς ἐν διπλάκεσσιν.
Hartung	ΠΛΑΓΚΤΟΙΣΕΝ -- ΔΙΠΛΑ ΚΕΣΣΙΝ | | || || || | | | |||| ΠΛΑΓΚΤ ---ΕΝΣΠ -Ι -ΛΑΔ-ΕΣΣΙΝ
	πλάγκτ᾽ ἐν σπιλάδεσσιν

Table 4: Character-by-character alignment (*Pers.* 280)

5. 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.

5.1 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 with other optimized algorithms for alignment.[9]

8 Kondrak 2002 explains the application of these algorithms for language reconstruction, and provides the code.

9 Optimized alignment algorithms with block moves, necessary to deal with transpositions, are discussed, e.g., in Tichy 1984 and in Cormode and Muthukrishnan 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 performed in two nested loops. The external one provides every combination of adjacent words of the verse(s) in the reference edition which are 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 should clarify 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᾽ ὄθῃ* Schuetz

the algorithm reconstructs 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.

5.2 Global alignment algorithm

The second step performs a global alignment[10] between the items of the variant reading and those of the subsequence of context identified in the previous step.

10 Navarro and Raffinot 2002 and Crochemore et al. 2007 provide detailed explanations about global (Needleman-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 in 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 the words.[11] According to Table 5, for example, *τε* and *ἐφαάνθην* are totally dissimilar (-1), *ἄρ᾽* and *ἆρ᾽* are identical[12] and - suitable result - *πατρῴᾳ*, even if different from *πατρίᾳ*, is evaluated very similar to it (0.75).
Even the cost of gaps can be tuned. In the current work insertions and deletions have a penalty of -1, i.e. the same penalty of a substitution with a totally dissimilar word.

	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
ἆρ᾽	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 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 $^{-0.48}$	-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 $^{-0.50}$
ἆρ᾽	-5 $^{-1}$	-3 $^{-0.33}$	-2.71 $^{-1}$	-2.21 $^{-0.50}$	-0.25 $^{-0.60}$	1.75 $^{-1}$	0.75 $^{-0.75}$
ἐφαάνθην	-6 $^{-1}$	-4 $^{-0.78}$	-3.71 $^{-1}$	-3.21 $^{-0.78}$	-1.25 $^{-0.56}$	0.75 $^{-0.78}$	1.19 $^{-0.56}$
-- *γᾶ τε πατρῴᾳ κακὸν ἄρ᾽ ἐγενόμαν* *καὶ γᾶ -- πατρίᾳ κακὸν ἆρ᾽ ἐφαάνθην*							

Table 6: Weight matrix

11 In the current work similarity values are rescaled from -1 to 1.
12 The evaluation is performed on the capitalized characters, i.e. excluding differences due to accents.

The weight matrix (Table 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).

5.3 Lemmatization

In order to improve the alignment performance, the similarity of words with a high probability of having the same lemma is scored 1. In fact, it is suitable that forms of the same paradigm were aligned independently according to 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 very high.

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 passed as a parameter to the morphological web-service provided by the Archimedes Project.[13]

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.

5.4 Discussion about the results of word-by-word alignment

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.

13 See http://archimedes.mpiwg-berlin.mpg.de/arch/xml-rpc.html (20.01.2008). The webservice can be accessed at http://archimedes.mpiwg-berlin.mpg.de:8098/RPC2

6. Annotation of positions and word distance issues

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 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.[14]
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 to 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.

6.1 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, repertories 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[15] from the repertory of Wecklein should illustrate the problem:

> 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*

14 I am grateful to Luigi Tessarolo, for a draft about the technical details of the search engine used by the Musisque Deoque Project.

15 This is an exceptional case, selected for its complexity. At present, the automatic parser is not yet able to deal correctly with these cases.

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,[16] the best solution is to reconstruct the minimal variant context for each conjecture, ignoring the left and right placeholders, as in Table 7:

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 7: Conjectures in the context of other conjectures

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

6.2 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[17] 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 Table 8.

16 In Wecklein's repertory on *Persians* they are only 25 in 1077 verses (and almost two thousand conjectures).
17 The case of discontinuous items is discussed below.

main reference ed. (Murray)	γᾶς ἀπ᾽ Ἀσίδος ἦλθετ᾽ -- -- -- αἰαῖ δᾴαν Ἑλλάδα χώραν										
reference ed. (Wecklein)				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 8: Offset (*Pers.* 273-274)

6.3 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:

$rp=p$ if $p \leq l$;
$rp=l+os$ if $p>l$ and $p<r$;
$rp=p+g$ if $p \geq r$;

In fact, if $p \leq l$, 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 $p \geq r$, the word occurs after the variant and it is necessary to add its global offset.
Finally, computation of word distance within 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, ἦλθετ᾽ 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.

6.4 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 9: Discontinuous conjecture (*Pers.* 43-44)

The parts of a discontinuous variant or conjecture 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 into 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.

7. Towards the linkage of multiple annotations

Given a reference system based on the word unit, the components of variants and conjectures can be annotated with the same features of the main corpus.
But even annotations can have variants. In particular, if annotations take into account, at least partially, information registered in commentaries, each word can be associated to multiple metrical, morphological or semantic interpretations.
If a textual variant, individuated by a unique identifier, can extend over several words, even interpretative variants, uniquely identified, can be associated with many word positions of the reference edition or with many word positions of textual variants and conjectures.
The scenario can be quite complex. For example, we have actually annotated the entire tragedy of *Persae,* in the reference edition of West, according to the metrical schemes related to the lyrical parts in the appendix of his edition. How is it possible to annotate an alternative metrical interpretation for an entire strophe, complicated by the fact that the text reconstructed differs by the reference edition, because it accepts some different variants and some different conjectures?

The relational framework based on the word unit deals with the problem in the following steps:

a) assigning a unique identifier to the metrical interpretation;
b) determining its boundaries, i.e. the word positions before and after the scope of the metrical interpretation;
c) reconstructing the text by the association of the unique identifier of the metrical interpretation with the array of the identifiers of the variants and conjectures that constitute the ultimate reading of the strophe;
d) associating the metrical structure of single words with the relative positions.

In this way one word position can be associated with many variant readings and, independently, to many different interpretations.

8. Conclusion

In conclusion, the word-by-word alignment of variants or conjectures with the text of the reference edition seems a suitable solution for the extension of annotated corpora, in particular if they were previously created using the word as a basic unit.

References

Boschetti, Federico. 2005. *Saggio di analisi linguistiche e stilistiche condotte con l'ausilio dell'elaboratore elettronico sui Persiani di Eschilo*. Trento, PhD Thesis. http://www.univ-lille3.fr/theses/BOSCHETTI_FEDERICO.pdf (20.01.2008)

Boschetti, Federico. 2007. Methods to Extend Greek and Latin Corpora with Variants and Conjectures. http://www.corpus.bham.ac.uk/corplingproceedings07 (20.01.2008)

Bozzi, Andrea — Anna Nikolova — Giuseppe Cappelli and Giuliana 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* 16: 155-179.

Bozzi, Andrea. 2004. Verso una filologia computazionale. *Euphrosyne* n.s. 32: 127-138.

Cormode, Graham and Shanmugavelayutham Muthukrishnan. 2007. The string edit distance matching problem with moves. *ACM Transactions on Algorithms* 3(1): 1-19.

Crochemore, Maxime — Christophe Hancart and Thierry Lecroq. 2007. *Algorithms on Strings*. New York: Cambridge University Press.

Dawe, Roger D. 1963. *The collation and investigation of the manuscripts of Aeschylus*. Cambridge (UK): Cambridge University Press.

Dawe, Roger D. 1965. *Repertory of conjectures on Aeschylus*. Leiden: Brill.

Kondrak, Grzegorz. 2002. *Algorithms for Language Reconstruction*. University of To-

ronto, Ph.D. Thesis. http://www.cs.ualberta.ca/~kondrak/papers/thesis.pdf (16.01.2008)

Mordenti, Raul. 2001. *Informatica e critica dei testi*. Roma: Bulzoni Editore.

Murray, Gilbert. 1955. *Aeschyli septem quae supersunt tragoediae*. Oxford (UK): Clarendon Press.

Navarro, Gonzalo and Mathieu Raffinot. 2002. *Flexible Pattern Matching in Strings*. New York: Cambridge University Press.

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

Wecklein, Nikolaus. 1885. *Aeschyli fabulae*. Berlin: S. Calvary.

Wecklein, Nikolaus. 1893. *Appendix propagata*. Berlin: S. Calvary.

West, Martin L. 1990. *Studies in Aeschylus*. Stuttgart: Teubner.

West, Martin L. 1998. *Aeschylus. Tragoediae cum incerti poetae Prometheo*. Lipsiae: Teubner.