

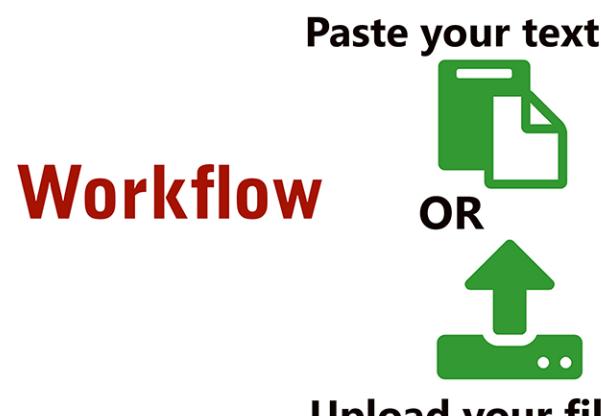
# iAligner A tool for syntax-based intra-language text alignment

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.

The screenshot shows the iAligner homepage. At the top, there's a navigation bar with links for Home, Get Started, Bibliography, and Contact. Below the navigation is a brief description of the tool's purpose: "The tool performs automatic syntax-based intra-language alignment. It performs automatic alignment of different versions of a text. Its concept is based on a modified version of the Needleman-Wunsch algorithm (for more information, see the Bibliography). The tool allows automatic alignment between parallel texts in the same language. Its purpose is to display various degrees of textual variants based on syntactic alignment." On the left, there's a section for uploading files, with a "Select File" button and options to ignore non-alphabetic characters, diacritics, or case sensitivity, along with a Levenshtein Distance checkbox. On the right, there's a text input field labeled "Enter Your Text" with a placeholder "Type your text here". Below the input field are several checkboxes for ignoring non-alphabetic characters, diacritics, case sensitivity, and Levenshtein Distance, followed by a "Reset" button and an "Align" button.



## User-refinement criteria

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

- > **Ignore nonalphabetic**: 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

τῆς δ' ἀσίας ἀπό κανάβου	ἐώς τανάδος ποταμοῦ	μετὰ τῶν κόλπων ὁ παράπλους	σταδίων μυριάδων δ και ρια
τῆς δέ ἀσίας ἀπό κανάβου	ἐώς τανάδος ποταμοῦ	μετὰ τῶν κόλπων ὁ παράπλους	σταδίων μυριάδων δ και ρια

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

## 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).

CLARK	46C	Paris1808	Tuebingen
#Σωκράτης και τιμώ, ούστερ και πρότερον: ὃν ἔν τοι παρόντι, εύ ισθι. διν οὐ μη βελτίω ξύμψουν λέγεν ἐν τοῖς παρόντι, εύ ισθι. διν οὐ μη σοι συγχωρήσουν οὐδὲ ὅπλει τῶν νῦν παρόντων, ἢ τὸν πολλὸν δύναμις: ὡστερ πατέσσας ήμας μορμολύπτηται δεσμούς, καὶ θανάτους ἐπιτέμπουσα- καὶ χρημάτων φραρέσες: πάς σύν μεριάστασα σκοπομέθα αύτοι: εἰ πράτον μὲν τοῖσιν τὸν λόγον ἀνάλθουμεν ὃν σύ λέγεις περὶ τῶν δοῦλων πότερον καλῶς ἐλέγετο ἐκάστοτε, ἥ οὐ δη ταῖς δέ τοι δοῦλων προσέσθεν τὸν νοῦν:	#Σωκράτης και τιμώ, ούστερ και πρότερον: ὃν ἔν τοι παρόντι, εύ ισθι. διν οὐ μη βελτίω ξύμψουν λέγεν ἐν τοῖς παρόντι, εύ ισθι. διν οὐ μη σοι συγχωρήσουν οὐδὲ ὅπλει τῶν νῦν παρόντων, ἢ τὸν πολλὸν δύναμις: ὡστερ πατέσσας ήμας μορμολύπτηται δεσμούς, καὶ θανάτους ἐπιτέμπουσα- καὶ χρημάτων φραρέσες: πάς σύν μεριάστασα σκοπομέθα αύτοι: εἰ πράτον μὲν τοῖσιν τὸν λόγον ἀνάλθουμεν ὃν σύ λέγεις περὶ τῶν δοῦλων πότερον καλῶς ἐλέγετο ἐκάστοτε, ἥ οὐ δη ταῖς δέ τοι δοῦλων προσέσθεν τὸν νοῦν:	#Σωκράτης και τιμώ, ούστερ και πρότερον: ὃν ἔν τοι παρόντι, εύ ισθι. διν οὐ μη βελτίω ξύμψουν λέγεν ἐν τοῖς παρόντι, εύ ισθι. διν οὐ μη σοι συγχωρήσουν οὐδὲ ὅπλει τῶν νῦν παρόντων, ἢ τὸν πολλὸν δύναμις: ὡστερ πατέσσας ήμας μορμολύπτηται δεσμούς, καὶ θανάτους ἐπιτέμπουσα- καὶ χρημάτων φραρέσες: πάς σύν μεριάστασα σκοπομέθα αύτοι: εἰ πράτον μὲν τοῖσιν τὸν λόγον ἀνάλθουμεν ὃν σύ λέγεις περὶ τῶν δοῦλων πότερον καλῶς ἐλέγετο ἐκάστοτε, ἥ οὐ δη ταῖς δέ τοι δοῦλων προσέσθεν τὸν νοῦν:	#Σωκράτης και τιμώ, ούστερ και πρότερον: ὃν ἔν τοι παρόντι, εύ ισθι. διν οὐ μη βελτίω ξύμψουν λέγεν ἐν τοῖς παρόντι, εύ ισθι. διν οὐ μη σοι συγχωρήσουν οὐδὲ ὅπλει τῶν νῦν παρόντων, ἢ τὸν πολλὸν δύναμις: ὡστερ πατέσσας ήμας μορμολύπτηται δεσμούς, καὶ θανάτους ἐπιτέμπουσα- καὶ χρημάτων φραρέσες: πάς σύν μεριάστασα σκοπομέθα αύτοι: εἰ πράτον μὲν τοῖσιν τὸν λόγον ἀνάλθουμεν ὃν σύ λέγεις περὶ τῶν δοῦλων πότερον καλῶς ἐλέγετο ἐκάστοτε, ἥ οὐ δη ταῖς δέ τοι δοῦλων προσέσθεν τὸν νοῦν:

#Σωκράτης και τιμώ, ούστερ και πρότερον: πάς σύν μεριάστασα αύτοι: εἰ πράτον μὲν τοῖσιν τὸν λόγον ἀνάλθουμεν ὃν σύ λέγεις περὶ τῶν δοῦλων πότερον καλῶς ἐλέγετο ἐκάστοτε, ἥ οὐ δη ταῖς δέ τοι δοῦλων προσέσθεν τὸν νοῦν:

#Σωκράτης και τιμώ, ούστερ και πρότερον: πάς σύν μεριάστασα αύτοι: εἰ πράτον μὲν τοῖσιν τὸν λόγον ἀνάλθουμεν ὃν σύ λέγεις περὶ τῶν δοῦλων πότερον καλῶς ἐλέγετο ἐκάστοτε, ἥ οὐ δη ταῖς δέ τοι δοῦλων προσέσθεν τὸν νοῦν:

#Σωκράτης και τιμώ, ούστερ και πρότερον: πάς σύν μεριάστασα αύτοι: εἰ πράτον μὲν τοῖσιν τὸν λόγον ἀνάλθουμεν ὃν σύ λέγεις περὶ τῶν δοῦλων πότερον καλῶς ἐλέγετο ἐκάστοτε, ἥ οὐ δη ταῖς δέ τοι δοῦλων προσέσθεν τὸν νοῦν:

#Σωκράτης και τιμώ, ούστερ και πρότερον: πάς σύν μεριάστασα αύτοι: εἰ πράτον μὲν τοῖσιν τὸν λόγον ἀνάλθουμεν ὃν σύ λέγεις περὶ τῶν δοῦλων πότερον καλῶς ἐλέγετο ἐκάστοτε, ἥ οὐ δη ταῖς δέ τοι δοῦλων προσέσθεν τὸν νοῦν:

#Σωκράτης και τιμώ, ούστερ και πρότερον: πάς σύν μεριάστασα αύτοι: εἰ πράτον μὲν τοῖσιν τὸν λόγον ἀνάλθουμεν ὃν σύ λέγεις περὶ τῶν δοῦλων πότερον καλῶς ἐλέγετο ἐκάστοτε, ἥ οὐ δη ταῖς δέ τοι δοῦλων προσέσθεν τὸν νοῦν:

## 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  $10 * m$ .

## 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