AI & DH (Part 2) @ Concordia January 16th

{LLMs for intertext detection}

Using LLMs on Greek and Latin

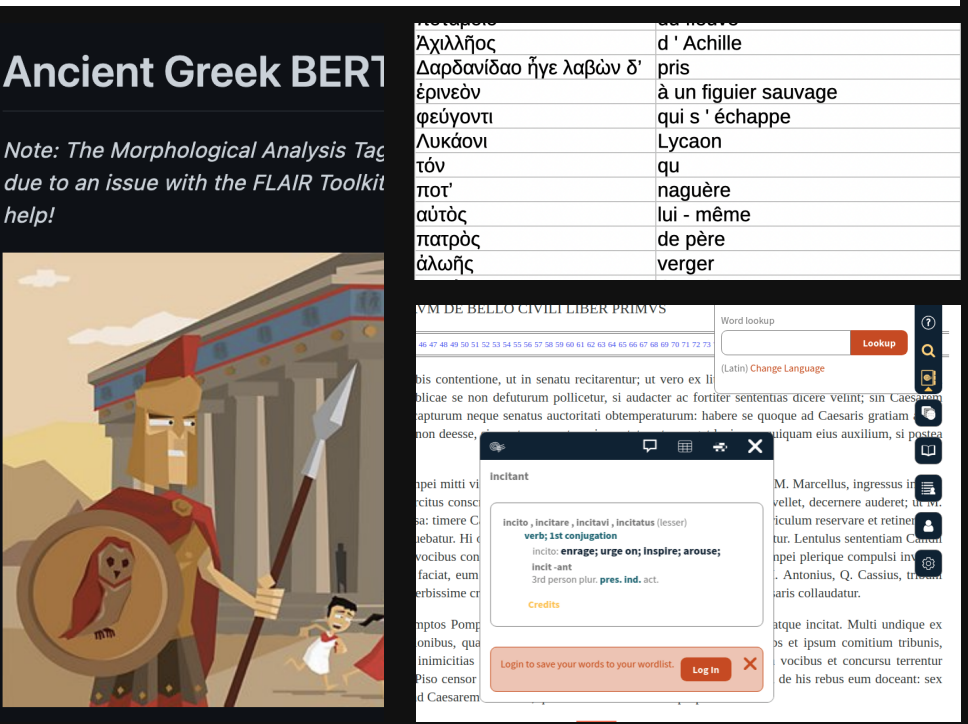
ENS DE LYON

ihrim

Université de Montréal

	Class	Frequency	3 Classes	Freque
yssey	Person	2.469	PER	2
	Place	698	LOC	
	Person	12.424	PER	12
	Place	2.305	LOC	2
	Ethnic	3.548	MISC	
	NoClass	2.263	MISC	
	Group	681	MISC	
pnosophists	Title	206	MISC	
	Festival	20	MISC	6
	Month	8	MISC	
	Language	7	MISC	
	Constellation	2	MISC	
al				24

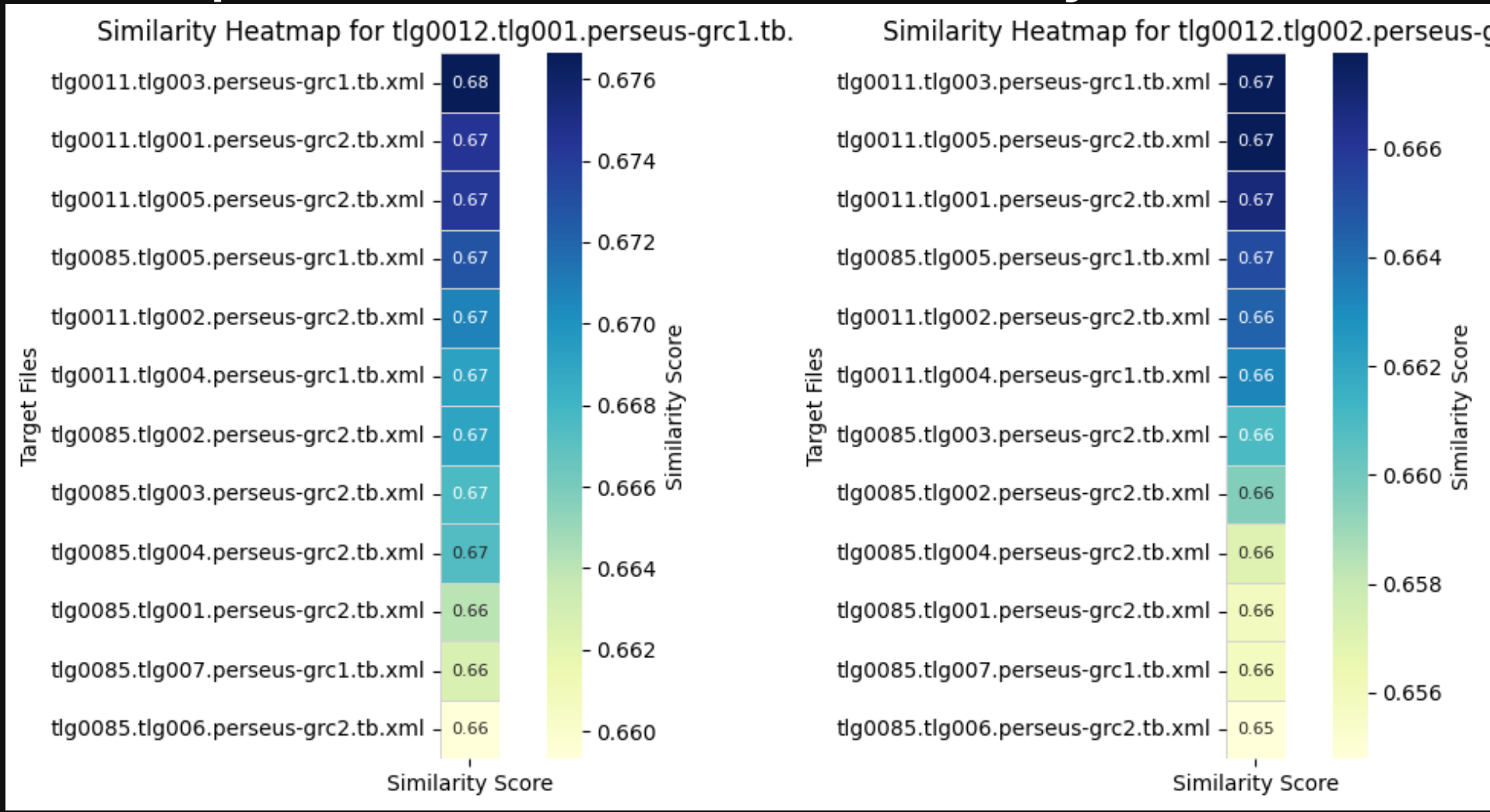
Table 1: An overview of the training data set.



Monolingual LLMs

Why ?

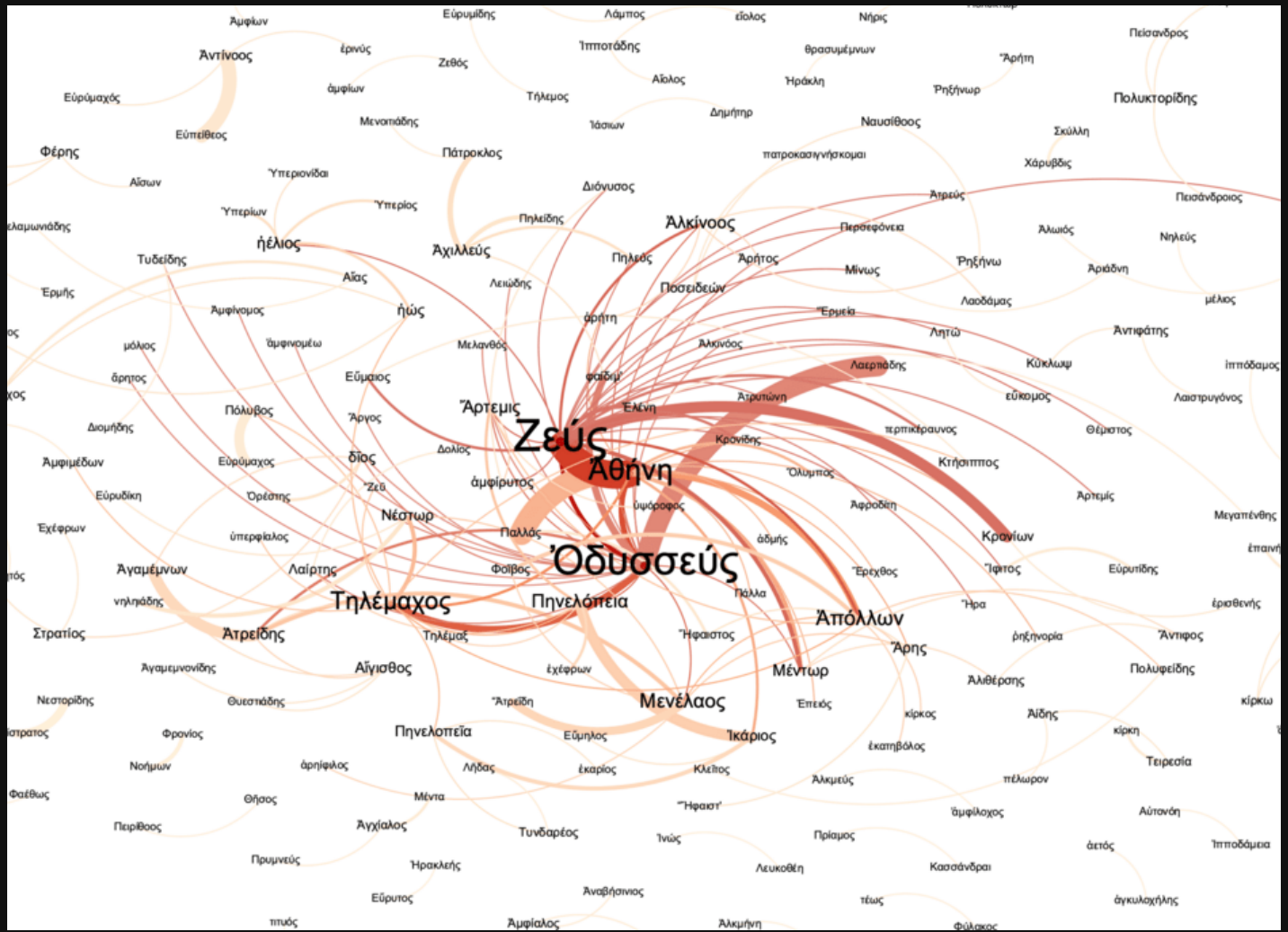
Example 1 : semantic similarity



https://github.com/OdysseusPolymetis/humanistica2023/blob/main/humanistica2023.ipynb

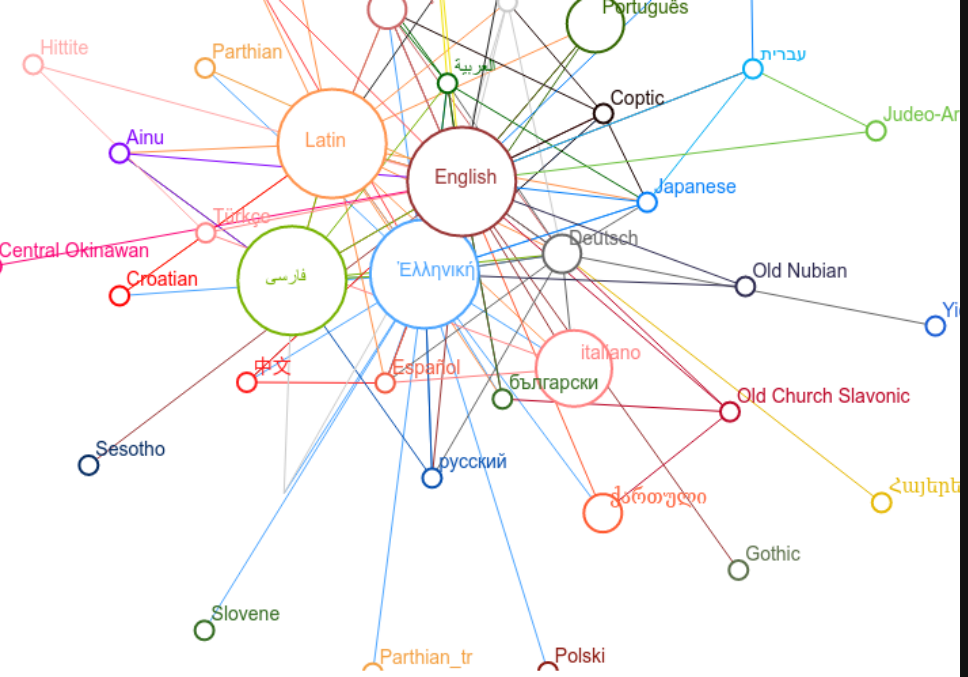
Example 2 : NER

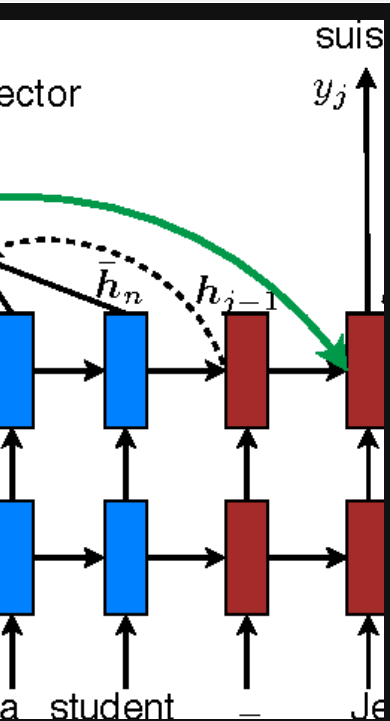
colab



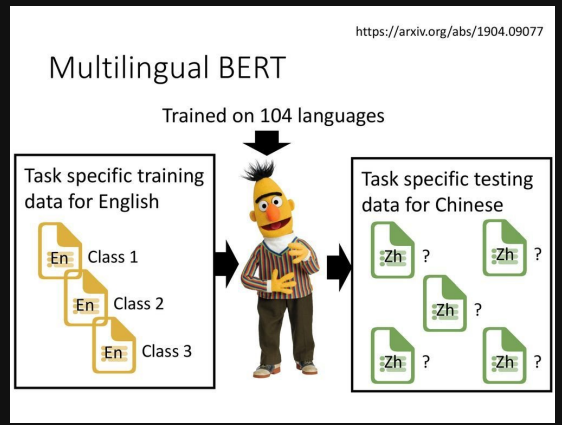
Example 3 : POStagging

https://github.com/OdysseusPolymetis/ia_et_shs/blob/main/nlp_greek_latin.ipynb





Ἀχιλλῆος	d ' Achille
Δαρδανίδαο ἥγε λαβὼν δ'	pris
ἐρινεὸν	à un figuier sauvage
φεύγοντι	qui s ' échappe
Λυκάονι	Lycaon
τόν	qu
ποτ'	naguère
αὐτὸς	lui - même
πατρὸς	de père
ἀλωῆς	verger



multilingual LLMs

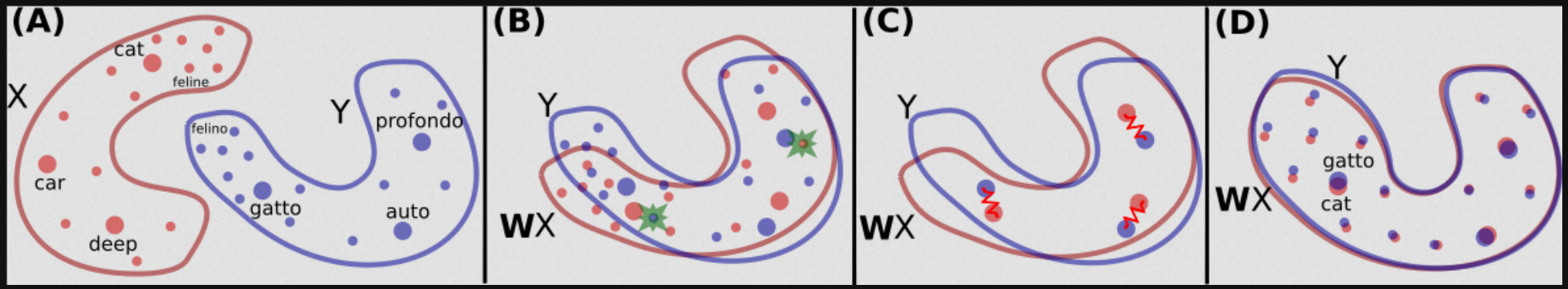
Even bigger problems

Why is it so complex ?

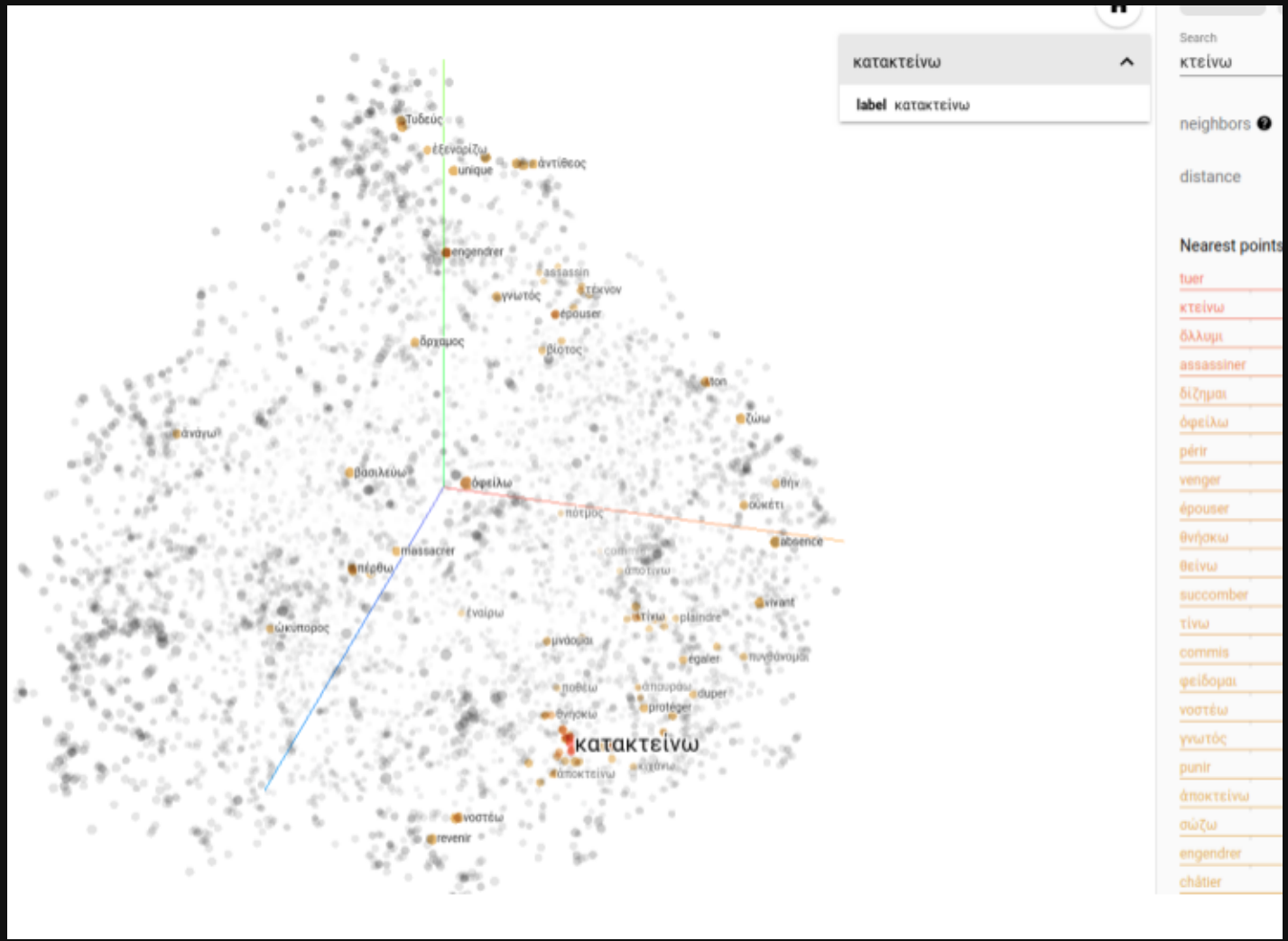
wmt19_translate/de-fr

- **Description** de la configuration : ensemble de données de tâche de traduction de-en WMT 2019.
- **Taille du téléchargement** : `9.71 GiB`

Solution 1 : mapping d'espaces monolingues



Solution 1 : mapping d'espaces monolingues



SOLUTION_2

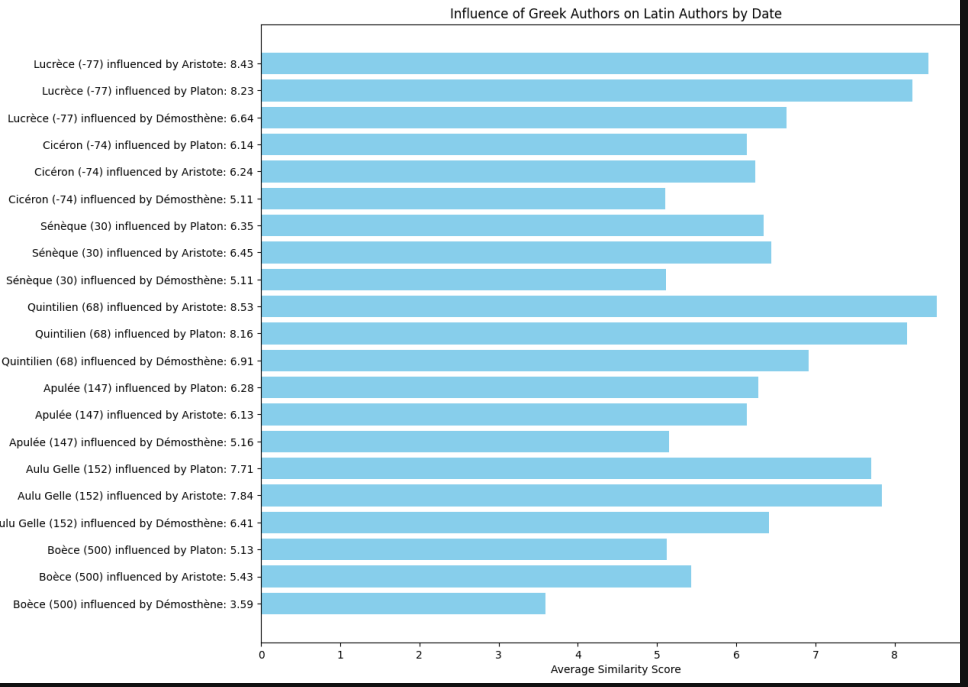
Similarité sémantique au token avec Multilingual BERT

```
greek_word = "λόγος"
greek_embedding = greek_filtered_embeddings[greek_word]

top_latin_words_with_scores = find_most_similar(greek_embedding, latin_filtered_embeddings)

for word, score in top_latin_words_with_scores:
    print(f"{word}: {score:.4f}")
```

```
oratio: 0.9894
τῷ: 0.9892
narratio: 0.9885
καὶ: 0.9878
liber: 0.9877
finis: 0.9877
scribo: 0.9877
loquor: 0.9875
ratio: 0.9875
argumentum: 0.9872
```




```
_grecs = [
halès", "tlg1705", -624, -546),
ythagore", "tlg0632", -570, -495),
éraclite", "tlg0626", -535, -475),
arménide", "tlg1562", -515, -450),
lcidamas", "tlg0610", -450, -400),
ntisthène", "tlg0591", -445, -365),
laton", "tlg0059", -427, -347),
émocrite", "tlg1304", -460, -370),
ristote", "tlg0086", -384, -322),
iogène de Sinope", "tlg1325", -413, -32
orgias de Léontium", "tlg0593", -480, -
émosthène", "tlg0014", -384, -322),
pictète", "tlg0557", 50, 135)

_latins = [
icéron", "phi0474", -106, -43),
ucrèce", "phi0550", -99, -55),
arron", "phi0684", -116, -27),
énèque", "phi1017", -4, 65),
énèque", "stoa0255", -4, 65),
énèque", "phi1014", -4, 65),
uintilien", "phi1002", 35, 100),
pulée", "phi1212", 124, 170),
ulu Gelle", "phi1254", 125, 180),
oèce", "stoa0058", 477, 524),
usonius Rufus", "tlg0628", 20, 101),
arc Aurèle", "tlg0562", 121, 180)
```

```
ile: tlg0086.tlg025.perseus-grc2.xml, Latin File: phi0550.phi001.perseus-la
 Phrase: δῆλον δὲ καὶ ἐκ τῶν τοιούτων λόγων·
 Phrase: id quod iam supra tibi paulo ostendimus ante.
arity Score: 0.9494874477386475

 Phrase: περὶ μὲν οὖν τούτων δεδήλωται καὶ πρότερον·
 Phrase: quo magis inceptum pergam pertexere dictis.
arity Score: 0.9490494132041931

 Phrase: ἔστι δὲ ἀπορία καὶ ἑτέρα περὶ αὐτῶν.
 Phrase: quae tibi posterius largo sermone probabo.
arity Score: 0.948077917098999

 Phrase: περὶ ὧν οὐδὲν ἧττον συμβαίνει τὰ αὐτὰ ἀπορεῖν.
 Phrase: id licet hinc quamvis hebeti cognoscere corde.
arity Score: 0.9477147459983826
```

8.12008

utation and Language

m victorem cepit. Detecting Latin Allusions to Ancie

Anette Frank

ivotal role in Classical Philology, with Latin authors frequently referencing Ancient Greek texts
xtual references has been constrained to monolingual approaches, seeking parallels solely wit
lingual Sentence-RoBERTa model tailored for Classical Philology, which excels at cross-lingua
ences across Ancient Greek, Latin, and English. We generate new training data by automatical
esent a case study, demonstrating SPhilBERTa's capability to facilitate automated detection of
https URL.

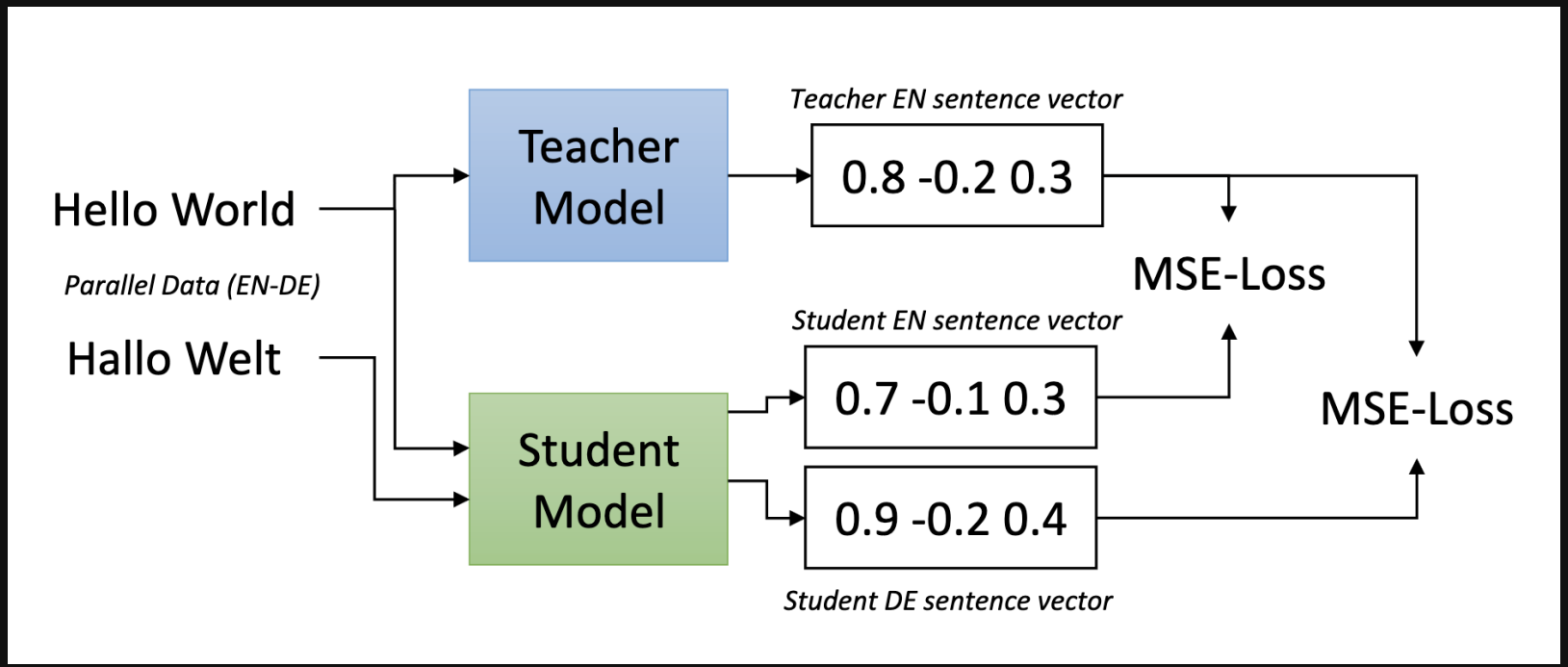
lication at the First Workshop on Ancient Language Processing (ALP) 2023; 9 pages, 5 tables
guage (cs.CL)

L]
[cs.CL] for this version)
50/arXiv.2308.12008

Experiment

Can we quantify proximity between Homer and Latin authors ?

The idea behind it :



Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Step 1 : corpus selection

Global open source data (First1KGreek, Perseus)

Homer
Iliad
Odyssey

~ 30 000 unique sentences

Virgil
Ovid
Horace
Lucan
Propertius
Terence
Cicero
Lucretius
Apuleius
Catullus
Juvenal
Plautus
and others

~ 100 000 unique sentences

Step 2 : topoi versus peculiarities

Docs	Word 1	Word 2	Word 3	Author
Doc1	3	0	1	A
Doc2	0	4	1	B

$$Eij=(Ni.\times N.j)/N$$

χ^2

Ni. : all words in class *i*.
N.j: all words *j* in all classes.
N : all words in all classes.

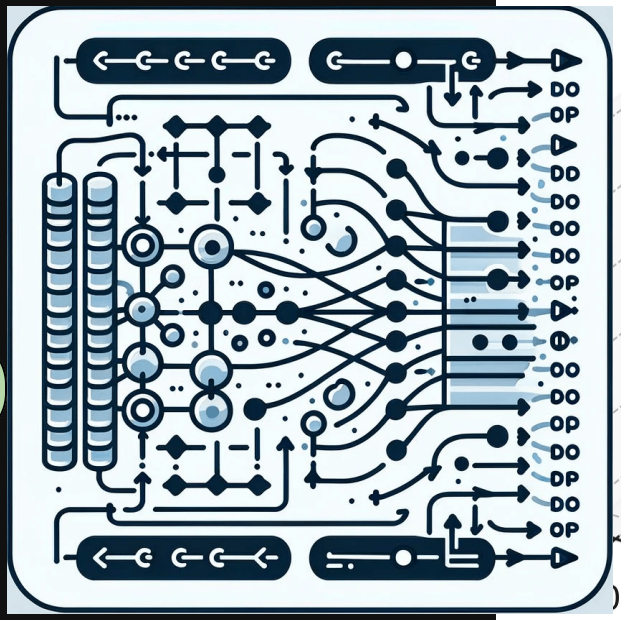
- Word 1 for author A : $E=(4*3)/9=1.33$.
- Word 1 for author B : $E=(5*3)/9=1.67$.

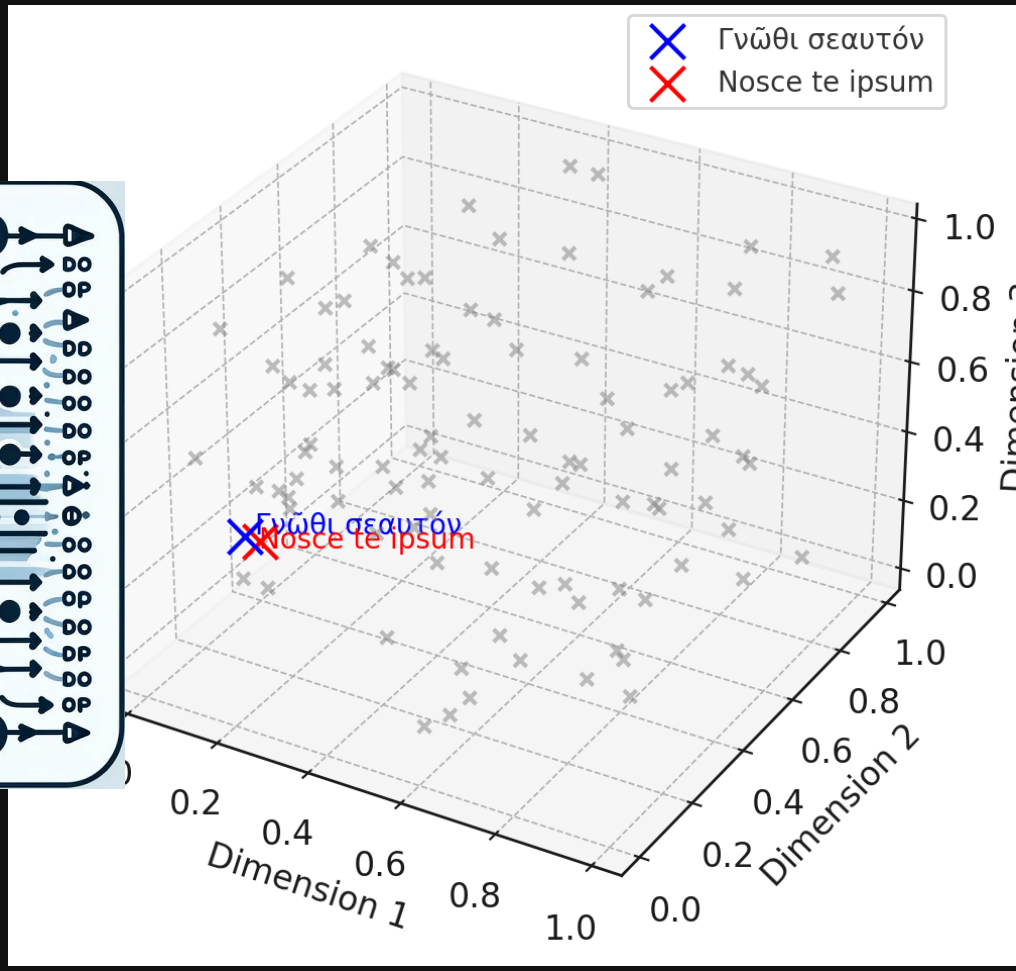
- $\chi^2 = \frac{(3-1.33)^2}{1.33} + \frac{(0-1.67)^2}{1.67} = \frac{2.77}{1.33} + \frac{2.79}{1.67} = 2.08 + 1.67 = 3.75$
- $\chi^2 = \frac{(0-1.78)^2}{1.78} + \frac{(4-2.22)^2}{2.22} = \frac{3.17}{1.78} + \frac{3.17}{2.22} = 1.78 + 1.43 = 3.21$
- $\chi^2 = \frac{(1-0.89)^2}{0.89} + \frac{(1-1.11)^2}{1.11} = \frac{0.0121}{0.89} + \frac{0.0121}{1.11} = 0.014 + 0.011 = 0.025$

Step 4 : encoding

latin sentence

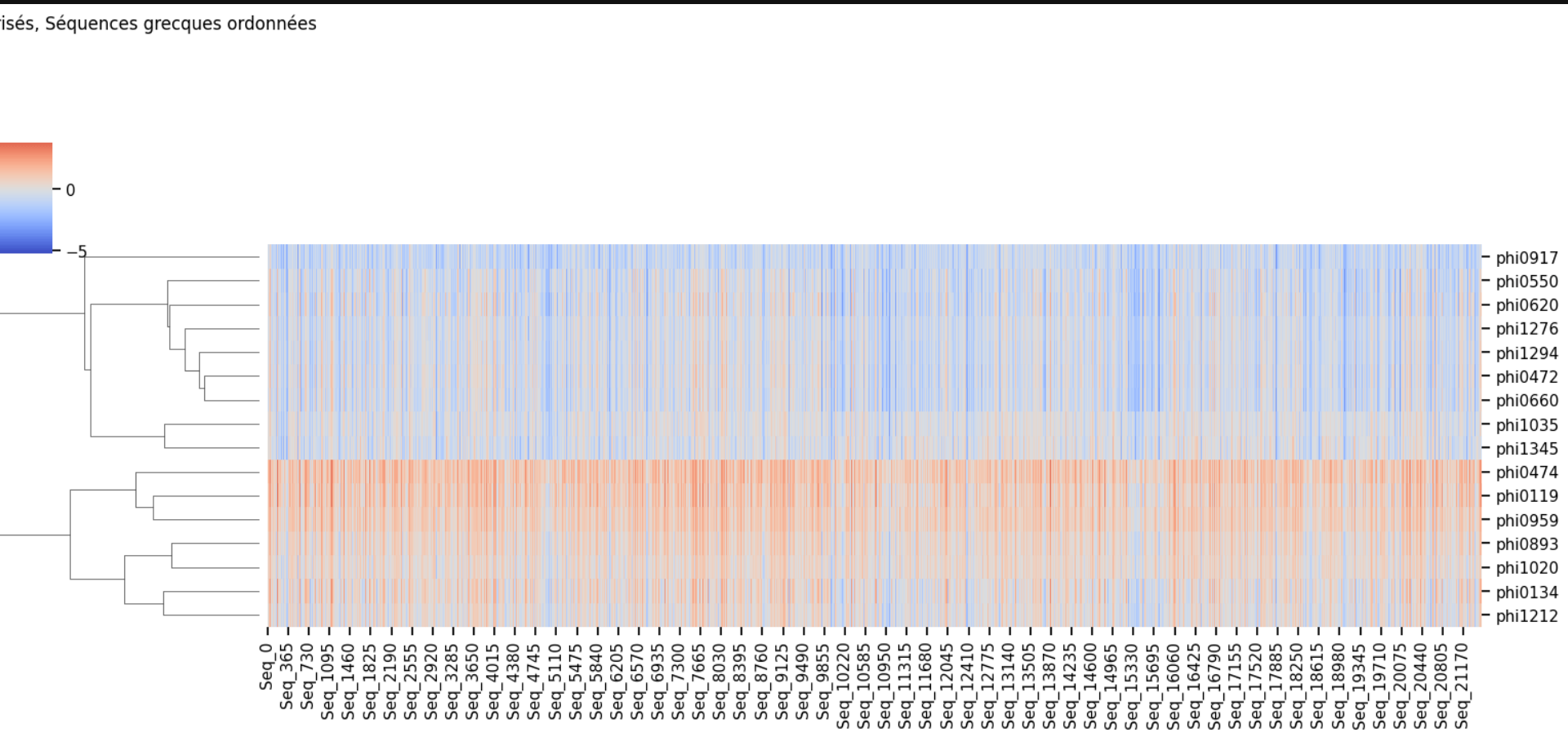
greek sentence





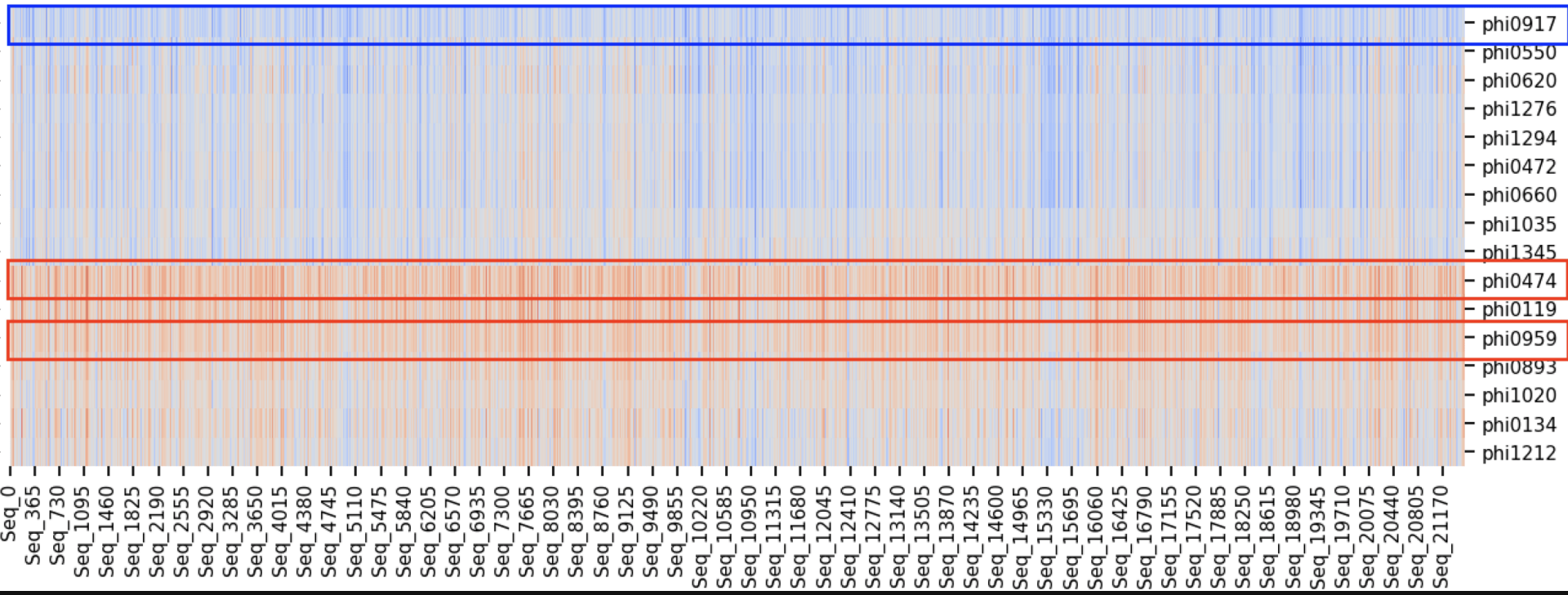
Step 5 : final results -- part 1

Iliad versus Odyssey



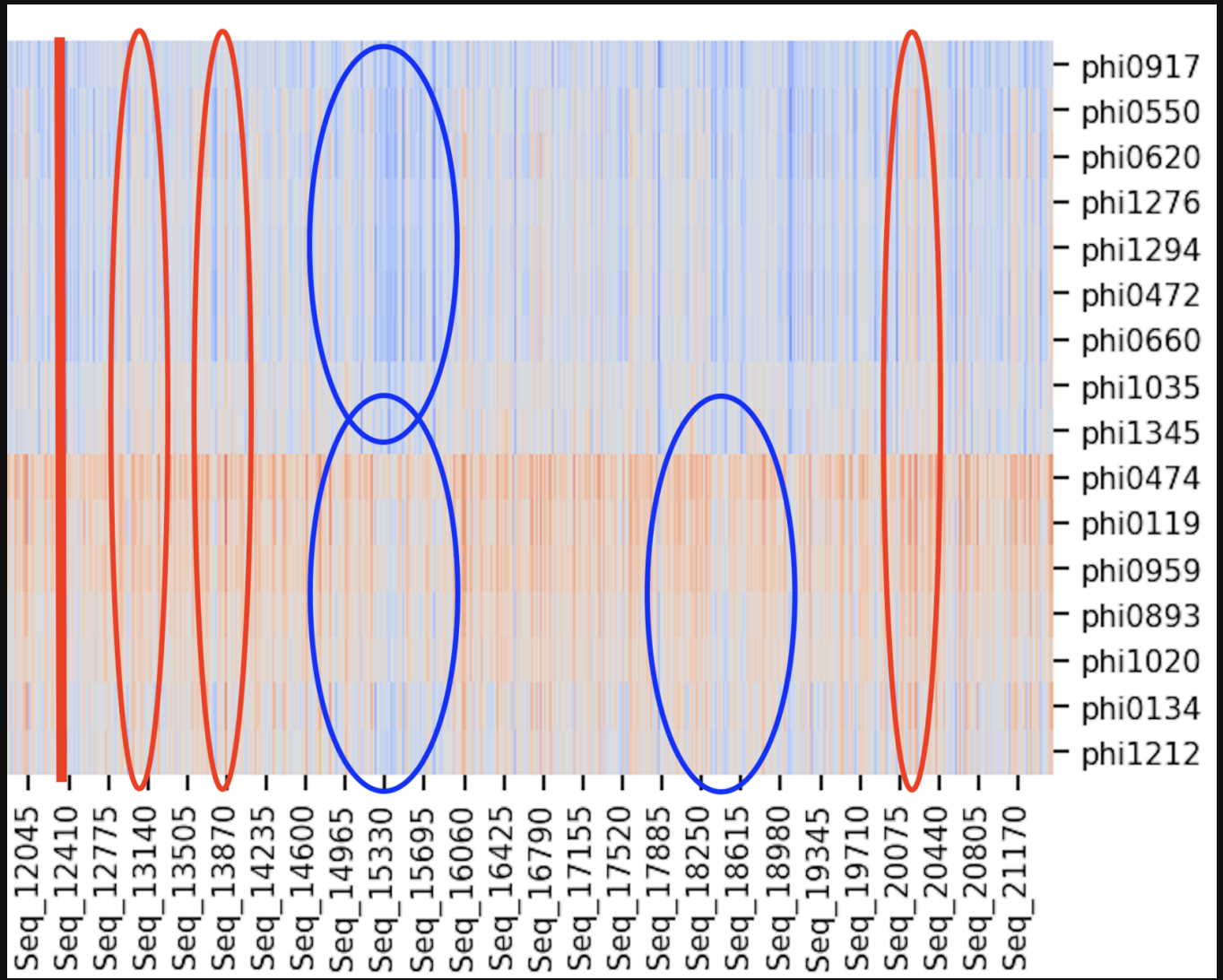
Step 5 : final results -- part 1

Iliad versus Odyssey

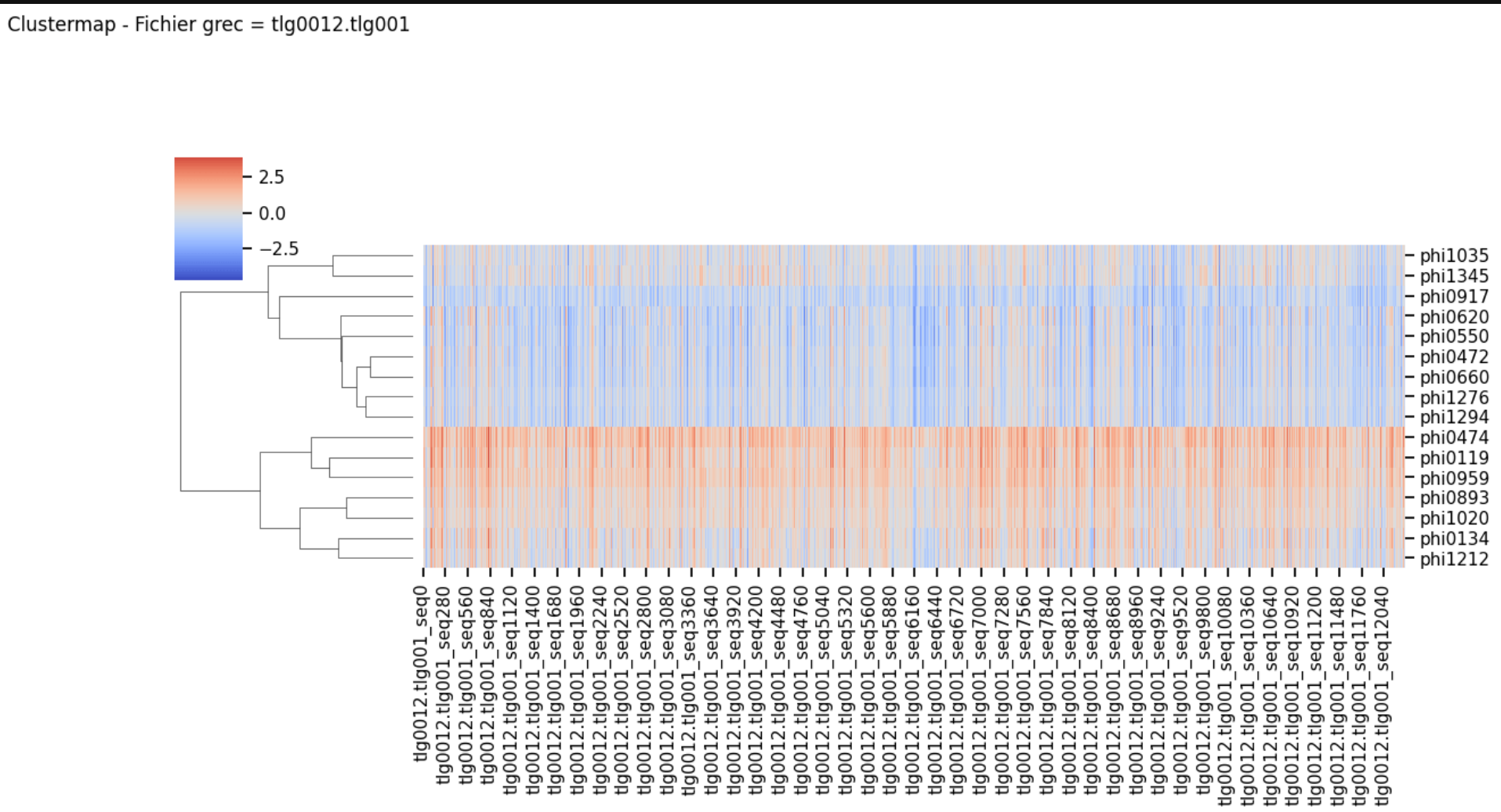


Step 5 : final results -- part 1

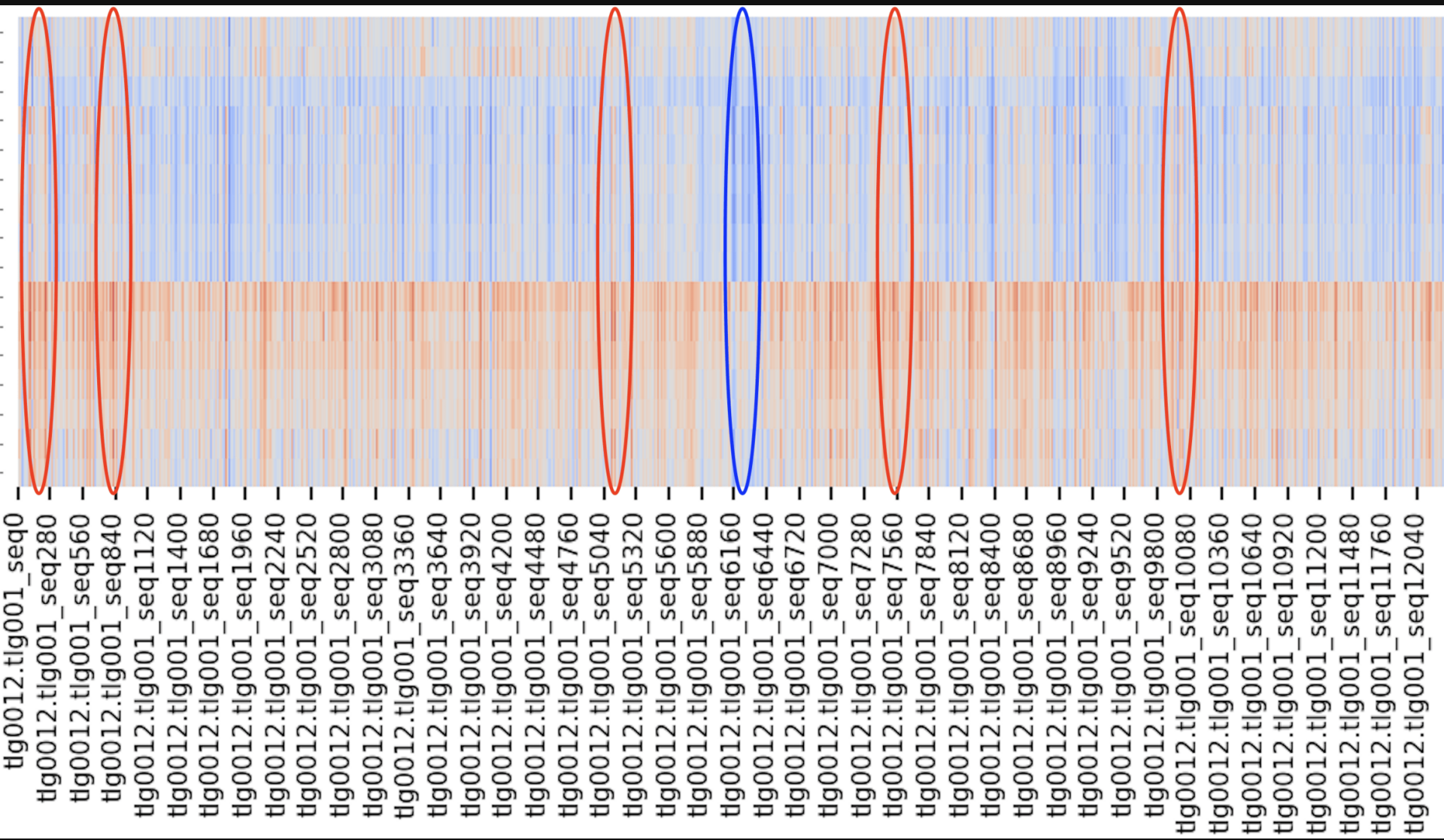
Iliad versus Odyssey



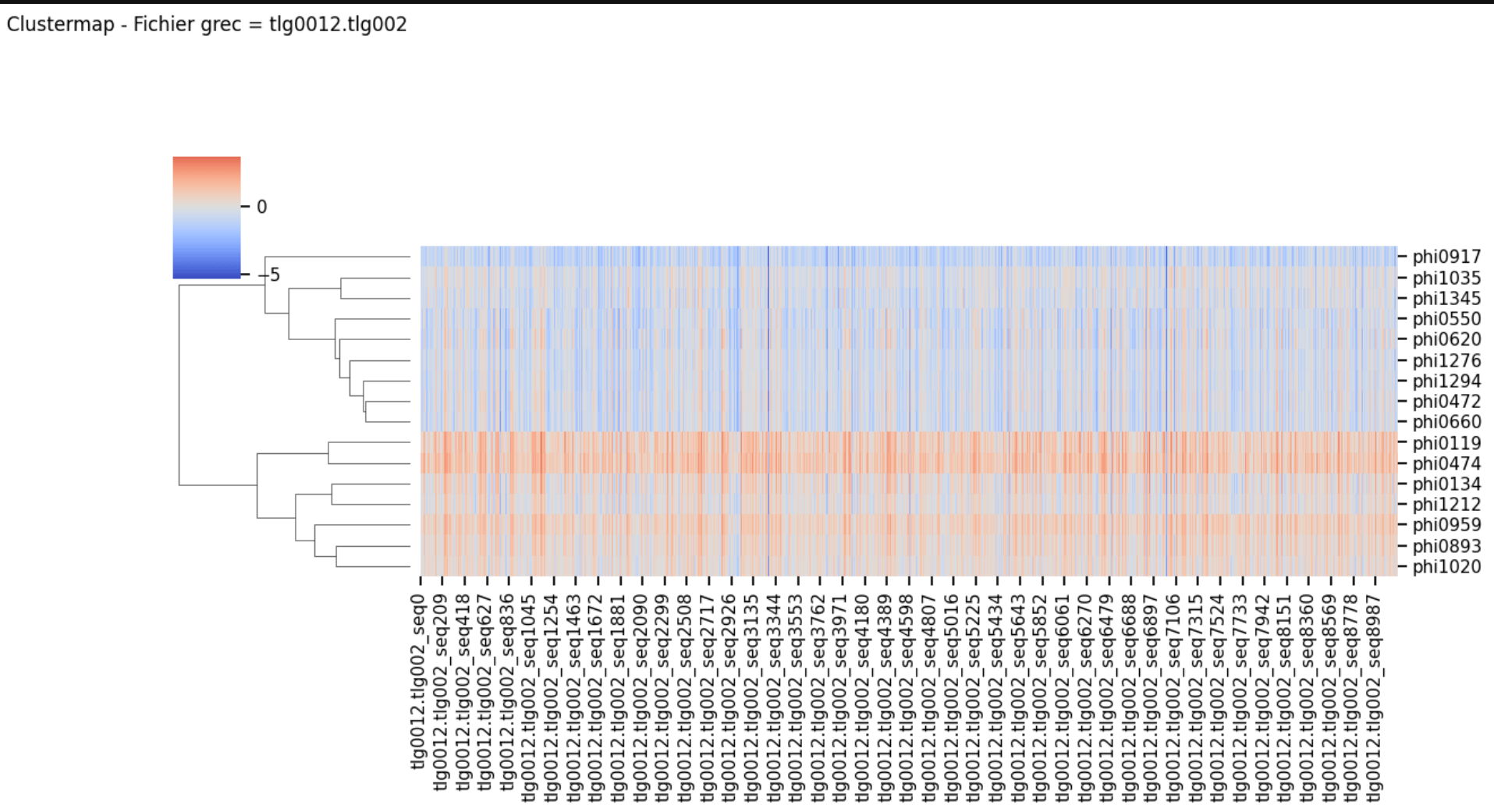
Step 5 : final results -- part 2



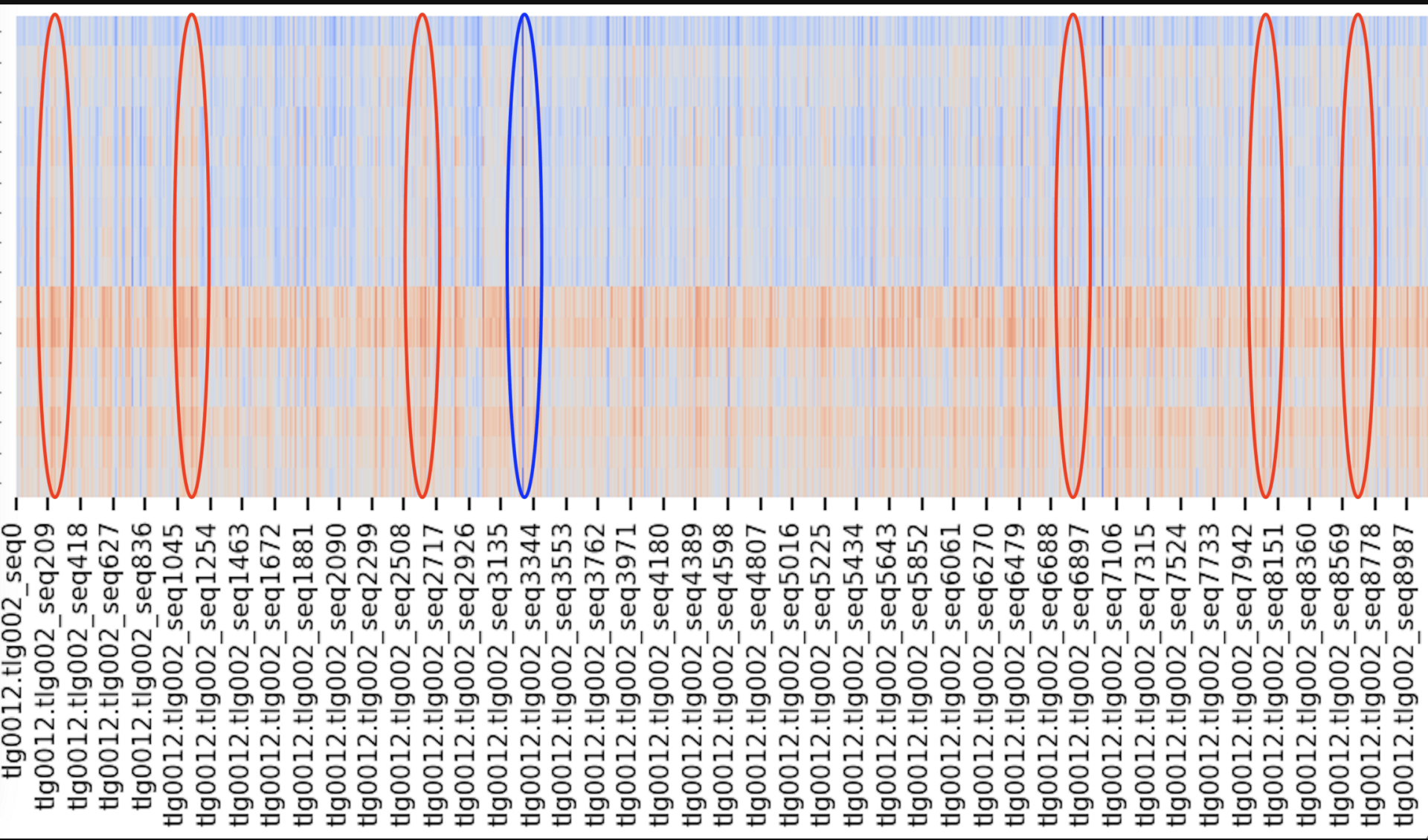
Step 5 : final results -- part 2



Step 5 : final results -- part 3



Step 5 : final results -- part 3



Results ? No and Yes

CONS :

- massive use of GPU : unable to run it on the whole similarities --> authors missing

- not appearing in the score matrix does not mean that there is no influence !

- similarity does not necessarily mean intertextuality (need to add stylometrics)

PROS :

- it is the best sentence aligner I've tested this far

- one may overcome the problem of detecting intertext using simple measures (stylos)

- it can be a good way to produce parallel data more easily

- the results can give us clues about what texts were conveyed and why

- the method can be applied to a much larger range of texts

Useful references

Yousef, T., Palladino, C., Shamsian, F., d'Orange Ferreira, A., & Ferreira dos Reis, M. (2022). An automatic model and Gold Standard for translation alignment of Ancient Greek. In Proceedings of the Language Resources and Evaluation Conference (pp. 5894–5905). European Language Resources Association. Marseille, France. Retrieved from https://aclanthology.org/2022.lrec-1.634

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Riemenschneider, F., & Frank, A. (2023). Graecia capta ferum victorem cepit. Detecting Latin Allusions to Ancient Greek Literature. In *Proceedings of the First Workshop on Ancient Language Processing*. Association for Computational Linguistics. Varna, Bulgaria. Retrieved from https://arxiv.org/abs/2308.12008

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

MZES Data Lab. (n.d.). BERT and explainable AI. Retrieved from https://www.mzes.uni-mannheim.de/socialsciencedatalab/article/bert-explainable-ai/

Contact

Thank you for your attention

https://github.com/OdysseusPolymetis/AI_DH_Concordia

marianne.reboul@ens-lyon.fr

github : OdysseusPolymetis