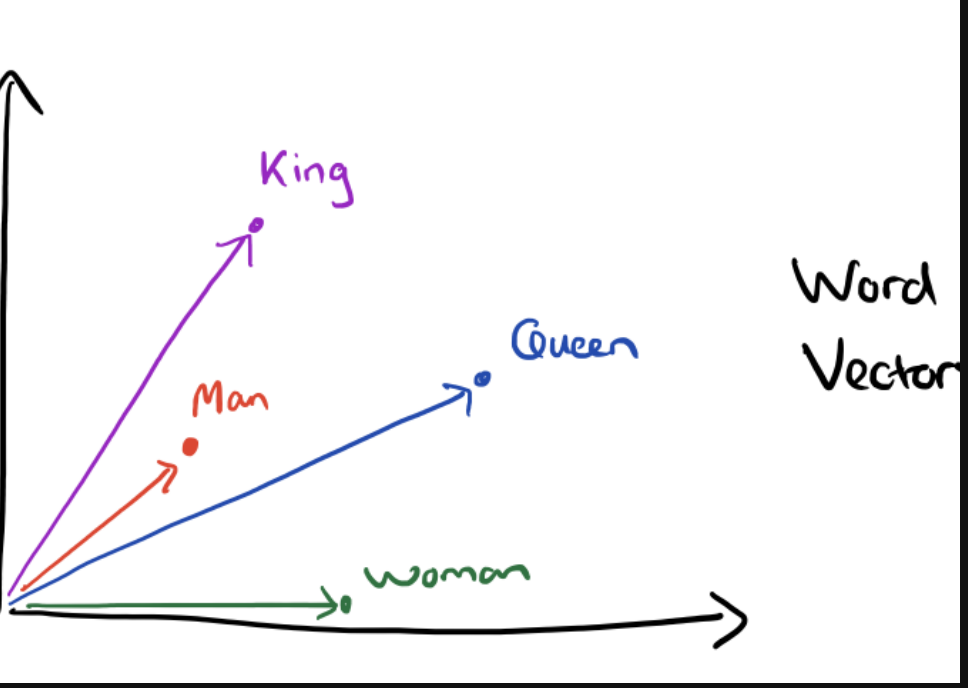
Marianne Reboul

2 avril 2024

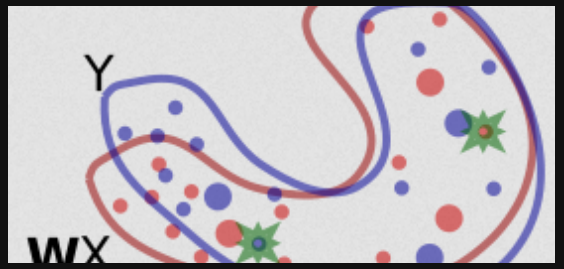
# {LLM, grec et latin}

*Constructions et usages de modèles monolingues et multilingues pour le latin et le grec ancien*

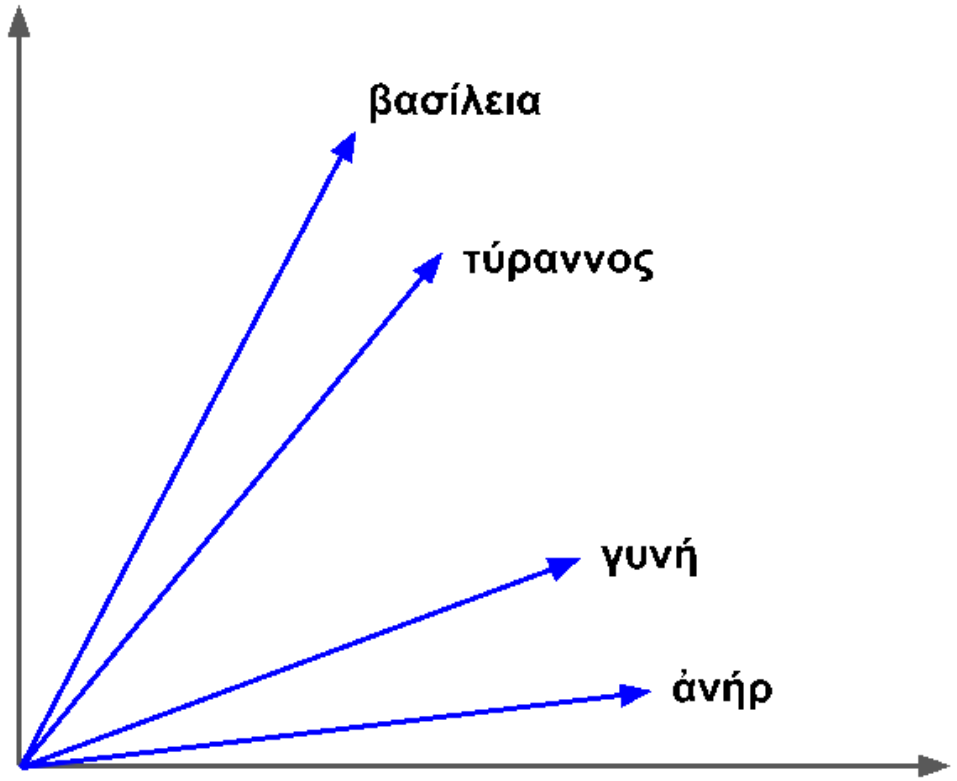



# Les bases

Vecteurs de mots statiques, contextuels...



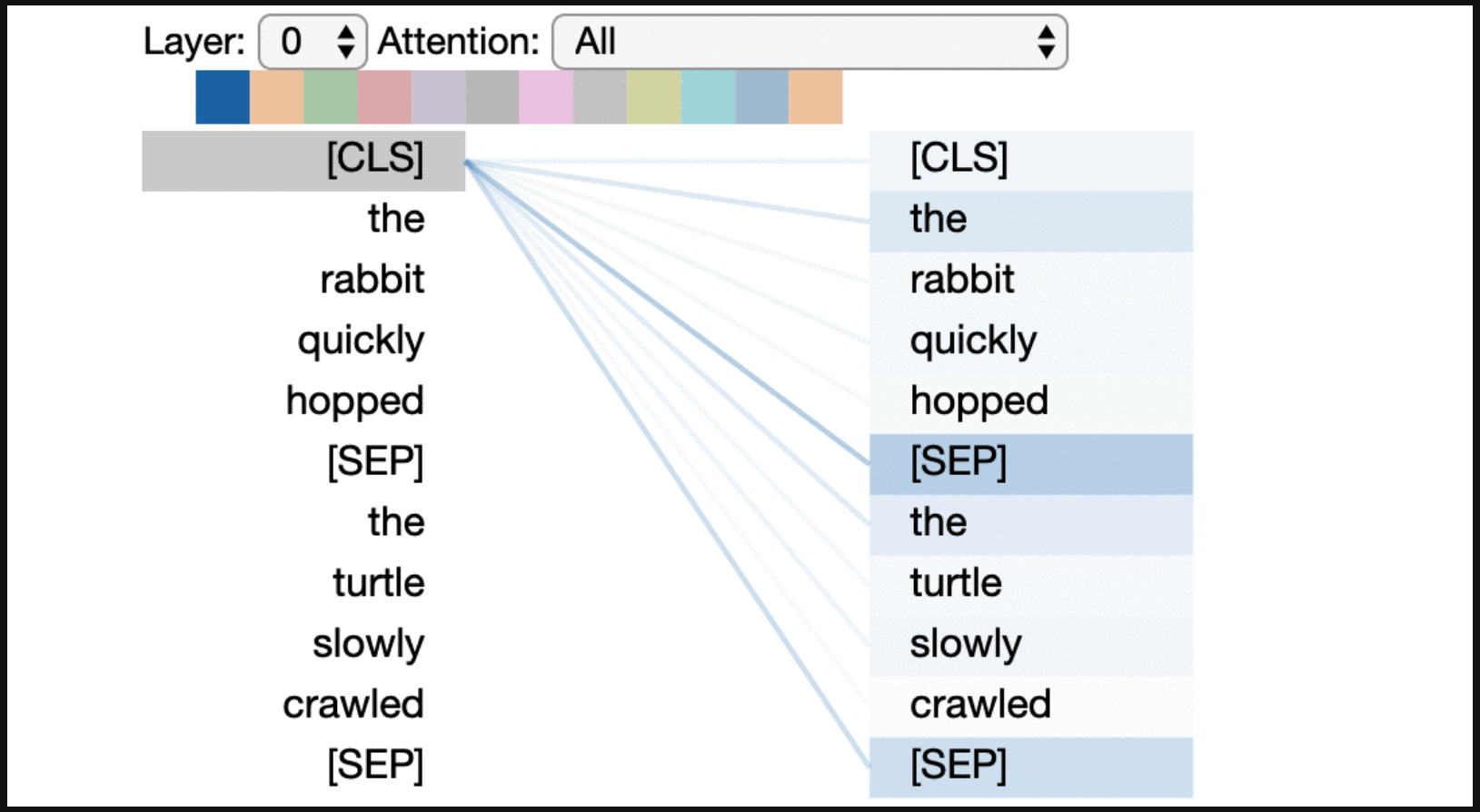

# Les vecteurs de mots



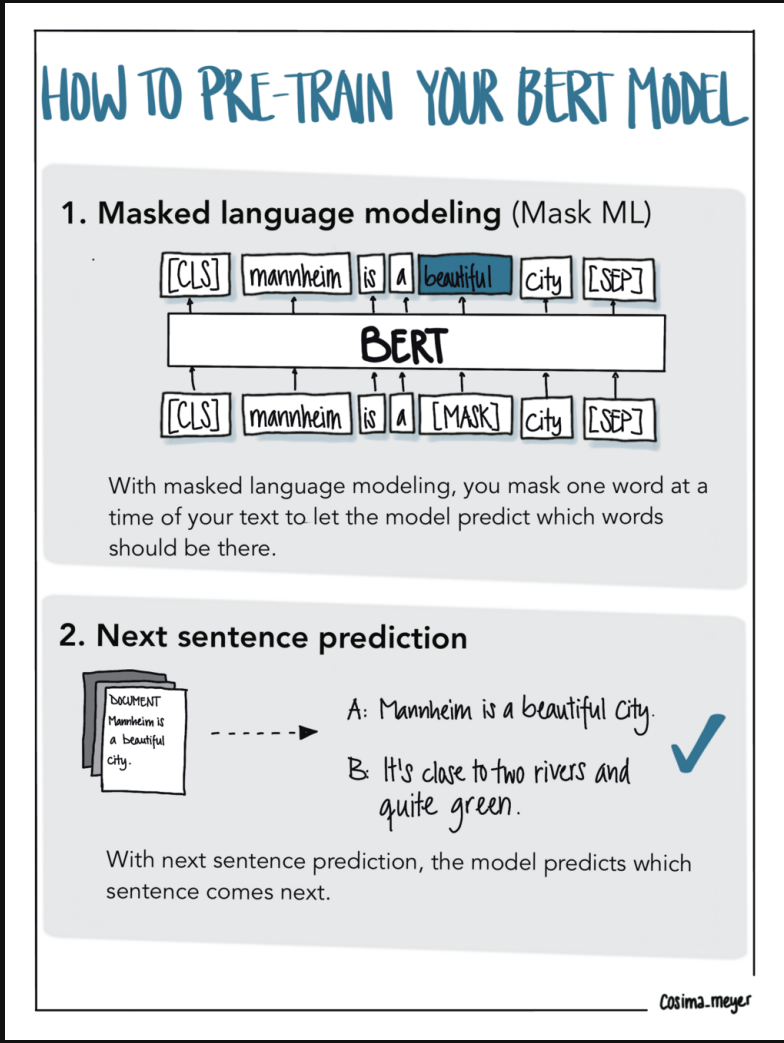
# Exemple sur un corpus conséquent

https://projector.tensorflow.org/

# Qu'est-ce qu'un grand modèle de langue (LLM) ?



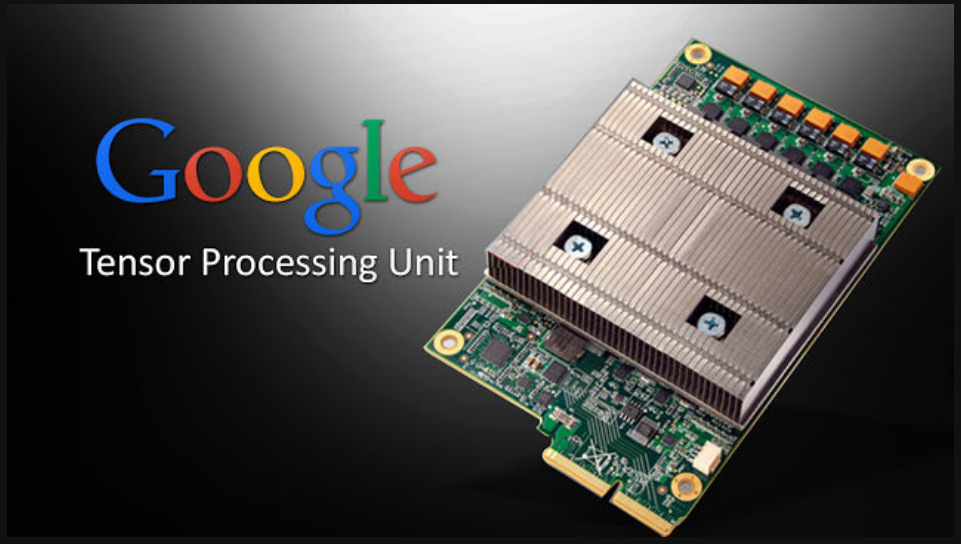
# Entraînement




https://www.mzes.uni-mannheim.de/socialsciencedatalab/article/bert-explainable-ai/#bert

# Problèmes ?

Petit problème matériel d'abord : les LLMs,
ça n'est pas pour tout le monde.



# Problèmes ?

Mais le problème principal :

- le très faible nombre de données

- le très faible nombre de données représentatives (hors âge classique)

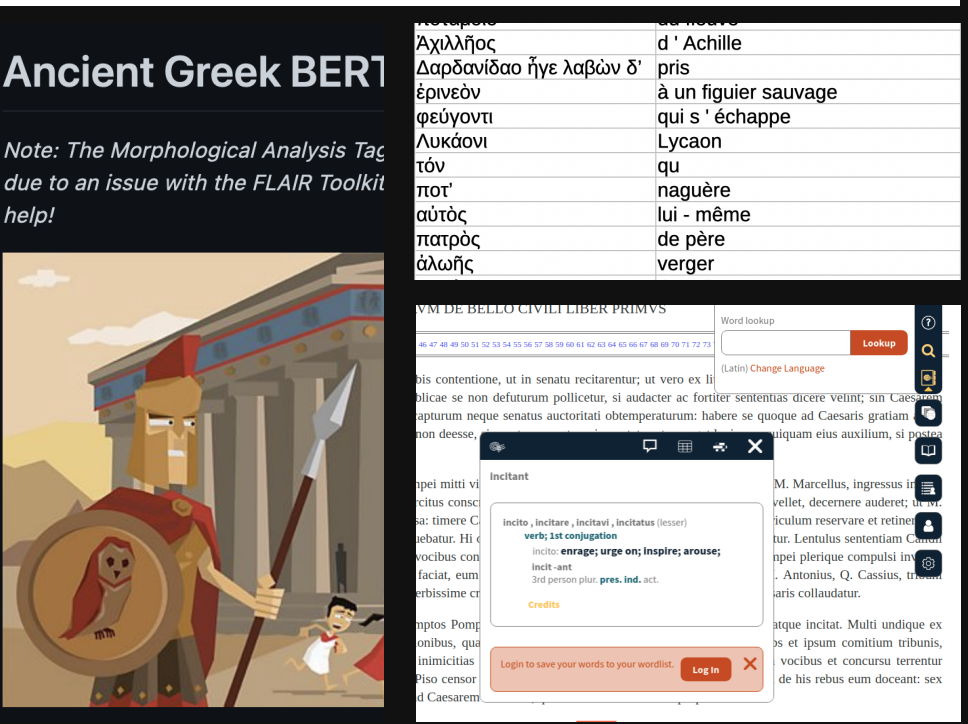
# MONOLINGUE

# Utilisation de LLMs monolingues

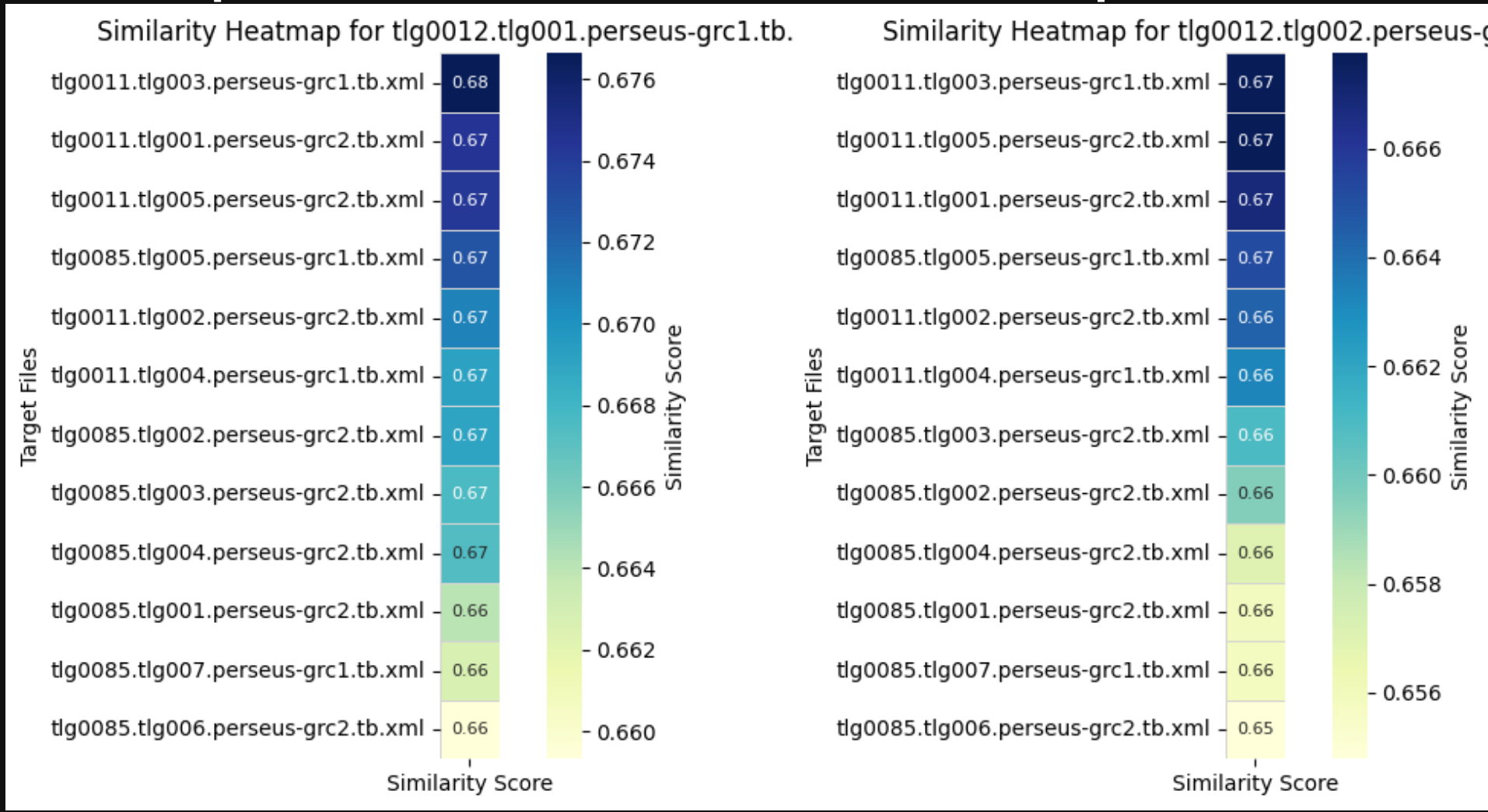
## À quoi cela peut servir ?

| | 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.

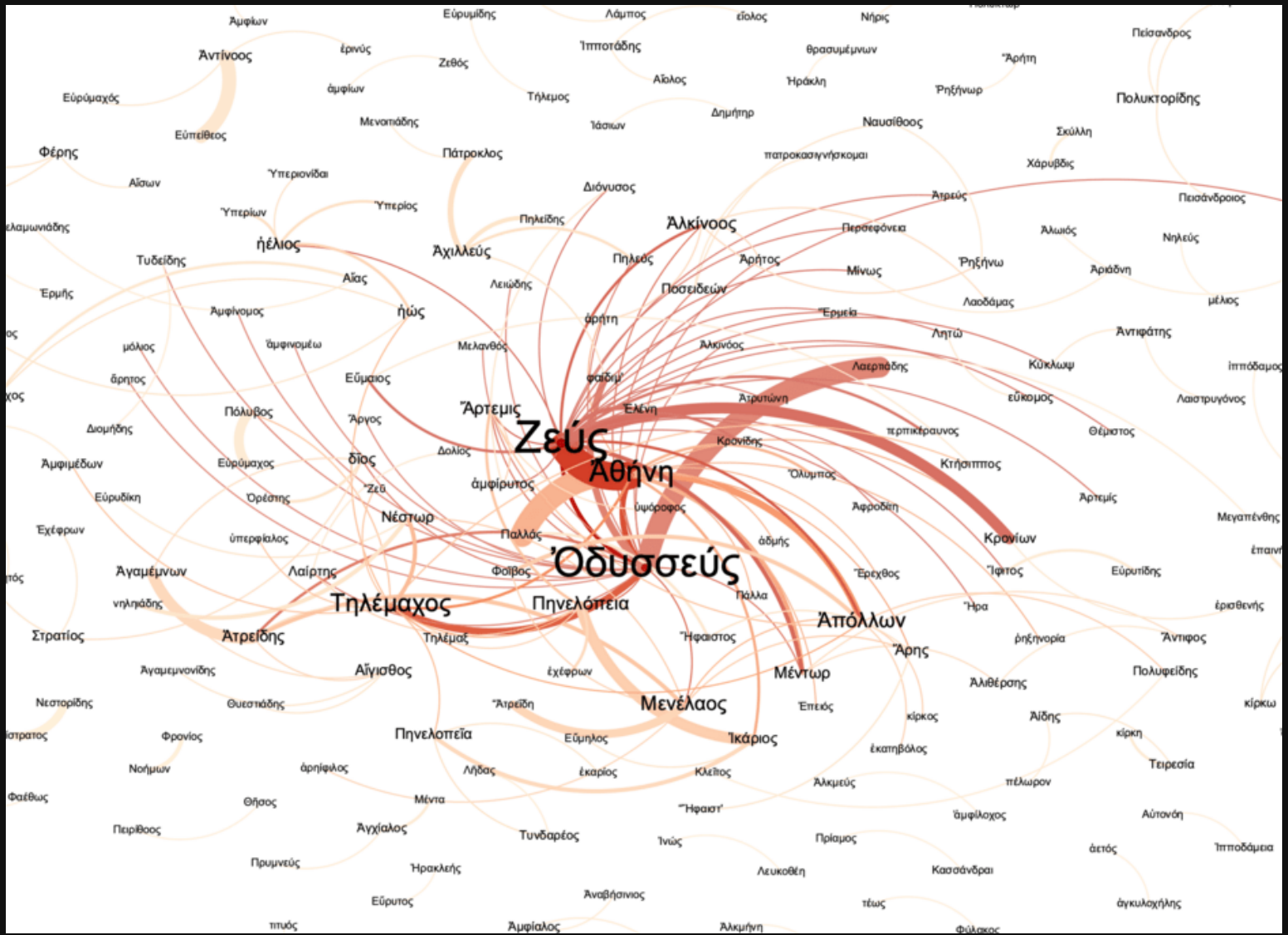


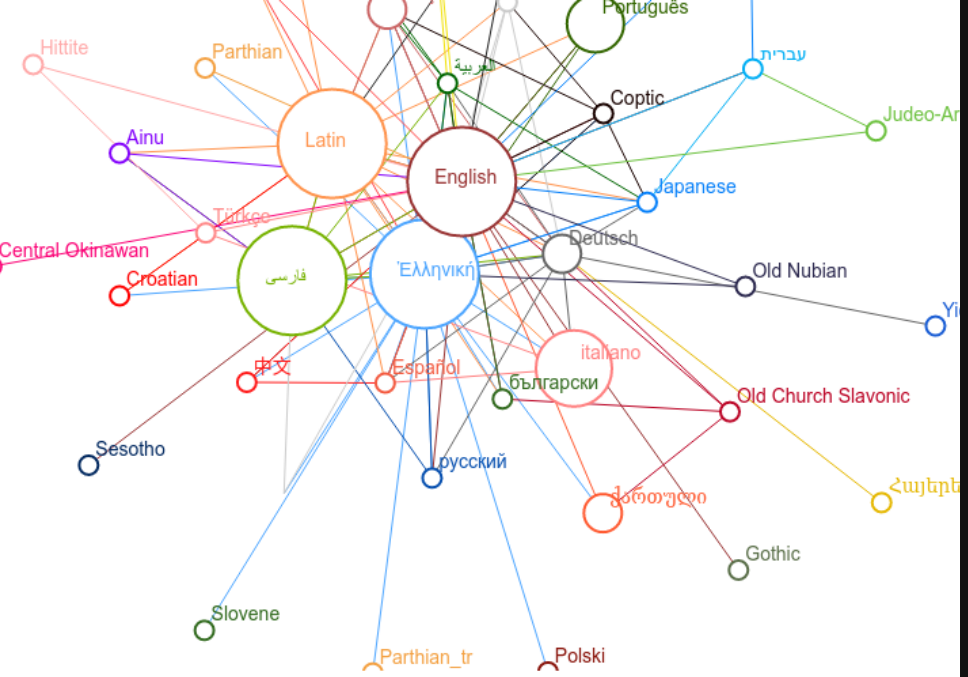
# Exemple 1 : similarité sémantique

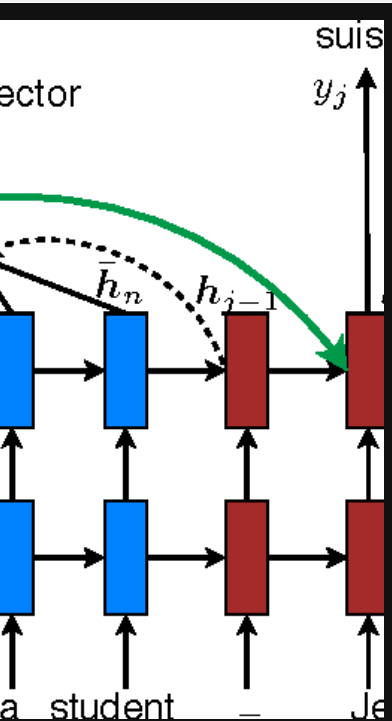



https://github.com/OdysseusPolymetis/humanistica2023/blob/main/humanistica2023.ipynb

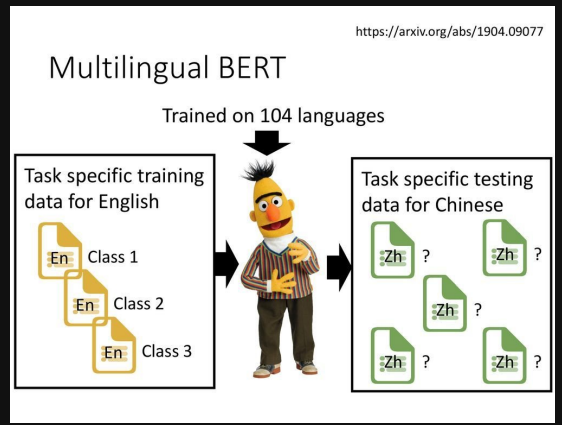
# Exemple 2 : NER






| | |
|---|---|
| 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 |




# Utilisation de LLMs multilingues

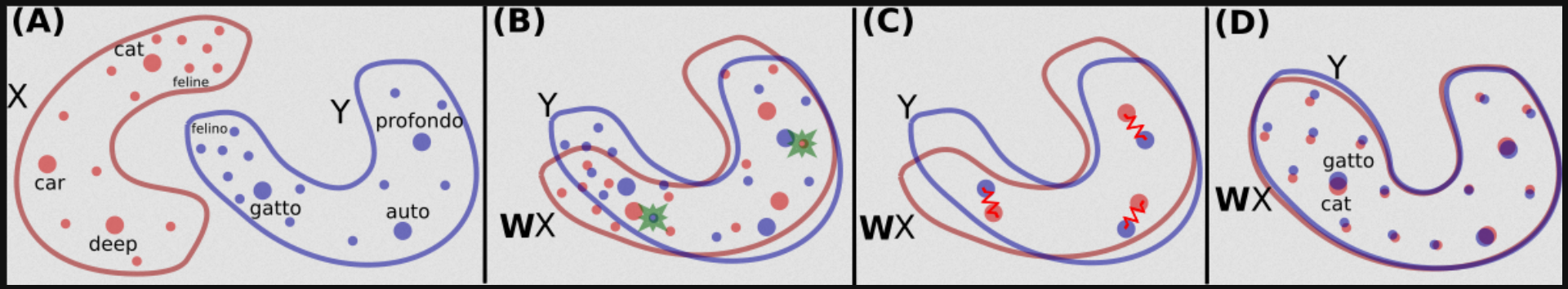
D'encore plus gros problèmes

# Pourquoi c'est encore plus compliqué

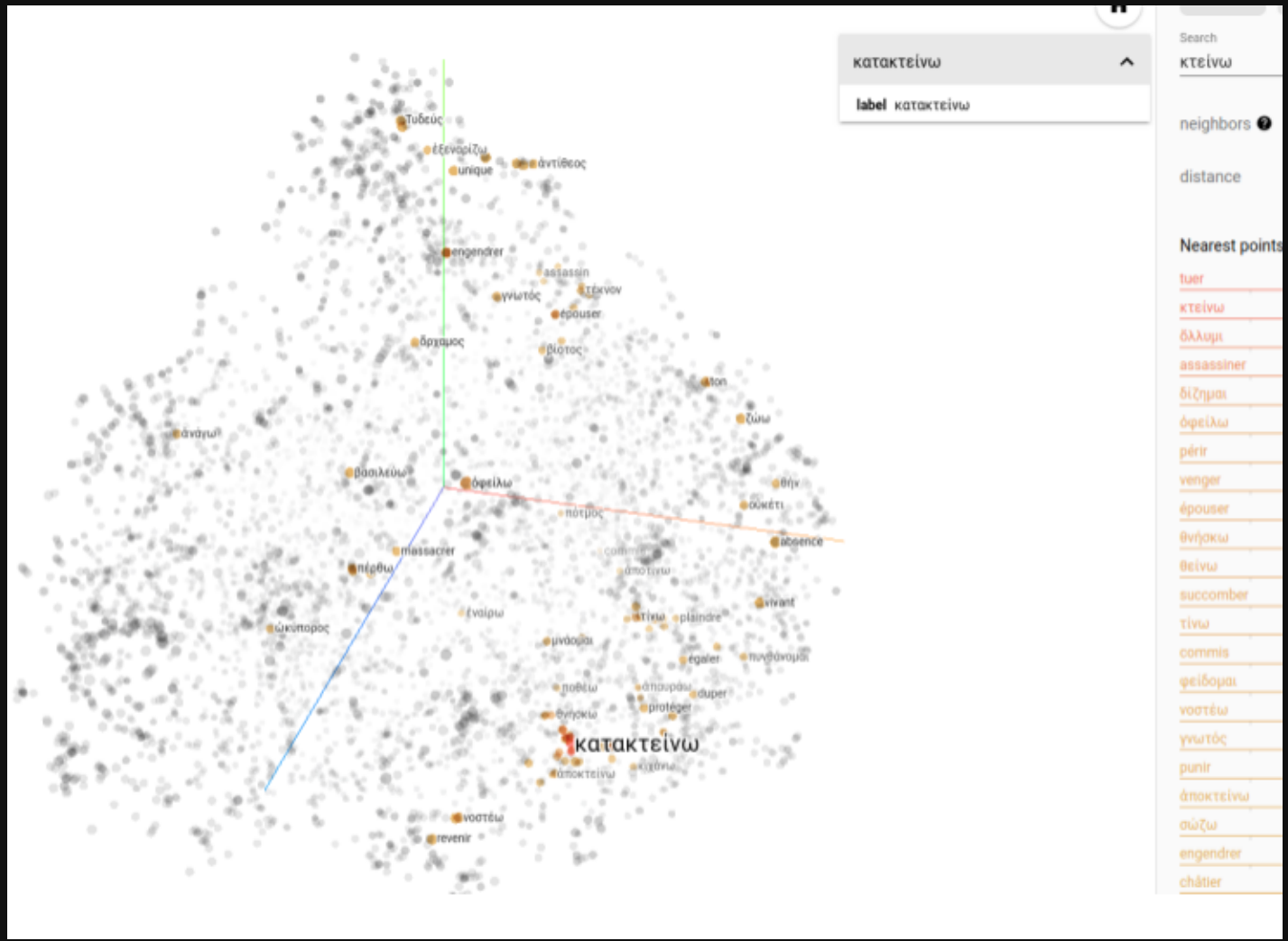
wmt19_translate/de-fr

- **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
```

# Alignement de séquences et détection d'intertexte

```
Greek File: tlg0086.tlg025.perseus-grc1.xml, Latin File: phi0550.phi001.perseus-lat1.xml
  Greek Phrase: δῆλον δ’ ἔσται τὸ λεγόμενον ἐκ τῶν ὕστερον μᾶλλον.
  Latin Phrase: id licet hinc quamvis hebeti cognoscere corde.
  Similarity Score: 0.9741785526275635

  Greek Phrase: δῆλον δ’ ἔσται τὸ λεγόμενον ἐκ τῶν ὕστερον μᾶλλον.
  Latin Phrase: id licet hinc quamvis hebeti cognoscere corde.
  Similarity Score: 0.9741785526275635

  Greek Phrase: περὶ μὲν οὖν τούτων δεδήλωται καὶ πρότερον·
  Latin Phrase: id quod iam supra tibi paulo ostendimus ante.
  Similarity Score: 0.9737507104873657

  Greek Phrase: δῆλον δ’ ἔσται τὸ λεγόμενον ἐκ τῶν ὕστερον μᾶλλον.
  Latin Phrase: quae tibi posterius largo sermone probabo.
  Similarity Score: 0.9700868129730225

  Greek Phrase: περὶ μὲν οὖν τούτων δεδήλωται καὶ πρότερον·
  Latin Phrase: id quod iam supera tibi saepe ostendimus ante.
  Similarity Score: 0.9692978858947754

  Greek Phrase: περὶ μὲν οὖν τούτων δεδήλωται καὶ πρότερον·
  Latin Phrase: id quod iam supera tibi paulo ostendimus ante.
  Similarity Score: 0.9681644439697266
```

# Contact

Je vous remercie.

https://github.com/OdysseusPolymetis/ganeshs_ia_02_04_24

marianne.reboul@ens-lyon.fr

github : OdysseusPolymetis

# Bibliographie

Yousef, T., Palladino, C., & Jänicke, S. (2022). Transformer-based named entity recognition for ancient greek.

Yousef, T., Palladino, C., Shamsian, F., Ferreira, A. D. O., & dos Reis, M. F. (2022, June). An automatic model and Gold Standard for translation alignment of Ancient Greek. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 5894-5905).

Riemenschneider, F., & Frank, A. (2023). Exploring large language models for classical philology. *arXiv preprint arXiv:2305.13698*.

Artetxe, M., Labaka, G., & Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297*.

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*.