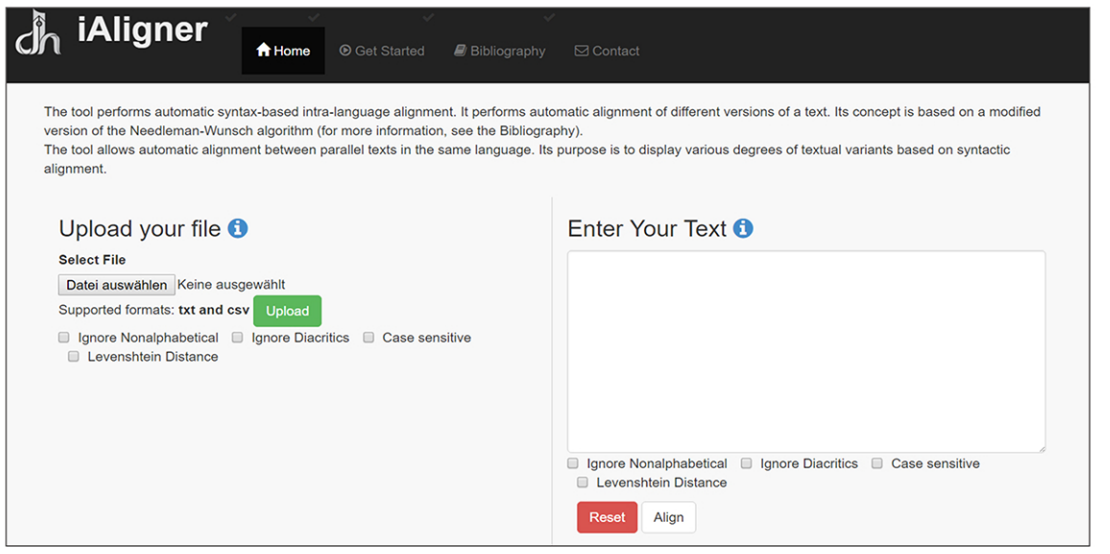
iAligner A tool for syntax-based intra-language text alignment

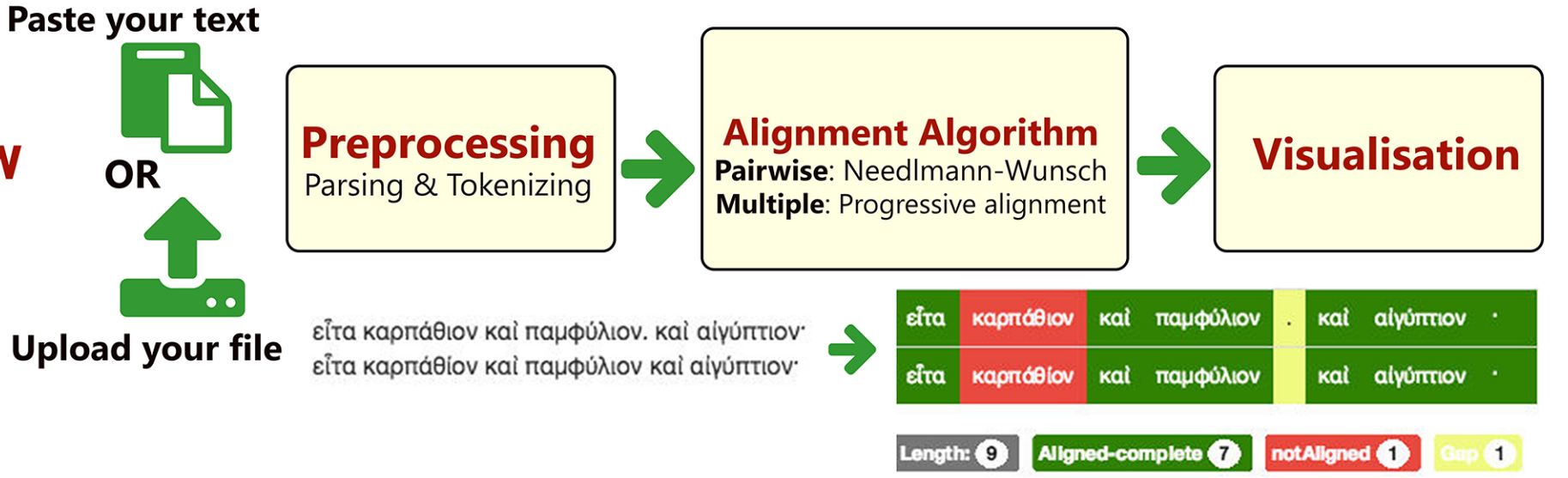
www.i-alignment.com

Tariq Yousef, University of Leipzig, tariq.yousef@uni-leipzig.de
Chiara Palladino, University of Leipzig & Bari, chiarapalladino1@gmail.com

Intra-language alignment is the alignment of texts in the same language. Algorithmic methods in this field have recently been applied in Textual Criticism, with the aim to detect textual variants across various witnesses, in order to support philological collation and the reconstruction of textual transmission. Current applications of alignment for semi-automated collation are particularly focused on the detection of multi-variants on modern texts, especially in cases where the authorial process is documented by multiple manuscript versions. However, ancient texts provide insights into a different situation, where authorial intervention can only be reconstructed from circumstantial evidence: in this case, the relations amongst various witnesses are not always clear, and instances of intertextual relations and reuse are decisive as well.



Workflow



User-refinement criteria

The tool provides additional refinement criteria, which can be chosen by the user:

> **Ignore nonalphabetical:** ignores symbols such as punctuation and numbers, anything that is not an alphabetical character.

> **Case sensitive:** detects variation between words in different cases.

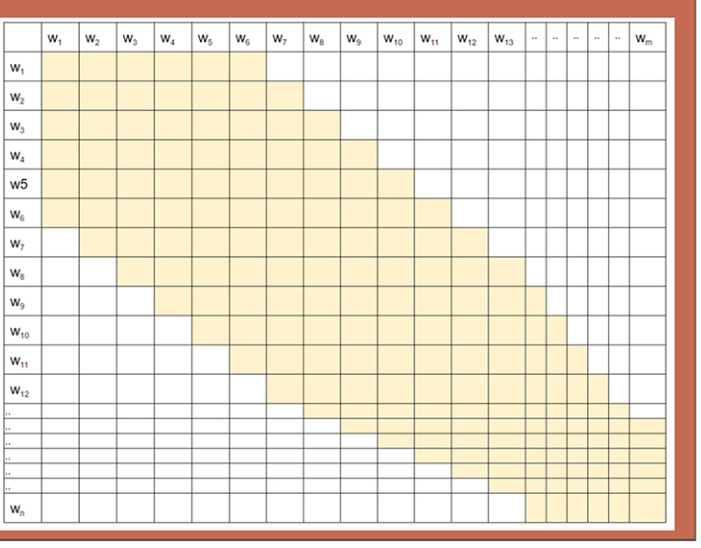
> **Ignore diacritics:** ignores any type of diacritical character, including punctuation.

> **Levensthein distance:** allows more tolerance on the alignment of similar words, based on a revised version of the Levensthein algorithm

Methodology

The main aim of the tool is to facilitate various degrees of textual comparison: in critical editorial practice, it allows the detection of manuscript variants across several witnesses, including non-literal variants in instances of textual re-use; it also provides comparison across multiple editions.

The alignment is performed through a modified version of the **Needleman-Wunsch algorithm.**
The algorithm is optimized by reducing the search space from n * m to 10 * m.



Alignment of Greek manuscripts

In the process of detection of manuscript variants, we did not apply any criteria of tolerance, in order to be able to individuate every minute difference between various exemplars: this approach was experimented on three manuscript witnesses of Plato's Crito (Clarkianus 39, Parisinus Graecus 1808, Tuebingensis Mb 14).

Future Work

In a further stage we are going to use the alignment workflow also for **post-correction of multiple OCR outputs**, i.e. texts resulting from Optical Character Recognition.
OCR on ancient Greek and Latin texts is still especially problematic, and it needs a stage of post-correction where three or more outputs of the same text are compared and the errors are manually removed. For this purpose, automatic alignment provides an ideal framework where variants in the outputs will be highlighted, making manual selection faster and easier. Moreover, we are going to provide a specific user environment where it will be also possible to choose the right variants.
We are currently experimenting this workflow on the OCR resulting from the massive digitization of the volumes of the **Patrologia Graeca**. We are also going to provide download options in XML format of the aligned text.

UNIVERSITÄT LEIPZIG

Digital Humanities
UNIVERSITÄT LEIPZIG

Quinto convegno annuale AIUCD
Fifth AIUCD Annual Conference 2016
7-9 settembre 2016 / Aula Magna S. Trentin, Ca' Dolfin, Dorsoduro 3825/e - 30123 Venezia