Apprentissage Machine et manque de données

L'apport des réseaux de neurones

Que fais-je ?

- Des **langues anciennes** (latin et grec essentiellement pour l'instant)
- Des **modèles** de langue, par "word embeddings"

Qu'est-ce que j'essaie de faire ?

- Projet en cours : voir l'**influence** sémantique des langues anciennes entre elles dans le temps
- Ce qui veut dire : la machine doit identifier des **correspondances** sémantiques entre les langues
- C'est (un peu) le même processus pour la **traduction automatique**,

MAIS

Pourquoi ?

- Trop de données pour une analyse strictement **humaine**
- Pas assez de données pour une analyse **machine** "traditionnelle"

Trop de données pour une analyse humaine

Rien que pour la **Perseus Digital Library** :

15 millions de mots en grec

12 millions de mots en latin

des traductions disponibles pour chaque texte

Pas assez de données pour les usages "traditionnels"

De quoi s'agit-il ?

La NMT

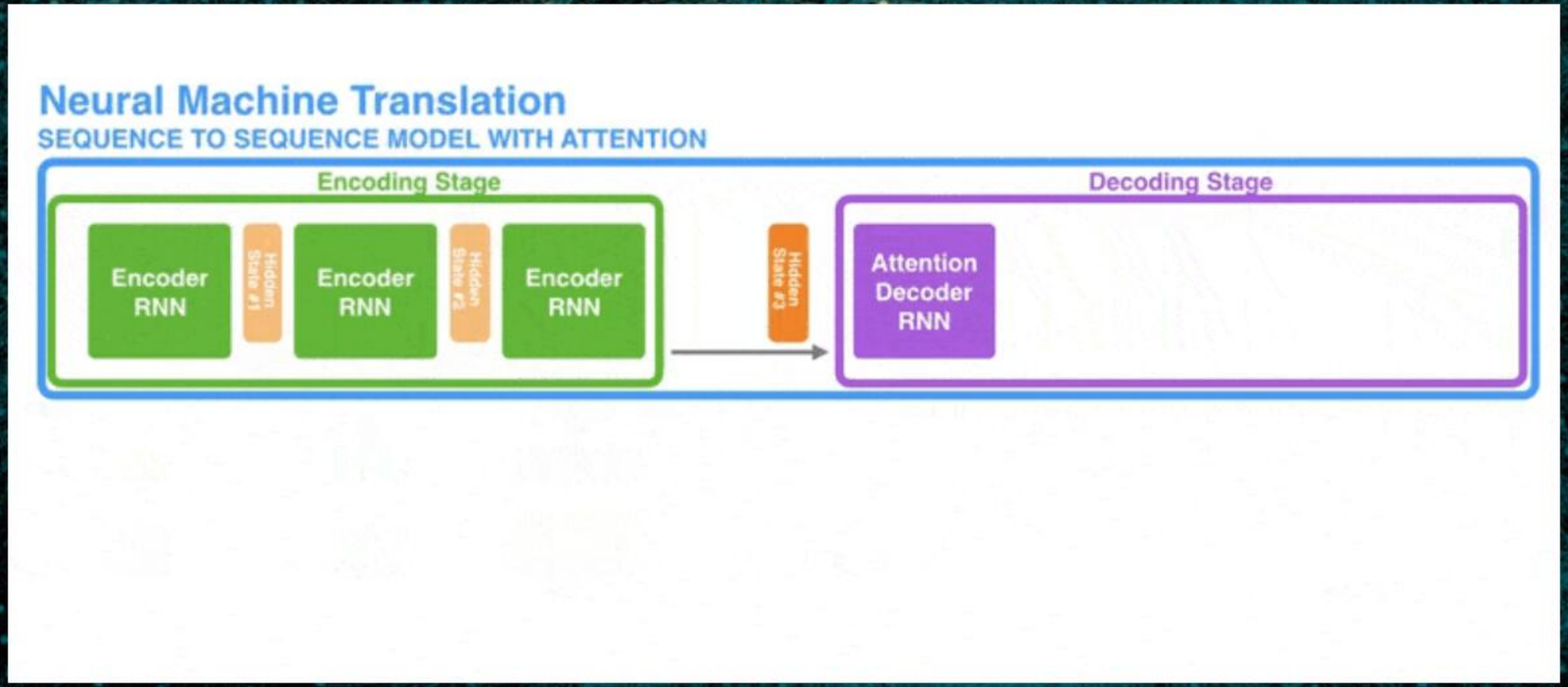
Rapidement, comment ça fonctionne

Deux phases :

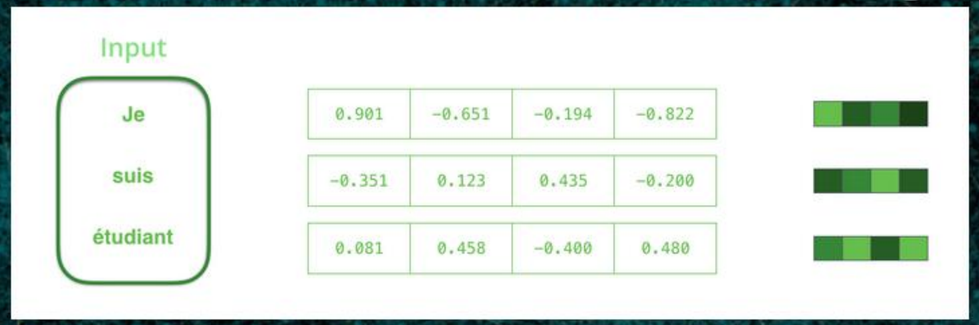
1, l'entraînement

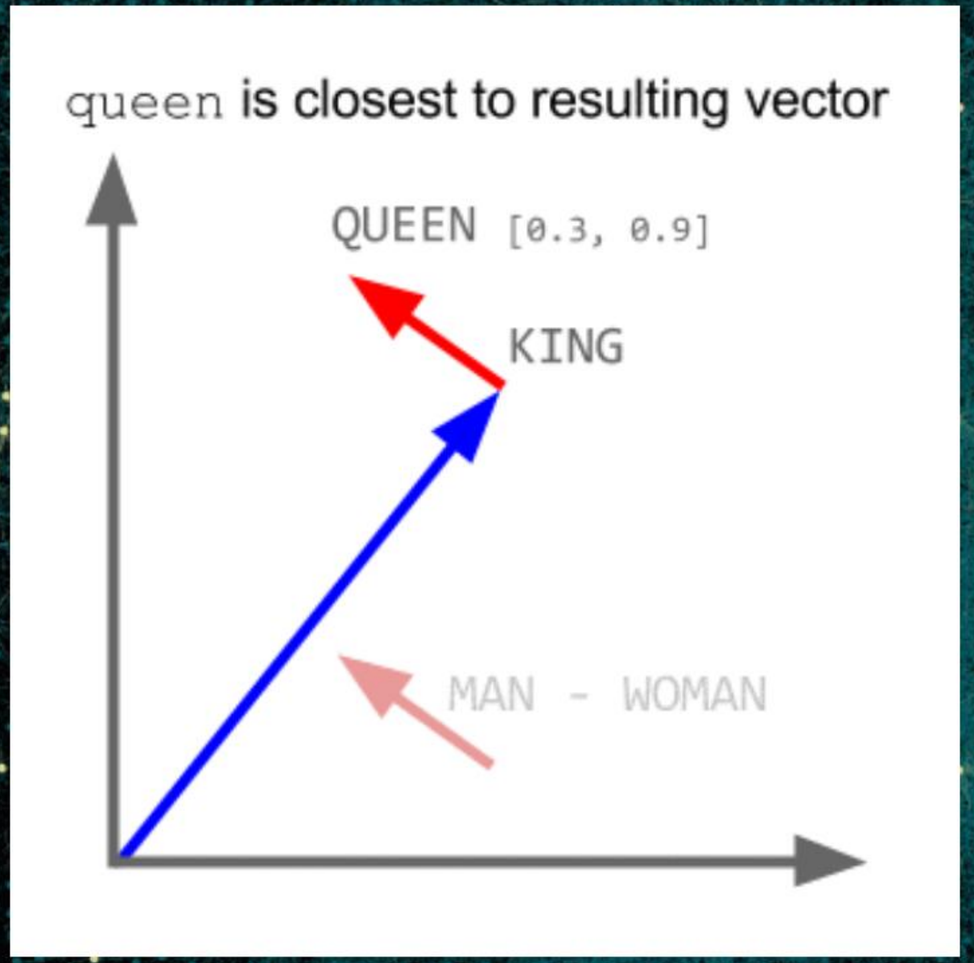
2, la prédiction

Création d'un modèle avec réseaux de neurones



Les mots en "embeddings"

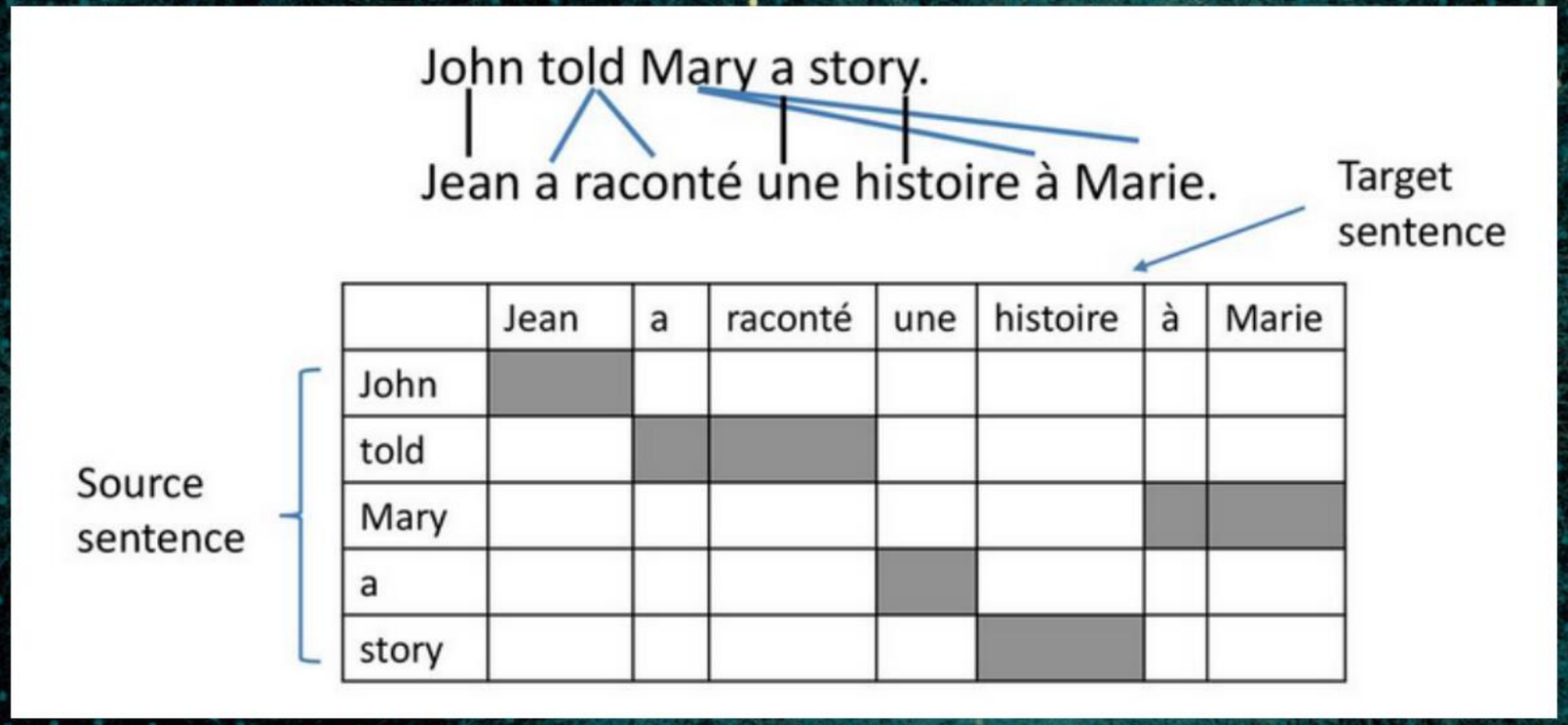




Problème de ce type d'approche pour nous

Nous avons des données, mais
pas de données parallèles.
On a beaucoup de traductions, mais on ne sait pas précisément quelles séquences sources correspondent aux séquences cibles.

Pourquoi ce n'est pas si simple ?

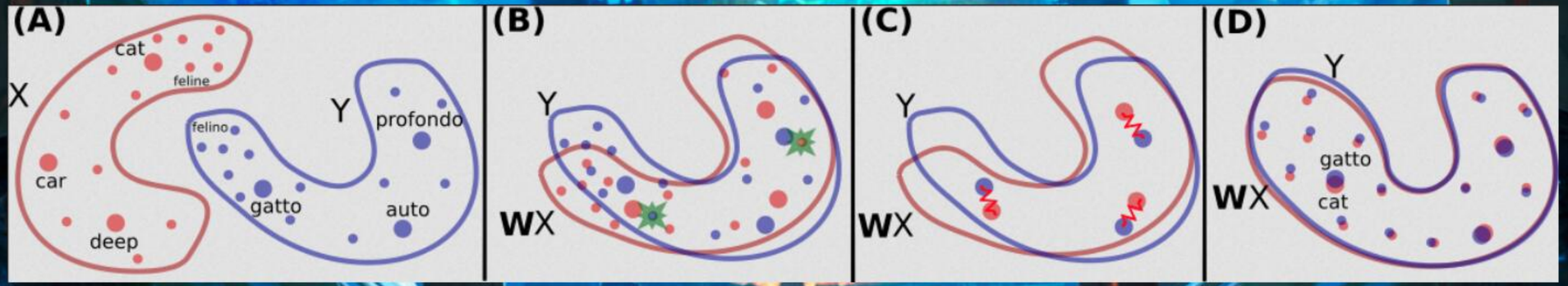


Une de mes tentatives précédentes

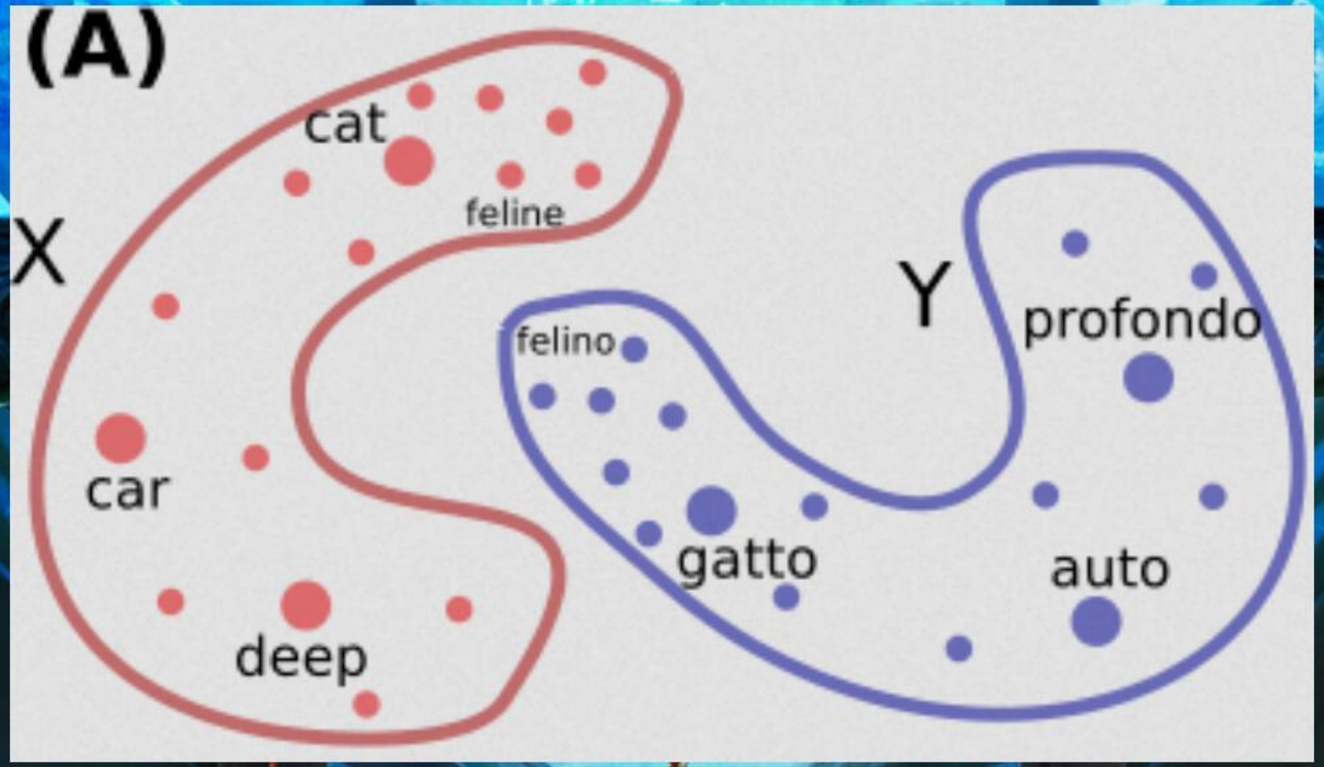

Du monolingue
au multilingue
Une solution pour les
langues rares ?


Un principe
fondateur de base
(discutable)
La ressemblance des
langues

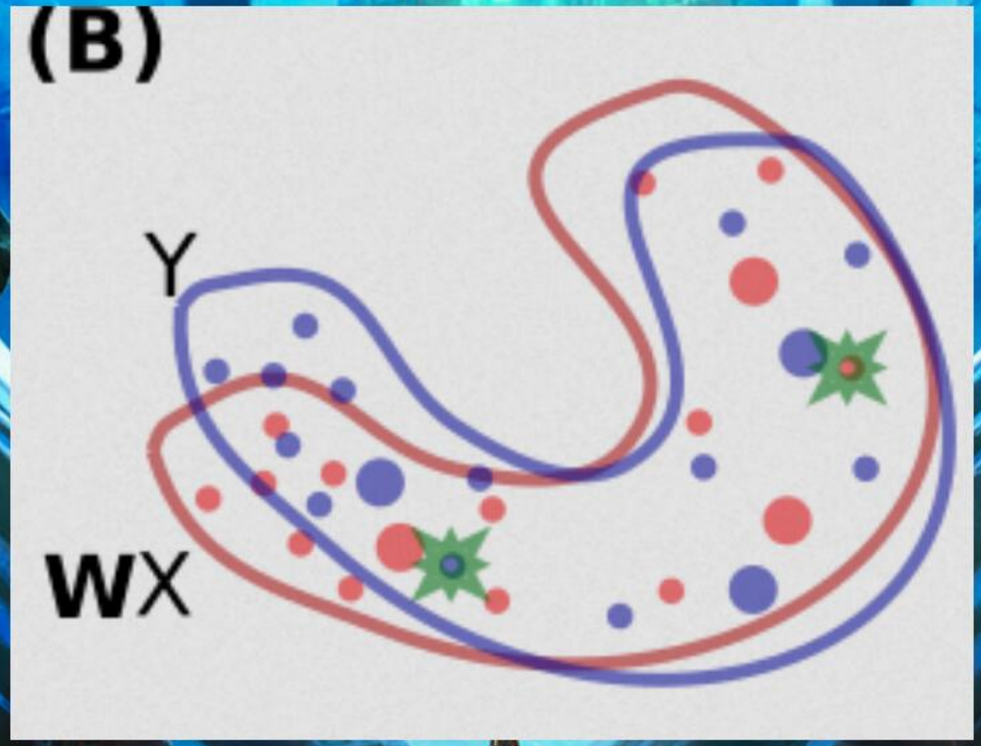
Comment fait-on ?



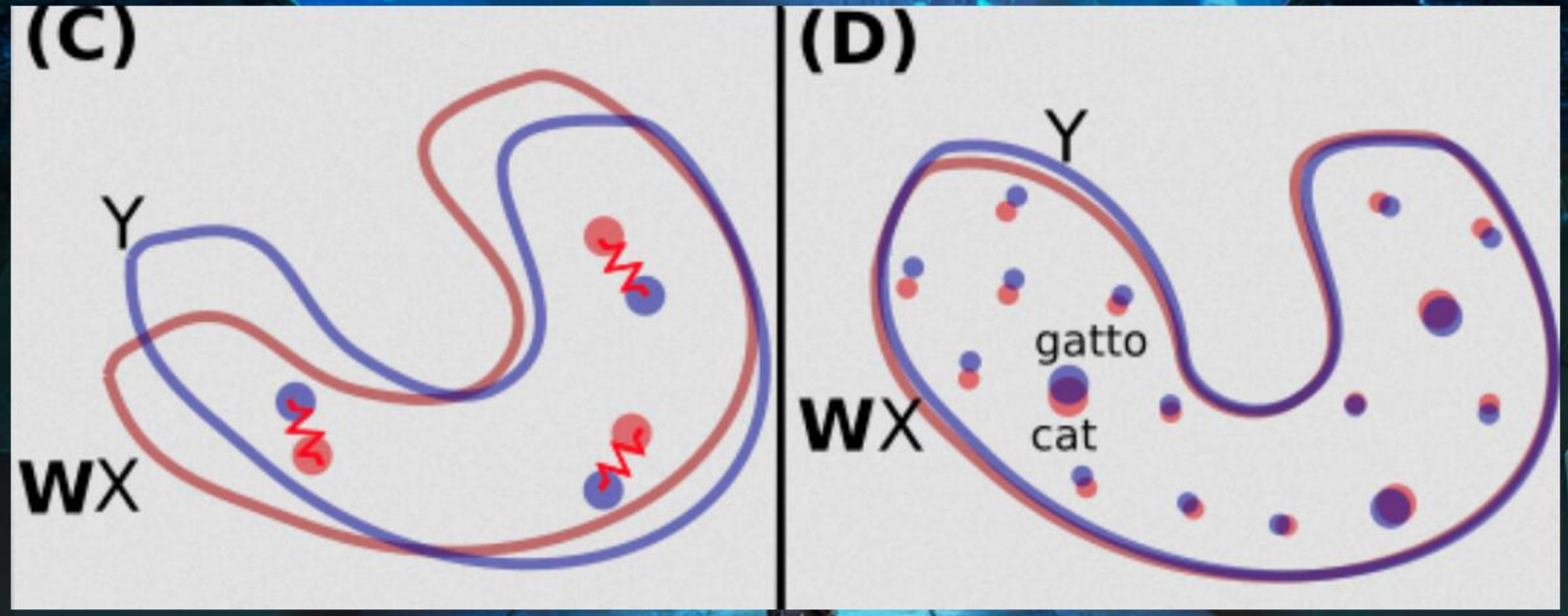
Similarité des espaces



Rotation de la source

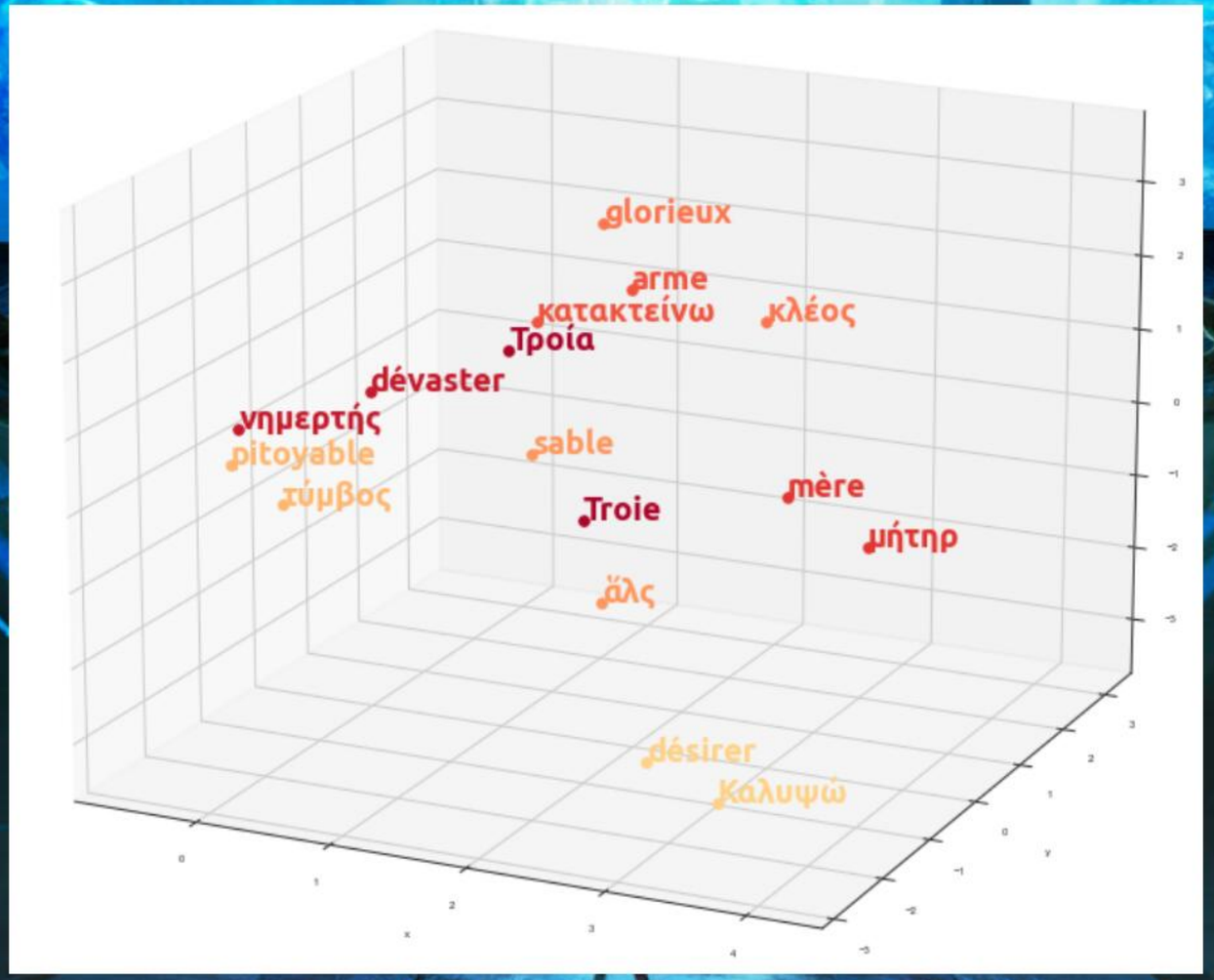


Rapprochement optimal et amplification




Expérience sur un micro-corpus
Entraînement sur l'*Iliade* et l'*Odyssée*
Stanford | NLP

Résultats préliminaires



Pourquoi c'est important

(et surprenant)

- on prouve que les langues sont **sémantiquement constituées de la même façon**,
- quand on a un **contenu similaire** (dans les diverses langues), on peut se passer davantage des données de masse,
- on peut mesurer **l'influence des mots** entre les langues.

Les Références

- **Conneau**, Alexis, et al. "XNLI: Evaluating cross-lingual sentence representations." *arXiv preprint arXiv:1809.05053* (2018). https://github.com/facebookresearch/XLM
- **Conneau**, Alexis, et al. "Word translation without parallel data." *arXiv preprint arXiv:1710.04087* (2017). https://github.com/facebookresearch/MUSE
- **Alammar**, Jay, "Visualizing A Neural Machine Translation Model (Mechanics of Seq2seq Models With Attention)", https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/
- **Perseus**, "Treebank Data", https://github.com/PerseusDL/treebank_data


Je vous remercie de votre attention.
marianne.reboul@ens-lyon.fr