Grenoble -- 18 janvier 2024

Annotation lexico-grammaticale du grec/latin

https://github.com/OdysseusPolymetis/philo_num_grenoble_24

ENS DE LYON

Qu'est-ce que "pré-traiter" ?

La lemmatisation

C'est un retour du mot à son entrée lexicale commune (« forme canonique » enregistrée dans les dictionnaires de la langue), c'est-à-dire sans flexions.

Exemple :

Quo	**usque**	**tandem**	**abutere**	**Catilina**
Qui	usque	tandem	abutor	Catilina

L'étiquetage de nature

On appelle ça "postagging". Cela consiste à rechercher la nature des mots.

Exemple :

Quo	**usque**	**tandem**	**abutere**	**Catilina**
Qui	usque	tandem	abutor	Catilina
PRON	ADP	ADV	VERB	NOUN

Autres tâches (variées)

Entre autres : la NER

Exemple :

Quo	**usque**	**tandem**	**abutere**	**Catilina**
Qui	usque	tandem	abutor	Catilina
PRON	ADP	ADV	VERB	NOUN
				PERS

Pourquoi traiter les textes

https://github.com/OdysseusPolymetis/philo_num_grenoble_24/blob/main/why_nlp.ipynb

Différents outils à comparer

Outil 1 : pie-extended

https://github.com/OdysseusPolymetis/philo_num_grenoble_24/blob/main/pie_for_comparison.ipynb

Outil 2 : cltk/stanza

https://github.com/OdysseusPolymetis/philo_num_grenoble_24/blob/main/cltk_for_comparison.ipynb

https://github.com/OdysseusPolymetis/philo_num_grenoble_24/blob/main/stanza_for_comparison.ipynb

Outil 3 : spacy

https://github.com/OdysseusPolymetis/philo_num_grenoble_24/blob/main/spacy_for_comparison.ipynb

Outil 4 : BERT

https://github.com/OdysseusPolymetis/philo_num_grenoble_24/blob/main/ancient_greek_bert_for_comparison.ipynb

En conclusion

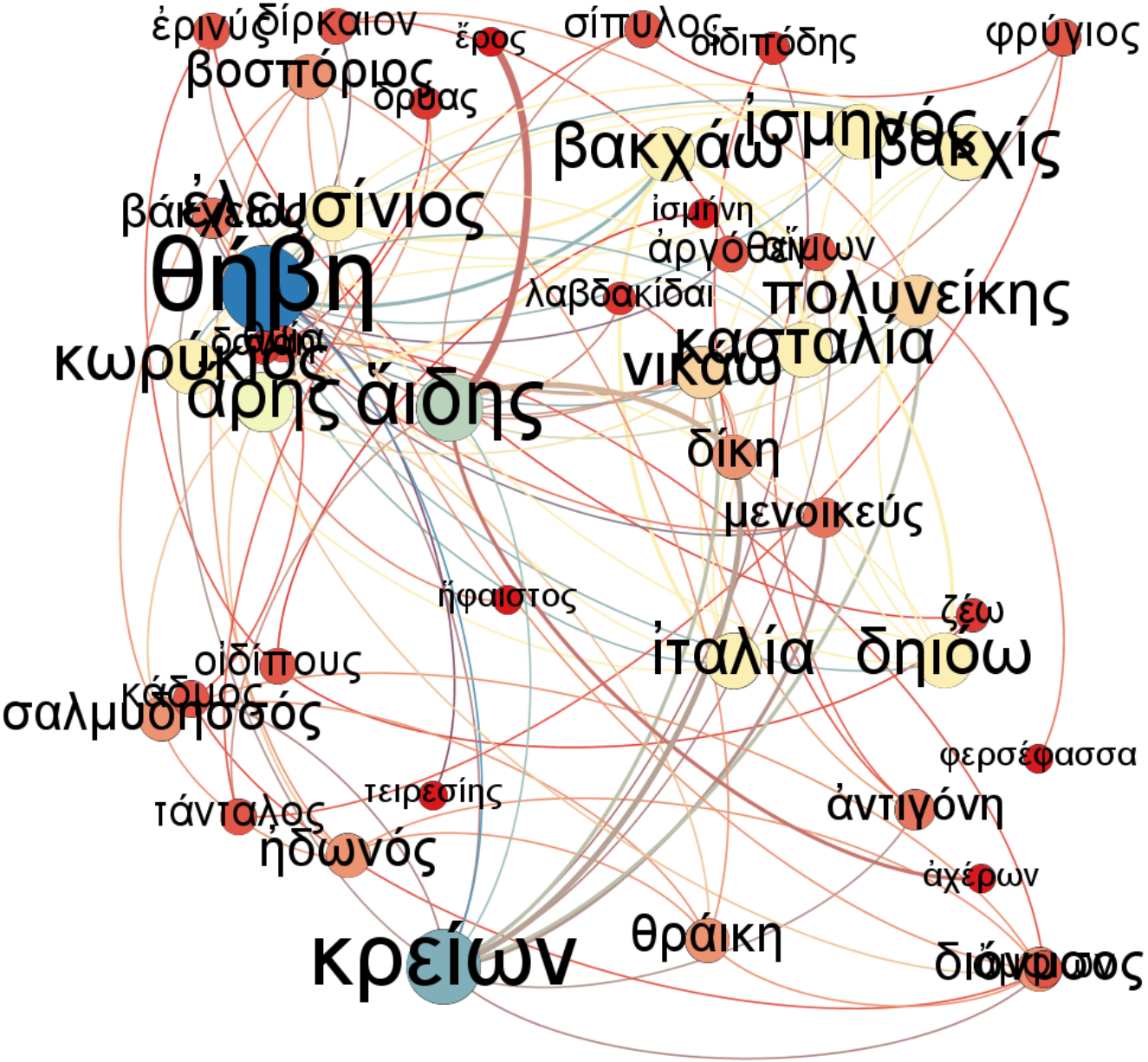
Je vous remercie.

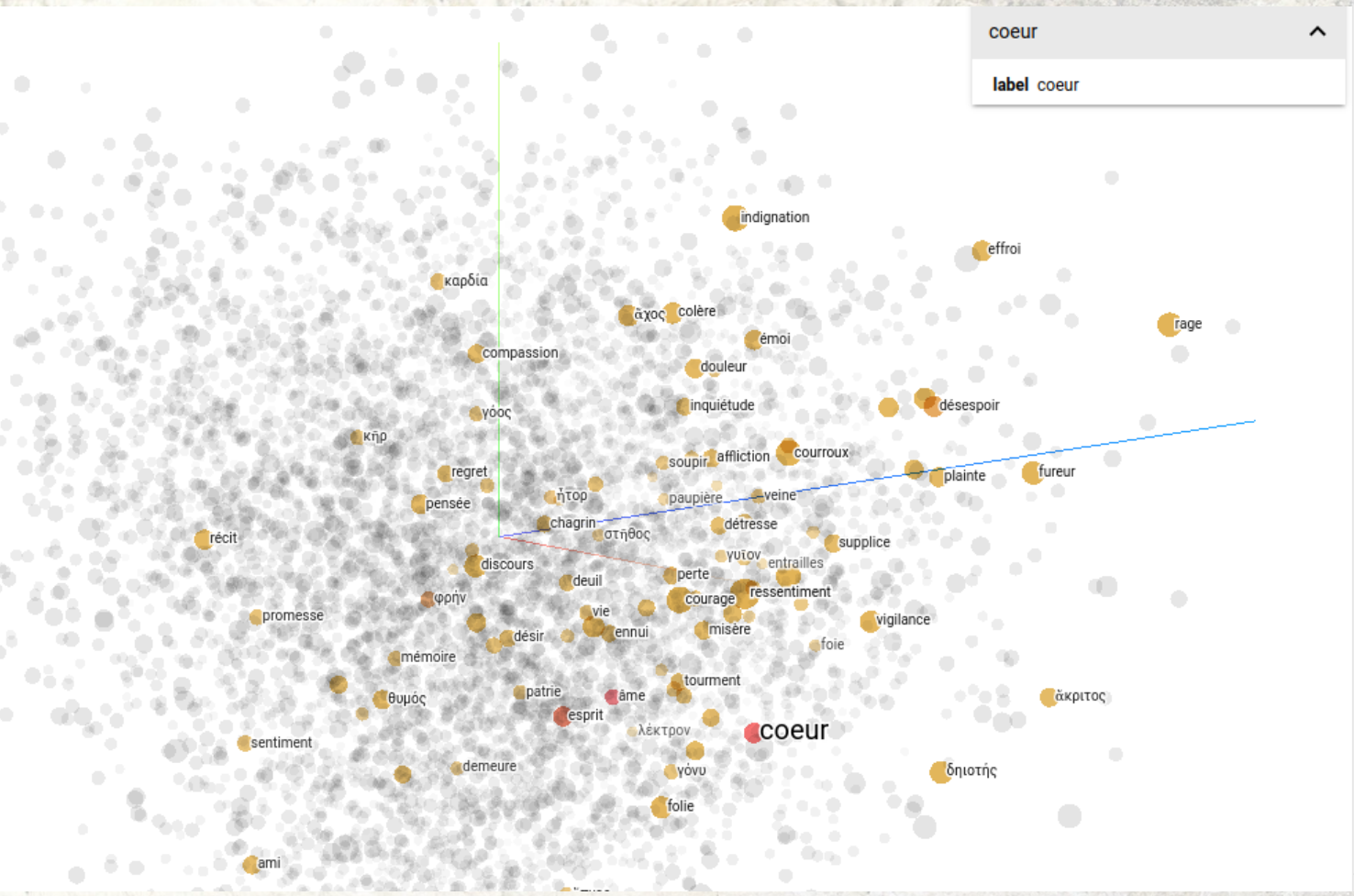
marianne.reboul@ens-lyon.fr

Références (et remerciements)

- deucalion : Clérice, T. (2018, November). Deucalion et Pyrrha: Environnement pour la lemmatisation et la postcorrection à l'École des chartes. In *Text Encoding: Latinists looking for new synergies*.
- pie-extended : Clérice, T. (2020). Pie Extended, an extension for Pie with pre-processing and post-processing. Zenodo. doi, 10.
- stanza : Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.
- ancient-greek-bert : Singh, P., Rutten, G., & Lefever, E. (2021). A pilot study for BERT language modelling and morphological analysis for ancient and medieval Greek. In *5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, co-located with EMNLP 2021* (pp. 128-137). Association for Computational Linguistics.
- spacy : Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.
- udpipe : Straka, M., Hajic, J., & Straková, J. (2016, May). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4290-4297).
- voyanttools : Sampsel, L. J. (2018). Voyant tools. *Music Reference Services Quarterly*, *21*(3), 153-157.
- cltk : Johnson, K. P., Burns, P. J., Stewart, J., Cook, T., Besnier, C., & Mattingly, W. J. (2021, August). The Classical Language Toolkit: An NLP framework for pre-modern languages. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing: System demonstrations* (pp. 20-29).

Addenda :
Quelques exemples de visualisations récentes


ἐρινύς
δίρκαιον
ἔρος
σίπυλος
διδιτόδης
φρύγιος
βοσπτόριος
ὄρθας
ἰσμηνός
βακχάω
βάκχις
ἐλευσίνιος
βάκχειος
ἰσμήνη
ἀργόθεν
θήβη
λαβδακίδαι
πολυνείκης
κωρύκιος
καστταλία
νικάω
ἄρης
ἄιδης
δίκη
μενοικεύς
ἥφαιστος
ζέω
ἰταλία
δηιόω
οἰδίπους
κάδμος
σαλμυδησσός
φερσέφασσα
τειρεσίης
τάνταλος
ἀντιγόνη
ἠδωνός
ἀχέρων
θράικη
κρείων
διόνυσος


coeur
label coeur
co
.*
by
neighbors
100
distance
COSINE
EUCLIDEAN
Nearest points in the original space:
âme 0.283
esprit 0.390
φρήν 0.506
désespoir 0.547
ἦτορ 0.585
καρδία 0.597
ἄχος 0.598
ressentiment 0.598
pensée 0.599
γόνυ 0.600
courroux 0.601
détresse 0.605
douleur 0.606
fureur 0.611
κῆρ 0.612
ennui 0.613
indignation
effroi
καρδία
ἄχος
colère
émoi
rage
compassion
douleur
inquiétude
désespoir
γόος
κῆρ
soupir
affliction
courroux
regret
plainte
fureur
pensée
ἦτορ
paupière
veine
chagrin
στῆθος
détresse
récit
supplice
γυῖον
entrailles
discours
deuil
perte
courage
ressentiment
φρήν
promesse
vie
ennui
misère
vigilance
désir
foie
mémoire
tourment
θυμός
patrie
âme
ἄκριτος
esprit
λέκτρον
coeur
sentiment
demeure
γόνυ
δηιοτής
folie
ami

Similarité globale entre Homère et les tragiques