iAligner : a tool for syntax-based intra-language alignment

Tariq Yousef & Chiara Palladino
University of Leipzig
Classical Philology Goes Digital
Potsdam
February 17, 2017

What is text alignment?

- Text alignment is the comparison of two or more parallel texts
- It tries to define correspondences/similarities and divergences/variants
- One of the most important tasks in Natural Language Processing: it can be performed automatically through algorithmic and dynamic programming methods

W1	This morning	the cat	observed	little	birds in the		trees		.
W2		The cat	was observing		birds in the	little	trees	this morning, it observed birds for two hours	.

Intra-Language alignment: alignment of texts in the same language

Cross-language alignment: alignment of texts in different languages

- Cross-language alignment is difficult to perform automatically
- It still needs training data from manual alignment

HYE
Ցաղագս թէ որպէս կայսրն Վաղէս հալածեալ զամենայն ուղղափառ Հաւատացեալս ,և կամ որպէս կամեցաւ յանդիմանութիւն առնել կայսրն ի մէջ ճշմարտութեանն Հաւատացելոց ,և ի մէջ չախող աղանդոյն արիանոսաց .

(29) 100% HYE - DEU

DEU
Darüber ,wie der Kaiser Vałēs alle rechtgläubigen Gläubigen verfolgte ,und noch wie der Kaiser beabsichtigte ,ein Streitgespräch unter den Gläubigen der Wahrheit und unter der bösen Sekte der Arianer zu machen .

(35) 100% HYE - DEU

Armenian-German aligned text, Alignment is done manually by Uta Koschmieder

CollateX (http://collatex.net/)

- A software for expressing textual variance with a graph-based data model
- It offers a preprocessing stage (tokenization, normalization, regularization) that is configurable by the user according to issues in language, orthography, punctuation
- Uses graphs and tables as graphic outputs

CollateX – Software for Collating Textual Sources
"In a language, in the system of language, there are only differences." – Jacques Derrida

» Home   » Download   » Documentation   » Demo   » About

CollateX is a software to

1. read **multiple (≥ 2) versions of a text**, splitting each version into parts (**tokens**) to be compared,

2. **identify similarities of and differences between the versions** (including moved/transposed segments) by aligning tokens, and

3. output the alignment results in a **variety of formats for further processing**, for instance

4. to support the **production of a critical apparatus** or the stemmatical analysis of a text's genesis.

It resembles software used to compute differences between files (e.g. **diff**) or tools for **sequence alignment** which are commonly used in Bioinformatics. While CollateX shares some of the techniques and algorithms with those tools, it mainly aims for a flexible and configurable approach to the problem of finding similarities and differences in texts, sometimes trading computational soundness or complexity for the user's ability to influence results.

As such it is primarily designed for use cases in disciplines like **Philology** or – more specifically – the field of **Textual Criticism** where the assessment of findings is based on interpretation and therefore can be supported by computational means but is not necessarily computable.

Latest Version

» **collatex-tools-1.7.1.jar**
Please refer to the **documentation** for detailed information about CollateX like its underlying concepts or usage instructions.
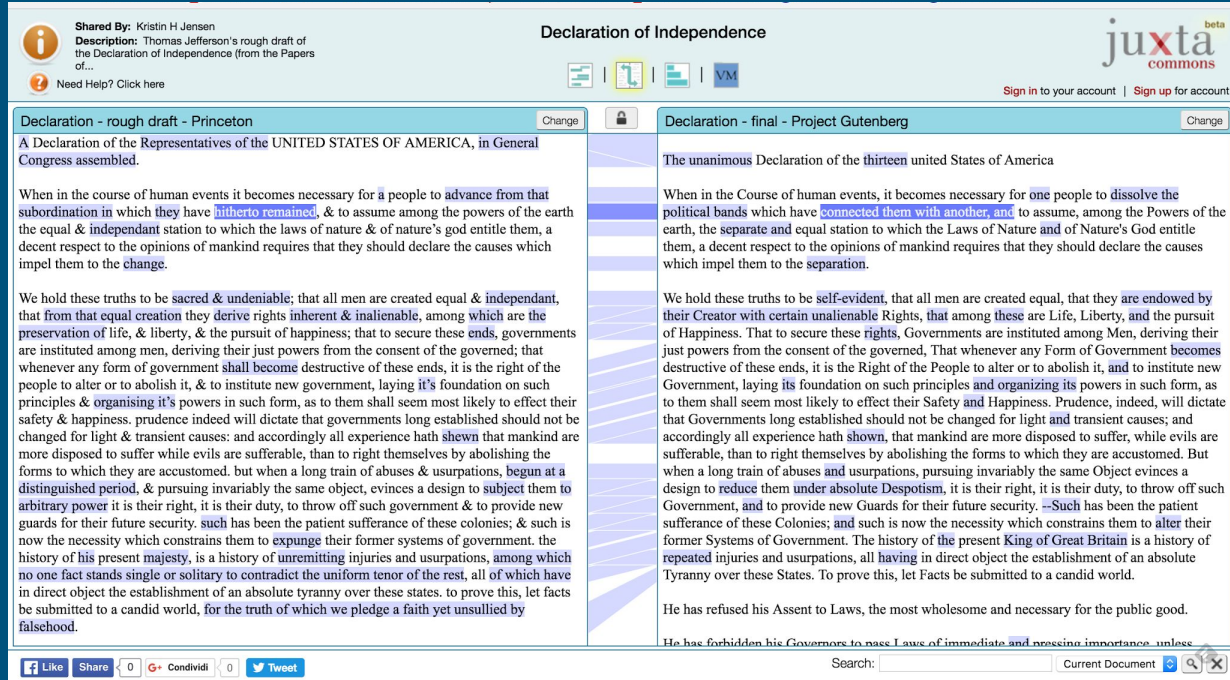
For alternative packages and license terms, please read the **download section**.

Snippet of a Variant Graph produced by CollateX

Copyright © 2010-2013 **The Interedition Development Group**. All rights reserved.

Juxta Commons (http://juxtacommons.org/)

- A group of Softwares for performing various alignment tasks
- Enables users to see changes in texts up to 20 versions
- Provides three visualization outputs: overlapping versions, parallel panels, bar histograms

The Versioning Machine (http://v-machine.org/)

- The comparison is visualized on text panels
- Allows comparison of diplomatic transcriptions or editions of the same text
- Provides a user space for annotations and bibliographical references, and an image viewer for manuscripts

eComparatio (www.ecomparatio.net)

- An ongoing project still at the experimental stage
- Will allow comparison of an arbitrary number of texts in Classical languages
- The visualization output will be in different panels with variations highlighted
- Aims at the creation of an apparatus of variants for born-digital critical editions

iAligner

A tool for automatic intra-language alignment

- **Automatic:** it is performed with algorithmic methods to reduce human intervention in the mechanical process of comparison.

- **Intra-language:** works with texts in the same language.

- **Pairwise or multiple:** works with two texts or with an unlimited number of multiple texts.

Algorithmic methods to produce alignment

The Needleman-Wunsch algorithm

- used in bioinformatics to align protein or nucleotide sequences.

- it uses Dynamic Programming to find the optimal alignment.

- divides a large problem into a series of smaller problems and uses the solutions to the smaller problems to reconstruct a solution to the larger problem.

- uses a score function and similarity matrix to represent all possible combinations of tokens and their resulting score.

The Needleman-Wunsch algorithm

- Aligning Bible Text John 1:1

NLT: In the beginning the Word already existed.

KJB: In the beginning was the Word

$$M[0][j] = j * indel \quad where \quad j \in [0, n],$$
$$M[i][0] = i * indel \quad where \quad i \in [0, m]$$

$$M[i][j] = max \begin{cases} M[i-1][j-1] + m \searrow & Matching \\ M[i-1][j-1] + mis \searrow & Mismatching \\ M[i][j-1] + indel \downarrow & Gap\ in\ S1 \\ M[i-1][j] + indel \rightarrow & Gap\ in\ S2 \end{cases}$$

$$Where \quad 0 < i <= m, \quad 0 < j <= n$$

The used score function (Matching = 5, Mismatching = -5, In/Del = -2)

		In	the	beginning	the	Word	already	existed	.
	0	-2→	-4→	-6→	-8→	-10→	-12→	-14→	-16→
In	-2↓								
the	-4↓								
beginning	-6↓								
was	-8↓								
the	-10↓								
Word	-12↓								
,	-14↓								

$$M[0][j] = j * indel \quad where \quad j \in [0, n],$$
$$M[i][0] = i * indel \quad where \quad i \in [0, m]$$

$$M[i][j] = max \begin{cases} M[i-1][j-1] + m \searrow & Matching \\ M[i-1][j-1] + mis \searrow & Mismatching \\ M[i][j-1] + indel \downarrow & Gap\ in\ S1 \\ M[i-1][j] + indel \rightarrow & Gap\ in\ S2 \end{cases}$$
$$Where \quad 0 < i <= m \ , \ 0 < j <= n$$

The used score function (Matching = 5, Mismatching = -5, In/Del = -2)

		In	the	beginning	the	Word	already	existed	.
	0	-2→	-4→	-6→	-8→	-10→	-12→	-14→	-16→
In	-2↓	**5** ↘	3→	1 →	-1 →	-3 →	-5 →	-7 →	-9 →
the	-4↓	3↓							
beginning	-6↓	1↓							
was	-8↓	-1↓							
the	-10↓	-3↓							
Word	-12↓	-5↓							
,	-14↓	-7↓							

$$M[0][j] = j * indel \quad where \quad j \in [0, n],$$
$$M[i][0] = i * indel \quad where \quad i \in [0, m]$$

$$M[i][j] = max \begin{cases} M[i-1][j-1] + m \searrow & Matching \\ M[i-1][j-1] + mis \searrow & Mismatching \\ M[i][j-1] + indel \downarrow & Gap\ in\ S1 \\ M[i-1][j] + indel \rightarrow & Gap\ in\ S2 \end{cases}$$
$$Where \quad 0 < i <= m,\ 0 < j <= n$$

The used score function (Matching = 5, Mismatching = -5, In/Del = -2)

		In	the	beginning	the	Word	already	existed	.
	0	-2→	-4→	-6→	-8→	-10→	-12→	-14→	-16→
In	-2↓	5 ↘	3→	1 →	-1 →	-3 →	-5 →	-7 →	-9 →
the	-4↓	3↓	10 ↘	8 →	6 ↘	4 →	2 →	0 →	-2 →
beginning	-6↓	1↓	8 ↓						
was	-8↓	-1↓	6 ↓						
the	-10↓	-3↓	4 ↓						
Word	-12↓	-5↓	2 ↓						
,	-14↓	-7↓	0 ↓						

$$M[0][j] = j * indel \quad where \quad j \in [0, n],$$
$$M[i][0] = i * indel \quad where \quad i \in [0, m]$$

$$M[i][j] = max \begin{cases} M[i-1][j-1] + m \searrow & Matching \\ M[i-1][j-1] + mis \searrow & Mismatching \\ M[i][j-1] + indel \downarrow & Gap \ in \ S1 \\ M[i-1][j] + indel \rightarrow & Gap \ in \ S2 \end{cases}$$
$$Where \quad 0 < i <= m \ , \ 0 < j <= n$$

The used score function (Matching = 5, Mismatching = -5, In/Del = -2)

		In	the	beginning	the	Word	already	existed	.
	0	-2→	-4→	-6→	-8→	-10→	-12→	-14→	-16→
In	-2↓	5 ↘	3→	1 →	-1 →	-3 →	-5 →	-7 →	-9 →
the	-4↓	3↓	10 ↘	8 →	6 ↘	4 →	2 →	0 →	-2 →
beginning	-6↓	1↓	8 ↓	15 ↘	13 →	11 →	9 →	7 →	5 →
was	-8↓	-1↓	6 ↓	13 ↓					
the	-10↓	-3↓	4 ↓	11↓					
Word	-12↓	-5↓	2 ↓	9 ↓					
,	-14↓	-7↓	-0 ↓	7 ↓					

$$M[0][j] = j * indel \quad where \ j \in [0, n],$$
$$M[i][0] = i * indel \quad where \ i \in [0, m]$$

$$M[i][j] = max \begin{cases} M[i-1][j-1] + m \searrow & Matching \\ M[i-1][j-1] + mis \searrow & Mismatching \\ M[i][j-1] + indel \downarrow & Gap\ in\ S1 \\ M[i-1][j] + indel \rightarrow & Gap\ in\ S2 \end{cases}$$
$$Where \quad 0 < i <= m \ , \ 0 < j <= n$$

The used score function (Matching = 5, Mismatching = -5, In/Del = -2)

New Living Translation		In	the	beginning		the	Word	already	existed	.	
King James Bible		In	the	beginning	was	the	Word				,

		In	**the**	**beginning**	**the**	**Word**	**already**	**existed**	**.**
	0	-2→	-4→	-6→	-8→	-10→	-12→	-14→	-16→
In	-2↓	5 ↘	3→	1 →	-1 →	-3 →	-5 →	-7 →	-9 →
the	-4↓	0↓	10 ↘	8 →	6 ↘	4 →	2 →	0 →	-2 →
beginning	-6↓	-2↓	8 ↓	15 ↘	13 →	11 →	9 →	7 →	5 →
was	-8↓	-4↓	6 ↓	13 ↓	11↓	9→	7→	5→	3→
the	-10↓	-8↓	4 ↓	11↓	18↘	16→	14→	12→	10→
Word	-12↓	-10↓	2 ↓	9 ↓	16↓	23↘	21→	19→	17→
,	-14↓	-12↓	-0 ↓	7 ↓	14↓	21↓	19↓	17↓	15↓

The modification to the algorithm

The goal is to optimize the algorithm by reducing the search space

compares a token `W` at the position i in `S1` with a range of tokens [*i-k*, *i+k*] in `S2` with length of *2k+1*.

The resulting search space is reduced from ($n * m$) to ($[2k +1]* m$) , where $k < n/2$

The modification to the algorithm

k = 14, n = 157, m = 134

*Search space = m*n = 21038*

after modification

*(2k+1)*m = 3886*

Multiple Sequence Alignment (In progress)

● Progressive alignment

builds up a final MSA by combining pairwise alignments beginning with the most similar pair and progressing to the most distantly related, it requires two stages:
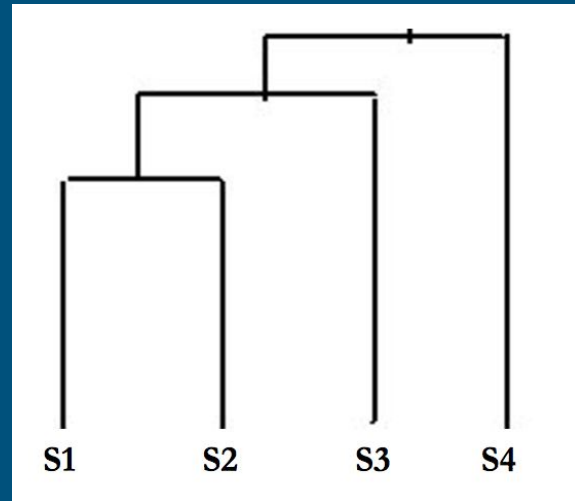
- creating the guide tree (clustering)
- adding the sequences sequentially to the growing MSA according to the guide tree

Multiple Sequence Alignment (In progress)

- Iterative alignment

The aim is to reduce the problem of a multiple alignment to an iteration of pairwise alignments.

How to align your texts with iAligner: copy your text on the editor

The text has to be parsed in sentences first

…Or upload it

Currently supports .txt and .csv files

Refinement criteria

Enter Your Text ℹ

1 The book of the generation of Jesus Christ, the son of David, the son of Abraham.
1 The book of the generation of Jesus Christ, the son of David, the son of Abraham.

2 Abraham begat Isaac; and Isaac begat Jacob; and Jacob begat Judah and his brethren;
2 Abraham begat Isaac; and Isaac begat Jacob; and Jacob begat Judas and his brethren;

3 and Judah begat Perez and Zerah of Tamar; and Perez begat Hezron; and Hezron begat Ram;
3 And Judas begat Phares and Zara of Thamar; and Phares begat Esrom; and Esrom begat Aram;

4 and Ram begat Amminadab; and Amminadab begat Nahshon; and Nahshon begat Salmon.

☐ Ignore Nonalphabetical   ☐ Ignore Diacritics   ☐ Case sensitive   ☐ Levenshtein Distance

Reset   Align

Upload your file ℹ

Select File

Scegli file   bible test.txt

Supported formats: **txt and csv**   Upload

☑ Ignore Nonalphabetical   ☐ Ignore Diacritics   ☐ Case sensitive
☐ Levenshtein Distance

- **Ignore non-alphabetical**: ignores symbols, such as punctuation and numbers, anything that is not an alphabetical character
- **Case sensitive**: if activated, detects variation across words according to the case
- **Ignore diacritics**: ignores any type of diacritical character (including punctuation marks)
- **Levenshtein distance**: applies a revised version of the Levenshtein algorithm and increases the tolerance threshold on the alignment of similar words.

The Levenshtein distance

The Levenshtein distance between two words is the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other. e.g

```
lev(George, Georg) = 1      Normalised Distance =1 - 1/6 = 0.833
lev(Prague, Prag ) = 2      Normalised Distance =1 - 2/6 = 0.666
 lev(Rome, Rom ) = 1        Normalised Distance =1 - 1/4 = 0.75
```

Mathematically, the Levenshtein distance between two strings { a,b} (of length |a| and |b| respectively) is given by $\text{lev}_{a,b}($ |a| , |b| $)$

$$
\text{lev}_{a,b}(i,j) =
\begin{cases}
\max(i,j) & \text{if } \min(i,j) = 0, \\
\min \begin{cases}
\text{lev}_{a,b}(i-1,j) + 1 \\
\text{lev}_{a,b}(i,j-1) + 1 \\
\text{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)}
\end{cases} & \text{otherwise.}
\end{cases}
$$

Modified Levenshtein Distance

Levenshtein distance is not very helpful in our case, because it is binary and there is no tolerance with errors produced by OCR or Transcription.

the distance between letters is not binary, but it is on scale. The cost of insertion or deletion depends on:

- Letter position

- Letter type (vowel or consonant)

Modified Levenshtein Distance

OCR		
Group elements		Distance within group elements
t	l	0.25
v	u	0.25
6	8	0.25
c	e	0.25
b	h	0.25
in	m	0.25
o	e	0.25
i	l	0.25
i	j	0.25

Transliteration	
Group elements	Distance within group elements
Vowel letters	0.25
C K Q	0.3
S C	0.2

lev(George, Georg) = 0.25
Normalised Distance = 1 - 0.25/6 = 0.958

lev(Prague, Prag) = 0.5
Normalised Distance = 1 - 0.5/6 = 0.916

A Greek text with no refinement criteria

τῆς δ' ἀσίας ἀπὸ κανώβου ἕως τανάιδος ποταμοῦ μετὰ τῶν κόλπων ὁ παράπλους, σταδίων μυριάδων δ καὶ ρια·
τῆς δὲ ἀσίας. ἀπὸ κανώβου. ἕως τανάιδος ποταμοῦ· μετὰ τῶν κόλπων ὁ παράπλους· στάδια Δ ρια·

| τῆς | δ' | ἀσίας | | ἀπὸ | κανώβου | | ἕως | τανάιδος | ποταμοῦ | | μετὰ | τῶν | κόλπων | ὁ | παράπλους | | , | | σταδίων | μυριάδων | δ | καὶ | ρια | · |
| τῆς | δὲ | ἀσίας | . | ἀπὸ | κανώβου | | ἕως | τανάιδος | ποταμοῦ | | μετὰ | τῶν | κόλπων | ὁ | | | παράπλους | · | | στάδια | Δ | | ρια | · |

Length: 24   Aligned-complete 13   notAligned 5   Gap 5   Aligned-case 1

The longest common substring

| μετὰ | τῶν | κόλπων | ὁ |
| μετὰ | τῶν | κόλπων | ὁ |

The same text with additional refinement criteria applied

τῆς δ' ἀσίας ἀπὸ κανώβου ἕως τανάιδος ποταμοῦ μετὰ τῶν κόλπων ὁ παράπλους, σταδίων μυριάδων δ καὶ ρια·
τῆς δὲ ἀσίας. ἀπὸ κανώβου. ἕως τανάιδος ποταμοῦ· μετὰ τῶν κόλπων ὁ παράπλους· στάδια Δ ρια·

| τῆς | δ' | ἀσίας | | ἀπὸ | κανώβου | | ἕως | τανάιδος | ποταμοῦ | | μετὰ | τῶν | κόλπων | ὁ | παράπλους | | , | | σταδίων | μυριάδων | δ | καὶ | ρια | · |
| τῆς | δὲ | ἀσίας | . | ἀπὸ | κανώβου | . | ἕως | τανάιδος | ποταμοῦ | | μετὰ | τῶν | κόλπων | ὁ | παράπλους | | · | | στάδια | | Δ | | ρια | · |

Length: 24   Aligned-complete 13   notAligned 1   Gap 5   Aligned-levenshtein 2   Aligned-removeddiacritics 1   Aligned-removedNonAlphanumeric 1   Aligned-case 1

The longest common substring

| μετὰ | τῶν | κόλπων | ὁ | παράπλους | | , | σταδίων |
| μετὰ | τῶν | κόλπων | ὁ | παράπλους | · | | στάδια |

Alignment output: a table-graph

iAligner displays all the nuances of variants according to a color-key:

- Completely aligned tokens (deep green)
- Tokens aligned by excluding case sensitivity or punctuation detection (light green)
- Gaps (yellow)
- Divergences (red)
- Tokens aligned by applying Levenshtein distance (blue-green)

What can you do with iAligner?
Some case studies

Manuscript alignment

Three manuscripts of Plato's *Crito* aligned (http://i-alignment.com/crito/)

OCR output alignment

Alignment of two OCR outputs from the *Patrologia Graeca*. The third column shows the overlapping sections and offers the user the choice between two variants where the two texts diverge.

OCR output alignment

Patrologia Latina: OCR output vs. correct version: http://www.i-alignment.com/pl/index.html

OCR output alignment

SANCTI AMBROSII MEDIOLANENSIS EPISCOPI

LIBER PRIMUS.
1113 [i]. Deduximus, fratres dilectissimi, hostiam meam, hostiam incontaminatam , hostiam Deo placentem ,
[b] domnum et fratrem meum Satyrum.
Memineram esse mortalem , nec fefellit opinio,

a Scripti anno 579.

b Omnes edit. ac mss. nonnum, *dominum et fra trem* ; alii mss. plures et potiores, *domnum et fratrem*. Observat autem Mcnardus noster in cap. 70 Regulæ S. Bened. vocem *Dominum* regiam et imperatoriam esse, nec ea salutari voluisse Augustum Cæsarem : postea autem a scriptoribus , ut humiliori vocabulo uterentur, ex τοῦ *dominus* factam esse vocem *dom nus* per syncopen. Quod ubi exemplis e Sev. Sulpi tio, et aliunde petitis probavit, addit apud Græcos

PATROL. XVI.

41

A

[a] sed superabundavit gratia. Itaque nihil habeo quod querar, et habeo inquo Deogratias agam; quia semper optavi, ut si quae perturbationes vet Ecclesiam vel me manerent, in me potius ac meam deciderent domum. Deo igitur gratias, quia in hoc omnium metu, cum omnia motibus sint suspecta Barbaricis, communem moerorem privato dolore transegi, et in nic conversum est [b] quidquid timebam omnibus. Atque utinam hic consummatum sit, ut dolor meas publici doloris redemptio sil 1
2. Nihil quidem habui, fratres charissimi, in rebus humanis tanto fratre pretiositis , hihil amabilius, nihil charius; sed praestant privatis publica. Ipsius quoque sententiam si quis sciscitaretur, [c] mallet occidl pro allis , quam sibi vivere; propterea enim

B

pro omnibus secundum carnem Christus est mortuus, ut itos non solis nobis vivere disceremus (ll) Cor. v, 15).
edit illud quod ingi atus divinitati esse non n : laetandum est euim magis quod talem fratreni im, quam dolendum quod fratrem amiserim ; im munus , hoc debitum est. Itaque Unctus sum, quamdiu licUit , 1114 commisso enOre : qui deposuit pignus, recepit. Nihil utrum abjures depositum, an doleas restitutum . que fidei ambiguum, a vitse periculum si pecuniam neges, culpa est : si hostiam [r] pietas est? cum pecuniæ fenerator illudi naturæ auctor, et necessitudinis creditor non queat. Itaque quanto uberior fenoris

A

[a] sed superabundavit gratia. Itaque nihil habeo quod querar, et habeo ✓ inquo Deogratias gam; quia semper optavi, ut si qua in quo Deo gratias Ecclesiam vel me manerent, in me potius ac meam deciderent domum. Deo igitur gratias, quia in hoc omnium metu, cum omnia motibus sint suspecta Barbaricis, communem moerorem privato dolore transegi, et in nic conversum est [b] quidquid timebam omnibus. Atque utinam hic consummatum sit, ut dolor meas

Alignment of editions

Three excerpted editions of Aeschylus' *Supplices* aligned.
www.i-alignment.com/Aeschylus

Future work

- Import and export options

- Language dependent options for Latin, Greek, Arabic

- Handling crossings and transpositions

Thanks for the attention!

Contact:
Tariq Yousef (tariq.yousef@dh.uni-leipzig.de)
Chiara Palladino (chiarapalladino1@gmail.com)