



ENS DE LYON

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

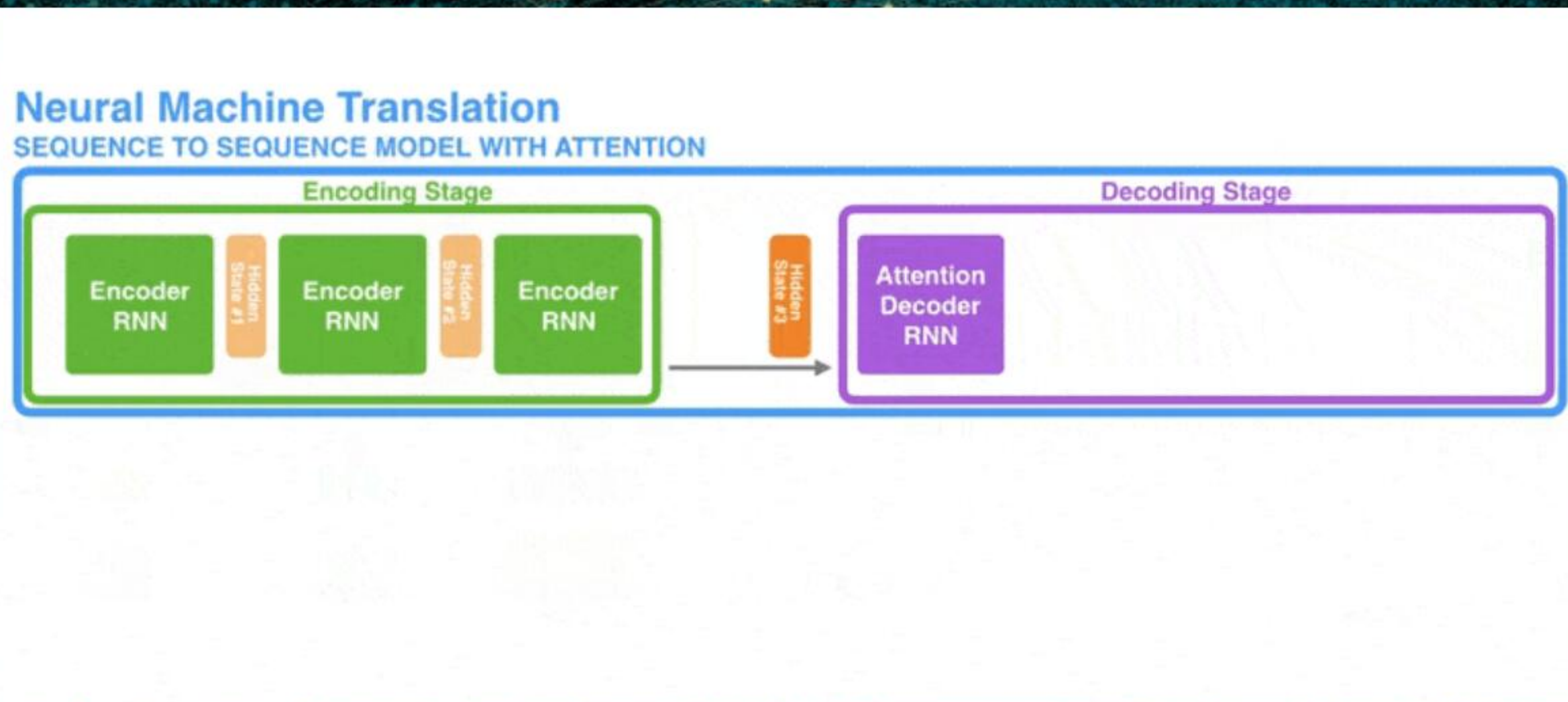
Deux phases :

1, l'entraînement

2, la prédiction



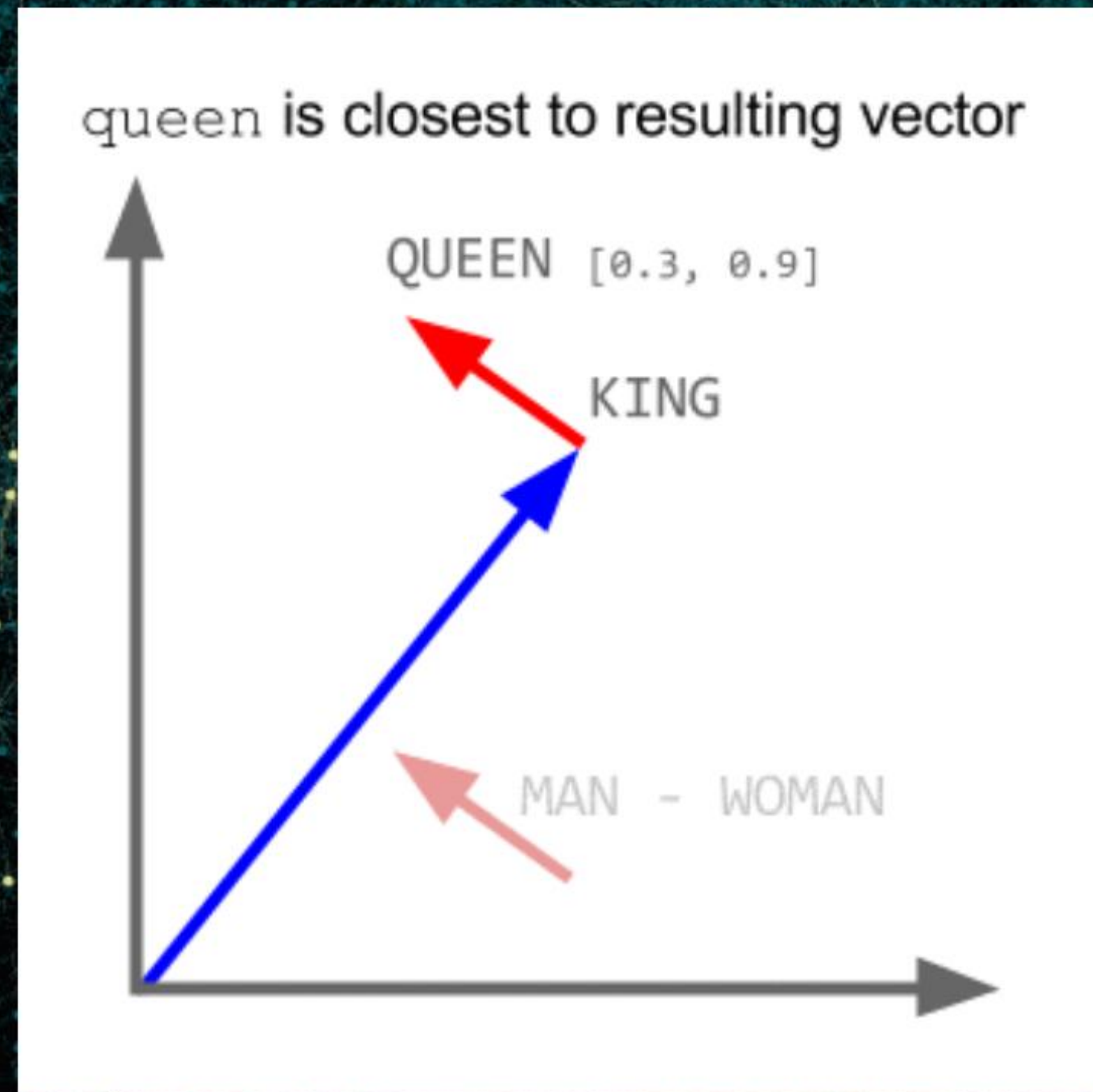
Création d'un modèle avec réseaux de neurones



Les mots en "embeddings"

Input

Je	0.901	-0.651	-0.194	-0.822	
suis	-0.351	0.123	0.435	-0.200	
étudiant	0.081	0.458	-0.400	0.480	



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 ?

John told Mary a story.

Jean a raconté une histoire à Marie.

Target sentence

Source sentence

	Jean	a	raconté	une	histoire	à	Marie
John							
told							
Mary							
a							
story							



Une de mes tentatives précédentes

L'Odyssée, chant . Ajouter une version

- High&Low Frequency
- Harax
- Potential text reuse

1000, Homère, chant 1

[1] ἄνδρα μοι ἔννεπε ,
[2] Μοῦσα , πολύτροπον , ὃς μάλα πολλὰ
πλάγχθη , ἐπεὶ
[3] Τροίης ἱερὸν πτολίεθρον ἔπερσεν · πολλῶν δ'
ἀνθρώπων ἴδεν ἄστεα καὶ νόον ἔγνω , πολλὰ δ' ὃ
γ' ἐν πόντῳ πάθεν ἄλγεα ὃν κατὰ θυμόν ,
ἀρνύμενος ἦν τε ψυχὴν καὶ νόστον ἐταίρων .
ἀλλ' οὐδ' ὧς ἐτάρους ἐρρύσατο , ἰέμενός περ ·
αὐτῶν νὰρ σφετέρῃσιν ἀτασθαλίῃσιν ὄλοντο .

×

1540, Jacques Peletier du Mans, chant 1

[1] Enseigne moi ,
[2] Muse le personnage
plein d' entreprise et savoir en son âge .
Lequel après qu' il a eu saccagé
[3] Troye la grand' , a longtemps voyagé
Et en errant les nilles a passées
D' hommes divers , et compris leurs pensée
qui a souffert maints travaux périlleux
dessus la mer . avec soin merveilleux

×

1584, /

[1]
[2]
Qui
usa
Qui
Dep
[3]
Il vi
De



A vibrant, blue-toned underwater scene. In the foreground, a warrior in a red and white tunic stands in a dark, stone-lined hallway, looking out towards a bright, sunlit underwater city. The city features tall, classical-style columns and buildings. In the background, a large, muscular merman with a beard and a crown-like headpiece stands on a ledge. Several fish, including a large whale-like creature, swim in the water. The overall atmosphere is mysterious and epic.

Du monolingue au multilingue

Une solution pour les langues rares ?



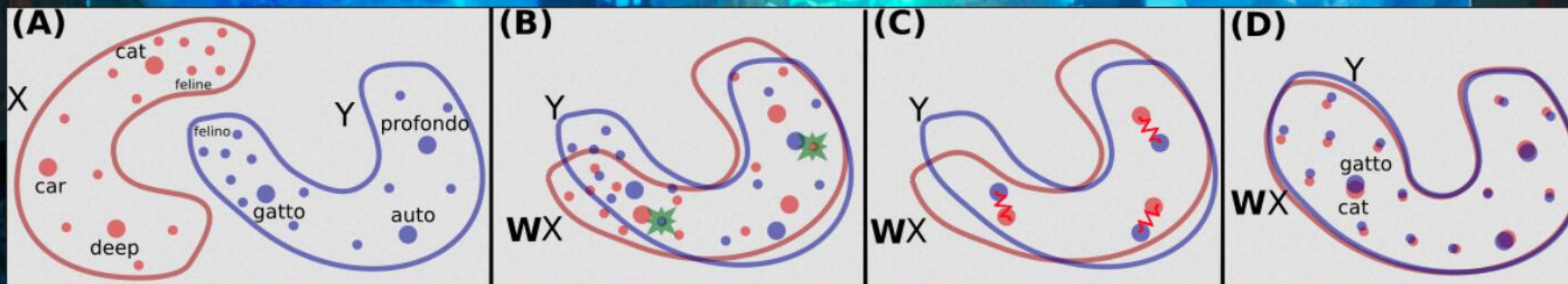
A vibrant, blue-toned underwater scene. In the foreground, a diver in a red and white suit stands on a stone platform. Behind them, a large, ornate archway frames a view of an ancient, submerged city with tall columns and domes. To the left, a muscular merman with a beard and a large fish tail is visible. The water is clear, with several large fish swimming in the background. The overall atmosphere is mysterious and ancient.

Un principe fondateur de base (discutable)

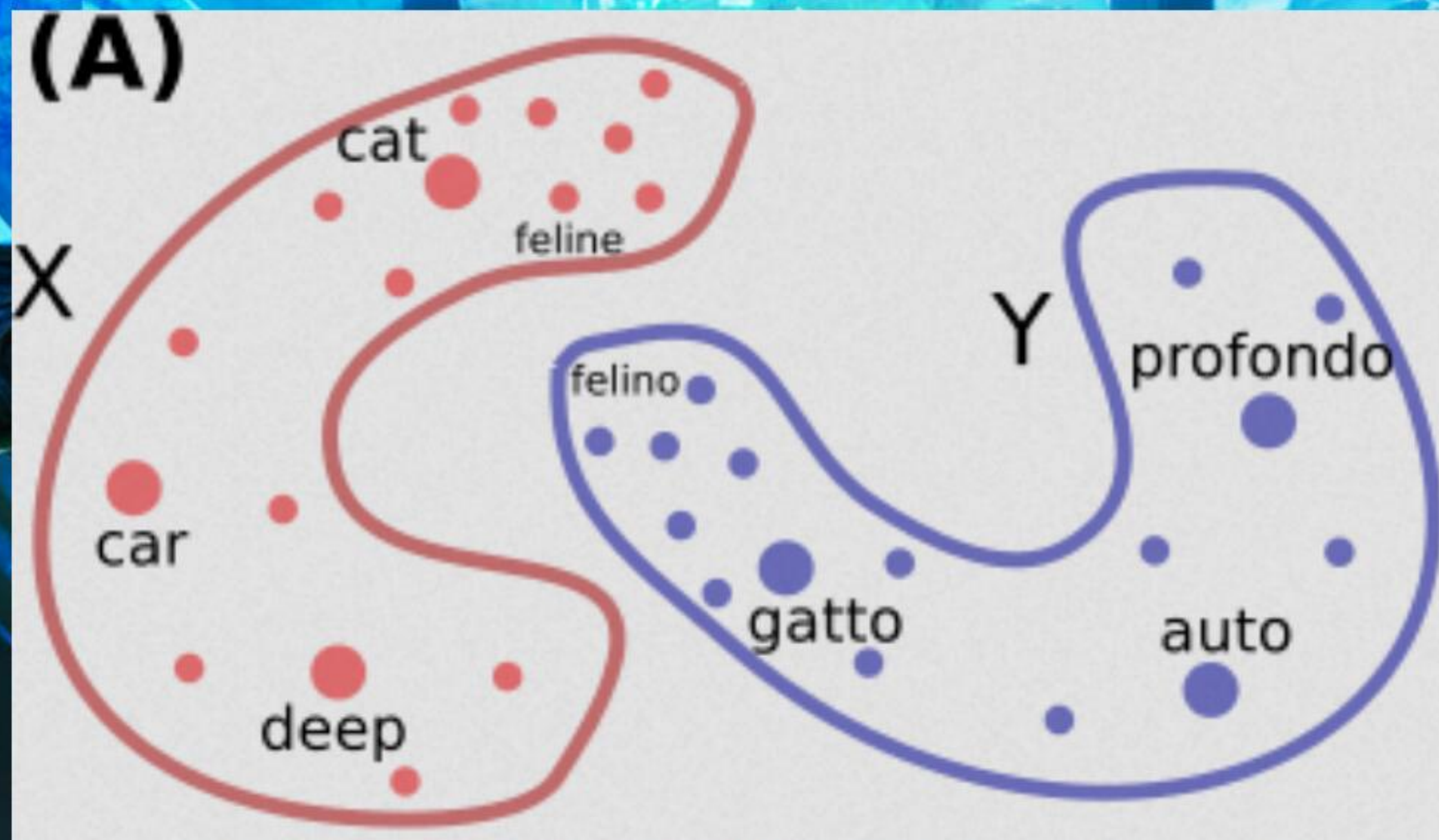
La ressemblance des
langues



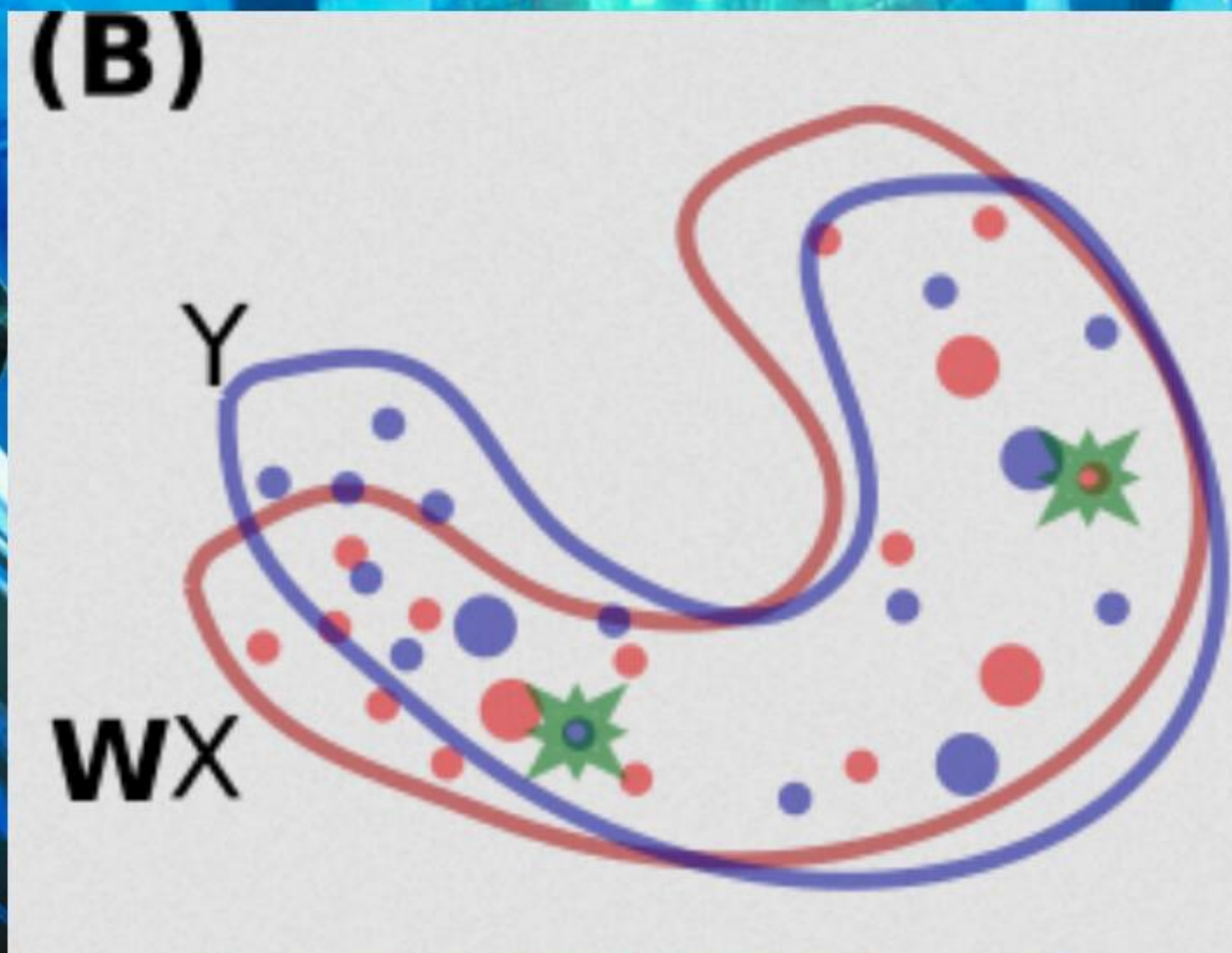
Comment fait-on ?



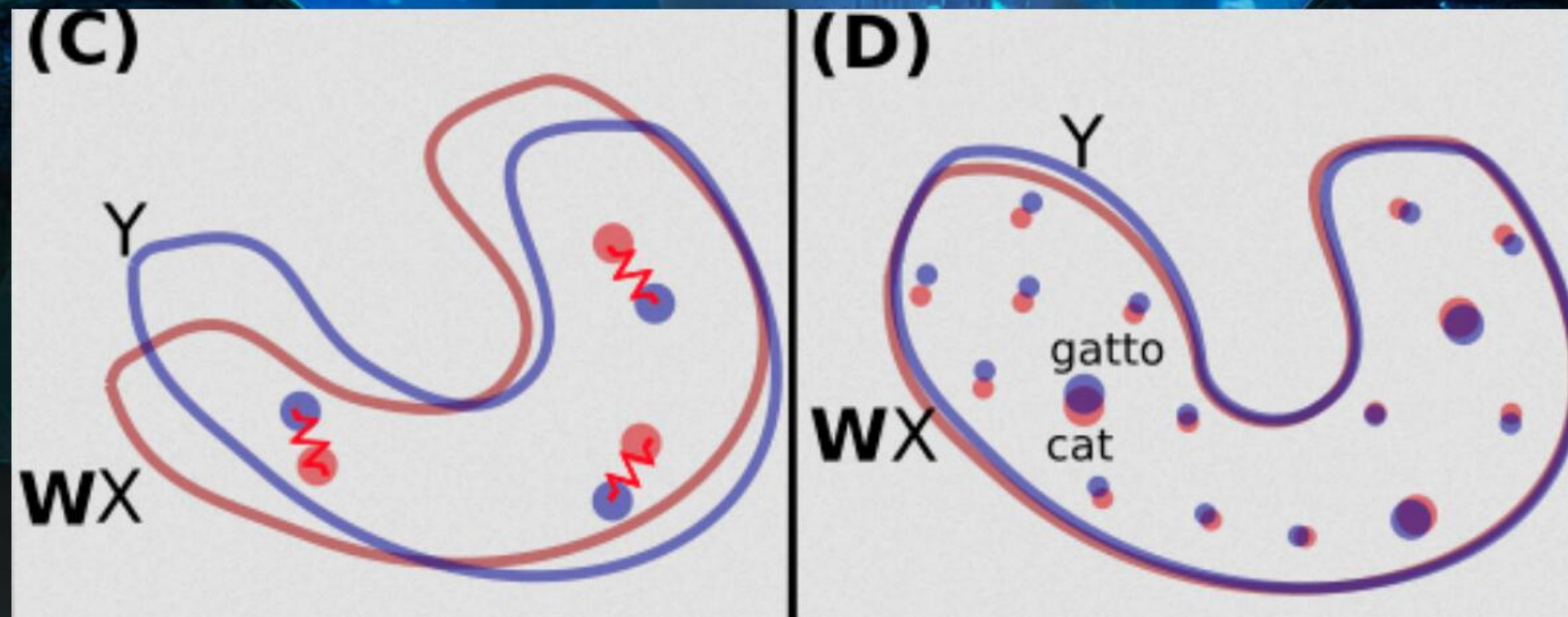
Similarité des espaces



Rotation de la source



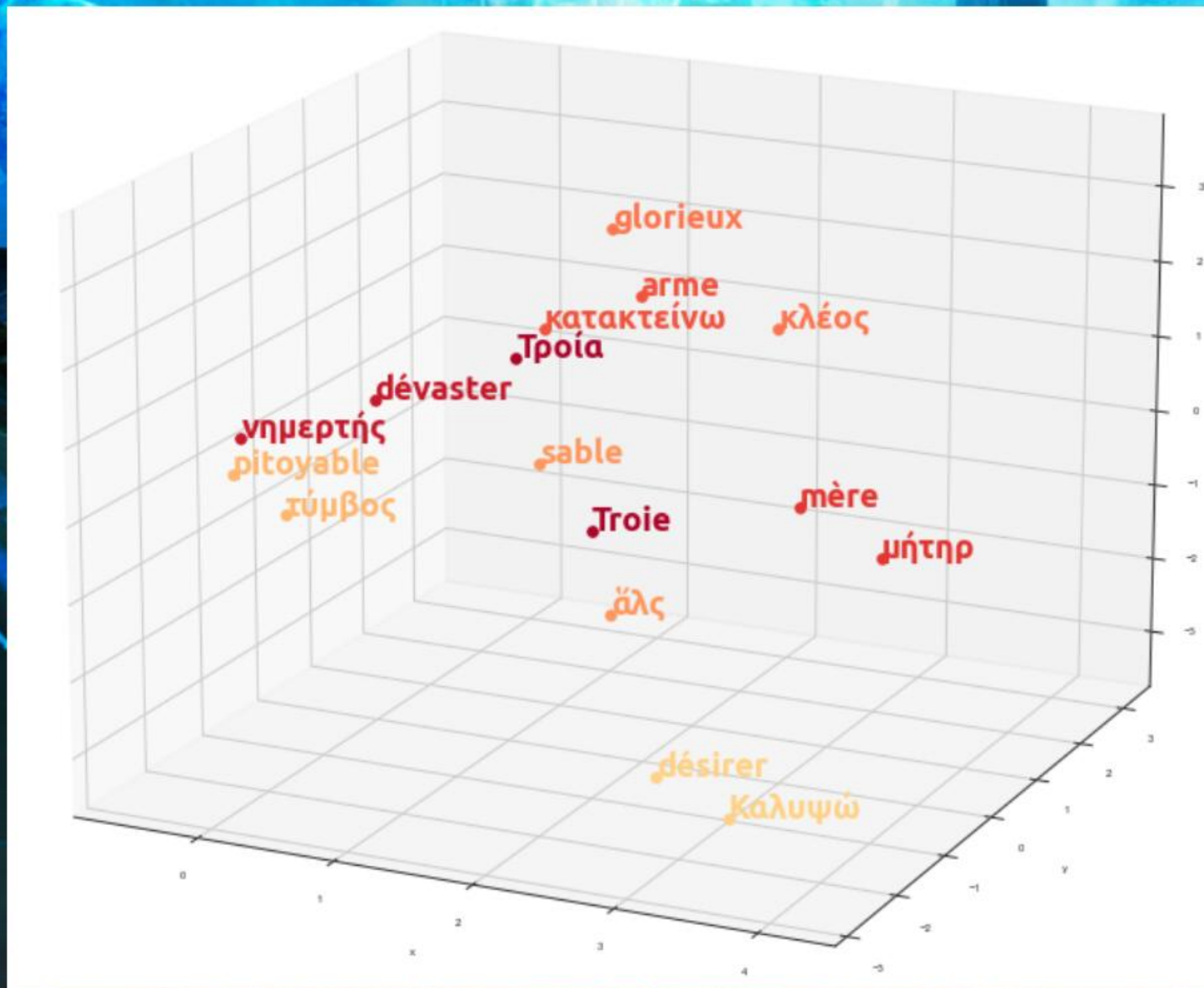
Rapprochement optimal et amplification



Expérience sur un
micro-corpus
Entraînement sur *Illiade* et
Odyssée



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

