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

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

1 Introduction

The aim of the poster is to introduce an in-development tool for intra-language and syntax-based text alignment.

Intra-language alignment is the alignment of texts in the same language. The topic was first raised within the Hear Homer project at DEVLAB (Haentjens Dekker et al. 2014). Recently, intra-language alignment methods have been applied in the field of Textual Criticism, with the aim to detect textual variants across various witnesses, in order to support the philological process of collation (Makedon 1998) and the reconstruction of textual transmission (West 1973). Current applications of alignment for semi-automated collation are particularly focused on the detection of multi-variants on “alive” texts, i.e. 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 tool is available as a web service at the address `http://www.i-alignment.com`. The actionable Python code is also provided in the GitHub repository (`https://github.com/OpenGreekAndLatin/ILA_python`).

2 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 (Needleman and Wunsch 1970), also used in Bioinformatics to perform optimal alignment of DNA sequences. The algorithm is optimized by reducing the search space: given two sentences **S1** and **S2** with length **n** and **m** respectively, the algorithm in its basic form compares each word of **S1** with each word of **S2**, producing a search space = **n * m**. As we do not need to compare each word of **S1** with each word of **S2**, our algorithm compares a word **W** in **S1** with words **[W-5, W+5]** in **S2**. Therefore, the search space is reduced from **n * m** to **10 * m** (Fig. 1). Various language-dependent refinement options are additionally chosen by the user: diacritics and punctuation can be detected as single tokens or ignored, and Levensthein distance metric can be applied to adjust the tolerance threshold, in order to restrict or amplify the tolerance in the detection of variants.

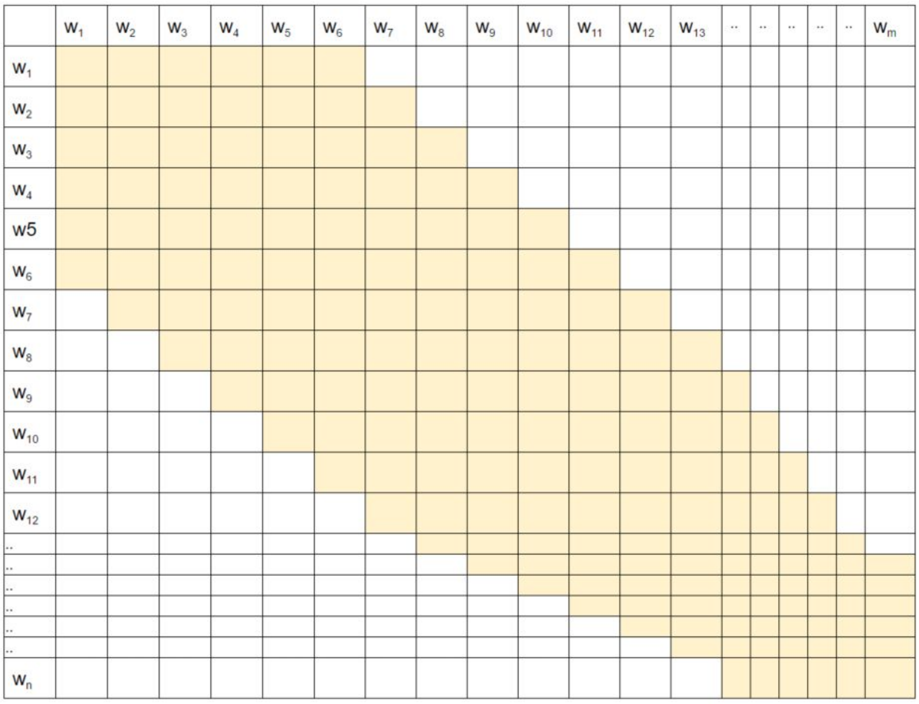


Figure 1: Reduction of the search space in the optimized Needleman-Wunsch algorithm. The white cells represent the search space of the algorithm in its normal form, the yellow cells the optimized search space.

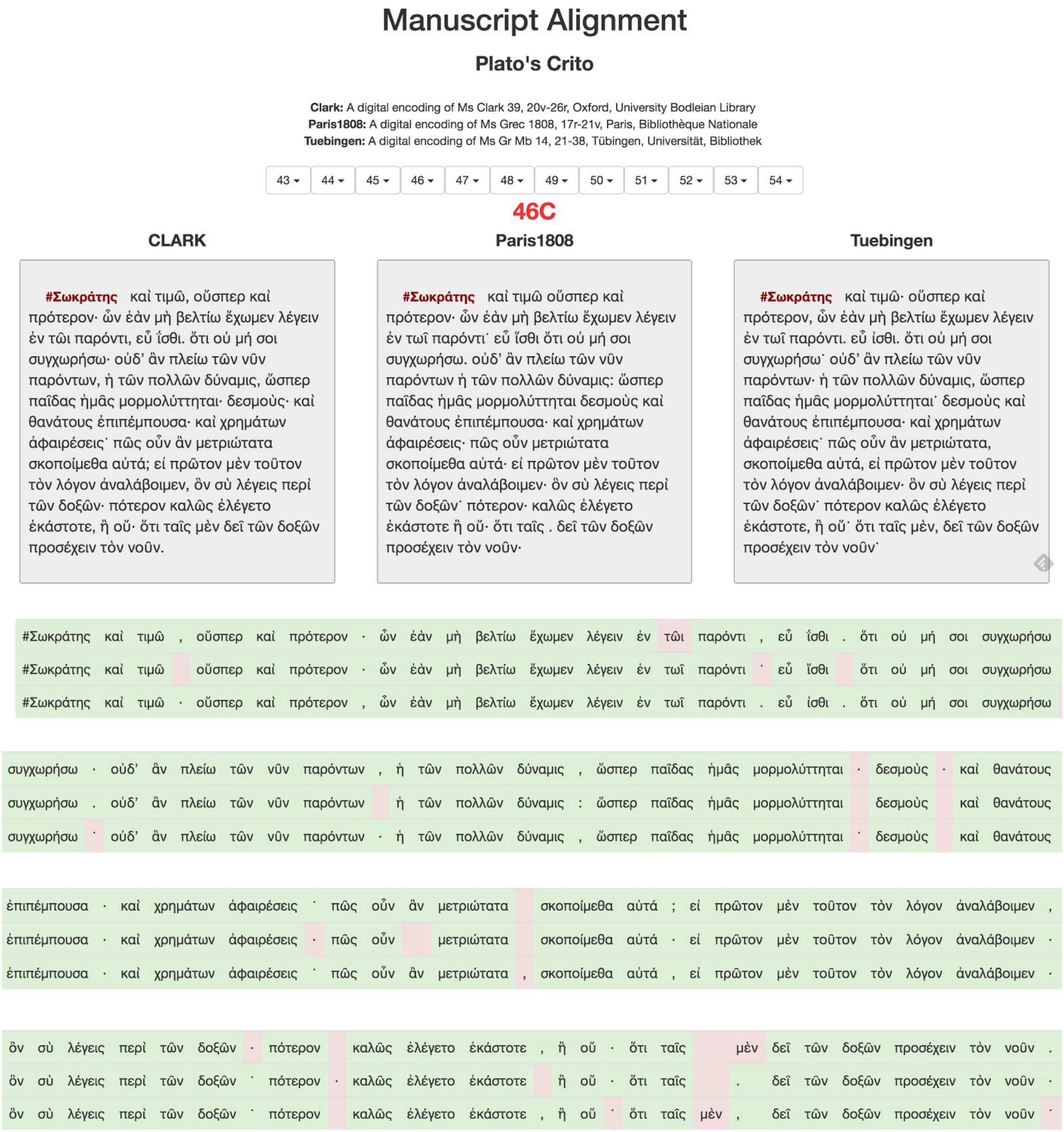
3 Workflow

The alignment can be performed either by uploading the text in tabular CSV format, or by pasting chosen junks in plain format. Further development will also allow direct upload of texts in XML and JSON format.

The uploaded file is parsed into a list of parallel sentences, which are then passed to tokenization and converted to a vector of single tokens: we used a simple tokenizer, which takes white spaces and punctuation marks as delimiters to split the sentence in a vector of single words. Once the parallel sentences are tokenized, they are processed by the alignment algorithm and further elaborated by the user according to the various refinement options. At present, the texts are supposed to be initially structured at paragraph or sentence level. Future implementations will also allow a preprocessing stage for initially not aligned texts, by means of algorithms on length-based methods (Thompson 1994).

4 Current results

The several refinement criteria are provided to the user in order to allow a specific performance of the algorithm, according to the purpose: for example, for the 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* (Fig. 2).



Manuscript Alignment

Plato's Crito

Clark: A digital encoding of Ms Clark 39, 20v-26r, Oxford, University Bodleian Library
Paris1808: A digital encoding of Ms Grec 1808, 17r-21v, Paris, Bibliothèque Nationale
Tuebingen: A digital encoding of Ms Gr Mb 14, 21-38, Tübingen, Universität, Bibliothek

46C

CLARK

#Σωκράτης καὶ τιμῶ, οὕσπερ καὶ πρότερον· ὧν ἐὰν μὴ βελτίω ἔχωμεν λέγειν ἐν τῶι παρόντι, εὖ ἴσθι. ὅτι οὐ μή σοι συγχωρήσω· οὐδ’ ἂν πλείω τῶν νῦν παρόντων, ἡ τῶν πολλῶν δύναμις, ὥσπερ παῖδας ἡμᾶς μορμολύττηται· δεσμοὺς· καὶ θανάτους ἐπιπέμπουσα· καὶ χρημάτων ἀφαιρέσεις˙ πῶς οὖν ἂν μετριώτατα σκοποίμεθα αὐτά; εἰ πρῶτον μὲν τοῦτον τὸν λόγον ἀναλάβοιμεν, ὃν σὺ λέγεις περὶ τῶν δοξῶν· πότερον καλῶς ἐλέγετο ἑκάστοτε, ἢ οὔ· ὅτι ταῖς μὲν δεῖ τῶν δοξῶν προσέχειν τὸν νοῦν.

Paris1808

#Σωκράτης καὶ τιμῶ οὕσπερ καὶ πρότερον· ὧν ἐὰν μὴ βελτίω ἔχωμεν λέγειν ἐν τωῖ παρόντι˙ εὖ ἴσθι ὅτι οὐ μή σοι συγχωρήσω. οὐδ’ ἂν πλείω τῶν νῦν παρόντων ἡ τῶν πολλῶν δύναμις: ὥσπερ παῖδας ἡμᾶς μορμολύττηται δεσμοὺς καὶ θανάτους ἐπιπέμπουσα· καὶ χρημάτων ἀφαιρέσεις· πῶς οὖν μετριώτατα σκοποίμεθα αὐτά· εἰ πρῶτον μὲν τοῦτον τὸν λόγον ἀναλάβοιμεν· ὃν σὺ λέγεις περὶ τῶν δοξῶν˙ πότερον· καλῶς ἐλέγετο ἑκάστοτε ἢ οὔ· ὅτι ταῖς . δεῖ τῶν δοξῶν προσέχειν τὸν νοῦν·

Tuebingen

#Σωκράτης καὶ τιμῶ· οὕσπερ καὶ πρότερον, ὧν ἐὰν μὴ βελτίω ἔχωμεν λέγειν ἐν τωῖ παρόντι. εὖ ἰσθι. ὅτι οὐ μή σοι συγχωρήσω˙ οὐδ’ ἂν πλείω τῶν νῦν παρόντων· ἡ τῶν πολλῶν δύναμις, ὥσπερ παῖδας ἡμᾶς μορμολύττηται˙ δεσμοὺς καὶ θανάτους ἐπιπέμπουσα· καὶ χρημάτων ἀφαιρέσεις˙ πῶς οὖν ἂν μετριώτατα, σκοποίμεθα αὐτά, εἰ πρῶτον μὲν τοῦτον τὸν λόγον ἀναλάβοιμεν· ὃν σὺ λέγεις περὶ τῶν δοξῶν˙ πότερον καλῶς ἐλέγετο ἑκάστοτε, ἢ οὔ˙ ὅτι ταῖς μὲν, δεῖ τῶν δοξῶν προσέχειν τὸν νοῦν˙

#Σωκράτης	καὶ	τιμῶ	,	οὕσπερ	καὶ	πρότερον	·	ὧν	ἐὰν	μὴ	βελτίω	ἔχωμεν	λέγειν	ἐν	τῶι	παρόντι	,	εὖ	ἴσθι	.	ὅτι	οὐ	μή	σοι	συγχωρήσω
#Σωκράτης	καὶ	τιμῶ		οὕσπερ	καὶ	πρότερον	·	ὧν	ἐὰν	μὴ	βελτίω	ἔχωμεν	λέγειν	ἐν	τωῖ	παρόντι	˙	εὖ	ἴσθι		ὅτι	οὐ	μή	σοι	συγχωρήσω
#Σωκράτης	καὶ	τιμῶ	·	οὕσπερ	καὶ	πρότερον	,	ὧν	ἐὰν	μὴ	βελτίω	ἔχωμεν	λέγειν	ἐν	τωῖ	παρόντι	.	εὖ	ἰσθι	.	ὅτι	οὐ	μή	σοι	συγχωρήσω

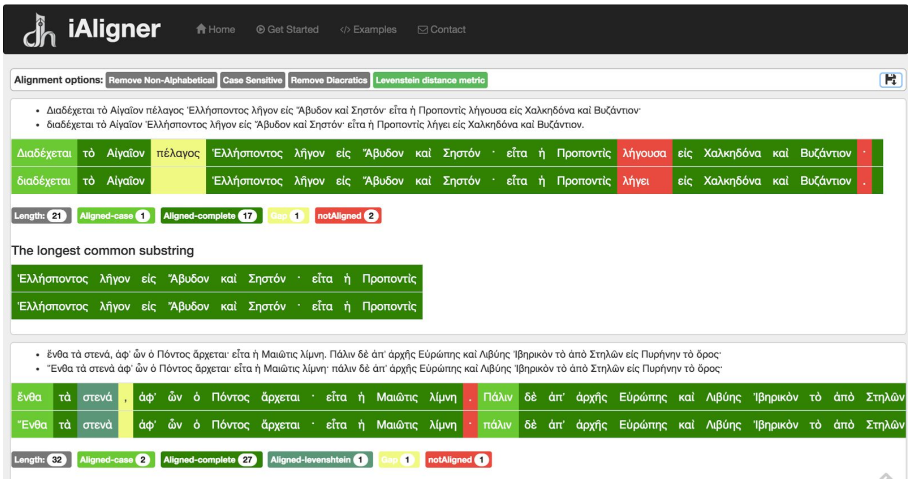
συγχωρήσω	·	οὐδ’	ἂν	πλείω	τῶν	νῦν	παρόντων	,	ἡ	τῶν	πολλῶν	δύναμις	,	ὥσπερ	παῖδας	ἡμᾶς	μορμολύττηται	·	δεσμοὺς	·	καὶ	θανάτους
συγχωρήσω	.	οὐδ’	ἂν	πλείω	τῶν	νῦν	παρόντων		ἡ	τῶν	πολλῶν	δύναμις	:	ὥσπερ	παῖδας	ἡμᾶς	μορμολύττηται		δεσμοὺς		καὶ	θανάτους
συγχωρήσω	˙	οὐδ’	ἂν	πλείω	τῶν	νῦν	παρόντων	·	ἡ	τῶν	πολλῶν	δύναμις	,	ὥσπερ	παῖδας	ἡμᾶς	μορμολύττηται	˙	δεσμοὺς		καὶ	θανάτους

ἐπιπέμπουσα	·	καὶ	χρημάτων	ἀφαιρέσεις	˙	πῶς	οὖν	ἂν	μετριώτατα		σκοποίμεθα	αὐτά	;	εἰ	πρῶτον	μὲν	τοῦτον	τὸν	λόγον	ἀναλάβοιμεν	,
ἐπιπέμπουσα	·	καὶ	χρημάτων	ἀφαιρέσεις	·	πῶς	οὖν		μετριώτατα		σκοποίμεθα	αὐτά	·	εἰ	πρῶτον	μὲν	τοῦτον	τὸν	λόγον	ἀναλάβοιμεν	·
ἐπιπέμπουσα	·	καὶ	χρημάτων	ἀφαιρέσεις	˙	πῶς	οὖν	ἂν	μετριώτατα	,	σκοποίμεθα	αὐτά	,	εἰ	πρῶτον	μὲν	τοῦτον	τὸν	λόγον	ἀναλάβοιμεν	·

ὃν	σὺ	λέγεις	περὶ	τῶν	δοξῶν	·	πότερον		καλῶς	ἐλέγετο	ἑκάστοτε	,	ἢ	οὔ	·	ὅτι	ταῖς		μὲν	δεῖ	τῶν	δοξῶν	προσέχειν	τὸν	νοῦν	.
ὃν	σὺ	λέγεις	περὶ	τῶν	δοξῶν	˙	πότερον	·	καλῶς	ἐλέγετο	ἑκάστοτε		ἢ	οὔ	·	ὅτι	ταῖς		.	δεῖ	τῶν	δοξῶν	προσέχειν	τὸν	νοῦν	·
ὃν	σὺ	λέγεις	περὶ	τῶν	δοξῶν	˙	πότερον		καλῶς	ἐλέγετο	ἑκάστοτε	,	ἢ	οὔ	˙	ὅτι	ταῖς	μὲν	,	δεῖ	τῶν	δοξῶν	προσέχειν	τὸν	νοῦν	˙

Figure 2: Alignment of manuscripts of Plato’s *Crito*.

On the other hand, user refinement criteria proved to be essential for more complex instances. We applied the workflow to the in-progress born-digital critical edition of Agathemerus’ *Sketch of Geography*, a Greek geographical work whose transmission offers a good and manageable example of case studies. Particular attention has been paid to the relationship between the manuscript tradition and three long excerpts of the same text found in the Expositio Fidei by the Syriac Patristic writer John of Damascus (†750 ca.). Being this tradition older than the direct witnesses, its textual contribution was particularly relevant. In this case, we experimented various options to test how the algorithm performed. The inspection of Excerpt b revealed a surprising correspondence between direct and indirect tradition [Image 3]: the application of Levensthein distance did not change the results of the alignment, as the text had already a good rate of overlap. On the other hand the Excerpt a [Image 4], was more problematic, being the only instance where a consistent section of the work is not transmitted in the indirect tradition, and where important differences between the two versions have been highlighted. Some cases of minor crossing were not handled properly because of the syntactical base of the algorithm, revealing a limitation in its application, which needs to be faced in future improvements.



Alignment options: Remove Non-Alphabetical Case Sensitive Remove Diacratics Levenstein distance metric

- Διαδέχεται τὸ Αἰγαῖον πέλαγος Ἑλλήσποντος λῆγον εἰς Ἄβυδον καὶ Σηστόν· εἶτα ἡ Προποντὶς λήγουσα εἰς Χαλκηδόνα καὶ Βυζάντιον·
- διαδέχεται τὸ Αἰγαῖον Ἑλλήσποντος λῆγον εἰς Ἄβυδον καὶ Σηστόν· εἶτα ἡ Προποντὶς λήγει εἰς Χαλκηδόνα καὶ Βυζάντιον.

Διαδέχεται	τὸ	Αἰγαῖον	πέλαγος	Ἑλλήσποντος	λῆγον	εἰς	Ἄβυδον	καὶ	Σηστόν	·	εἶτα	ἡ	Προποντὶς	λήγουσα	εἰς	Χαλκηδόνα	καὶ	Βυζάντιον	·
διαδέχεται	τὸ	Αἰγαῖον		Ἑλλήσποντος	λῆγον	εἰς	Ἄβυδον	καὶ	Σηστόν	·	εἶτα	ἡ	Προποντὶς	λήγει	εἰς	Χαλκηδόνα	καὶ	Βυζάντιον	.

Length: 21 Aligned-case 1 Aligned-complete 17 Gap 1 notAligned 2

The longest common substring

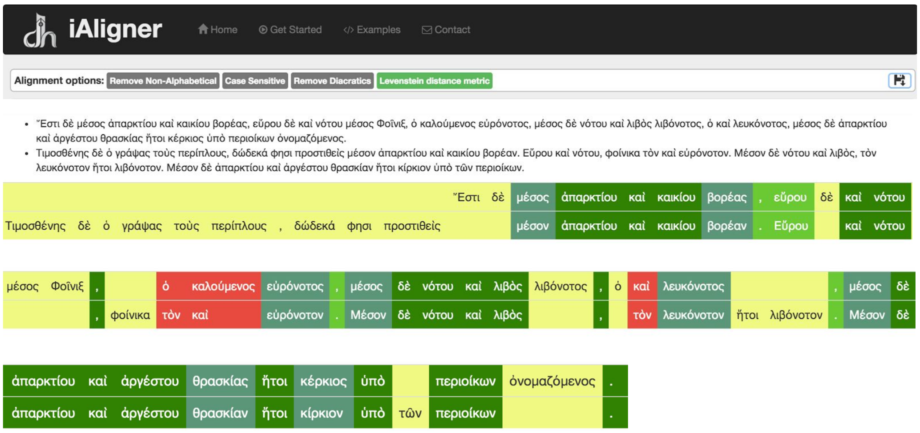
Ἑλλήσποντος	λῆγον	εἰς	Ἄβυδον	καὶ	Σηστόν	·	εἶτα	ἡ	Προποντὶς
Ἑλλήσποντος	λῆγον	εἰς	Ἄβυδον	καὶ	Σηστόν	·	εἶτα	ἡ	Προποντὶς

- ἔνθα τὰ στενά, ἀφ' ὧν ὁ Πόντος ἄρχεται· εἶτα ἡ Μαιῶτις λίμνη. Πάλιν δὲ ἀπ' ἀρχῆς Εὐρώπης καὶ Λιβύης Ἰβηρικὸν τὸ ἀπὸ Στηλῶν εἰς Πυρήνην τὸ ὄρος·
- Ἔνθα τὰ στενὰ ἀφ' ὧν ὁ Πόντος ἄρχεται· εἶτα ἡ Μαιῶτις λίμνη· πάλιν δὲ ἀπ' ἀρχῆς Εὐρώπης καὶ Λιβύης Ἰβηρικὸν τὸ ἀπὸ Στηλῶν εἰς Πυρήνην τὸ ὄρος·

ἔνθα	τὰ	στενά	,	ἀφ'	ὧν	ὁ	Πόντος	ἄρχεται	·	εἶτα	ἡ	Μαιῶτις	λίμνη	.	Πάλιν	δὲ	ἀπ'	ἀρχῆς	Εὐρώπης	καὶ	Λιβύης	Ἰβηρικὸν	τὸ	ἀπὸ	Στηλῶν
Ἔνθα	τὰ	στενὰ		ἀφ'	ὧν	ὁ	Πόντος	ἄρχεται	·	εἶτα	ἡ	Μαιῶτις	λίμνη	·	πάλιν	δὲ	ἀπ'	ἀρχῆς	Εὐρώπης	καὶ	Λιβύης	Ἰβηρικὸν	τὸ	ἀπὸ	Στηλῶν

Length: 32 Aligned-case 2 Aligned-complete 27 Aligned-levenshtein 1 Gap 1 notAligned 1

Figure 3: An extract from alignment performed on Excerpt b: the first sentence comes from the text of John of Damascus, the second from the *Sketch of Geography*.



Alignment options: Remove Non-Alphabetical Case Sensitive Remove Diacratics Levenstein distance metric

- Ἔστι δὲ μέσος ἀπαρκτίου καὶ καικίου βορέας, εὔρου δὲ καὶ νότου μέσος Φοῖνιξ, ὁ καλούμενος εὐρόνοτος, μέσος δὲ νότου καὶ λιβὸς λιβόνοτος, ὁ καὶ λευκόνοτος, μέσος δὲ ἀπαρκτίου καὶ ἀργέστου θρασκίας ἤτοι κέρκιος ὑπὸ περιοίκων ὀνομαζόμενος.
- Τιμοσθένης δὲ ὁ γράψας τοὺς περίπλους, δώδεκά φησι προστιθεὶς μέσον ἀπαρκτίου καὶ καικίου βορέαν. Εὔρου καὶ νότου, φοίνικα τὸν καὶ εὐρόνοτον. Μέσον δὲ νότου καὶ λιβὸς, τὸν λευκόνοτον ἤτοι λιβόνοτον. Μέσον δὲ ἀπαρκτίου καὶ ἀργέστου θρασκίαν ἤτοι κίρκιον ὑπὸ τῶν περιοίκων.

									Ἔστι	δὲ	μέσος	ἀπαρκτίου	καὶ	καικίου	βορέας	,	εὔρου	δὲ	καὶ	νότου
Τιμοσθένης	δὲ	ὁ	γράψας	τοὺς	περίπλους	,	δώδεκά	φησι	προστιθεὶς		μέσον	ἀπαρκτίου	καὶ	καικίου	βορέαν	.	Εὔρου		καὶ	νότου

μέσος	Φοῖνιξ	,		ὁ	καλούμενος	εὐρόνοτος	,	μέσος	δὲ	νότου	καὶ	λιβὸς	λιβόνοτος	,	ὁ	καὶ	λευκόνοτος			,	μέσος	δὲ
		,	φοίνικα	τὸν	καὶ	εὐρόνοτον	.	Μέσον	δὲ	νότου	καὶ	λιβὸς		,		τὸν	λευκόνοτον	ἤτοι	λιβόνοτον	.	Μέσον	δὲ

ἀπαρκτίου	καὶ	ἀργέστου	θρασκίας	ἤτοι	κέρκιος	ὑπὸ		περιοίκων	ὀνομαζόμενος	.
ἀπαρκτίου	καὶ	ἀργέστου	θρασκίαν	ἤτοι	κίρκιον	ὑπὸ	τῶν	περιοίκων		.

Figure 4: An extract from aligned Excerpt a.

5 Future work

Current results encourage both applications on scholarly editorial practice and on larger efforts for the detection of a high amount of variants. A further stage is going to establish a workflow for automatic alignment of OCR outputs for postcorrection. Current limits in the alignment output, due to the use of a syntax-based algorithm, will also be addressed combining syntactic and semantic matching by means of lexical information (e.g. lemma, synonyms and PoS tagging).

Various options for the graphic display of the alignment are currently evaluated, including visualization of the syntax with different coloring options according to the user's refinement criteria, highlight of individual variants by means of graphs, and adaptations of customisable existing graphs for comparison based on one reference text (Jänicke 2014).

References

Brown, P. F. 1991. «Aligning sentences in parallel corpora». In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 169–176. Berkeley.

Haentjens Dekker, R., et al. 2014. «Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project». *Digital Scholarship in the Humanities*. doi:10.1093/llc/fqu007.

Jänicke, S. 2014. «Improving the Layout for Text Variant Graphs». In *VisLR workshop at LREC Conference*. Reykjavík.

Levenshtein, V. I. 1966. «Binary codes capable of correcting deletions, insertions, and reversals». *Soviet Physics Doklady* 10:707–710.

Makedon, F. 1998. «HEAR HOMER: A multimedia-data access remote prototype for ancient texts». In *Proceedings of ED-MEDIA'98*. Freiburg.

Needleman, S. B., and C. D. Wunsch. 1970. «A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins». *Journal of Molecular Biology* 48:443–453.

Thompson, J. 1994. «CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice». *Nucleic Acids Research* 22:4673–4680.

West, M. 1973. *Textual criticism and editorial technique*. Stuttgart: Teubner.