Vers une extraction automatique de structures spatiales statiques pour le français

Application au corpus parallèle EN80jours

Antoine TARONI
Ludovic MONCLA
Frédérique LAFOREST

Atelier Humanités Numériques Spatialisées à l'ère des graphes de connaissances et des grands modèles de langage, Conférence SAGEO - Avignon, 21 mai 2025

Les Constructions Locatives de Base (CLB)

- Objets d’étude récurrents dans les études sur le langage et la cognition spatiale (Levinson, Wilkins, 2006)

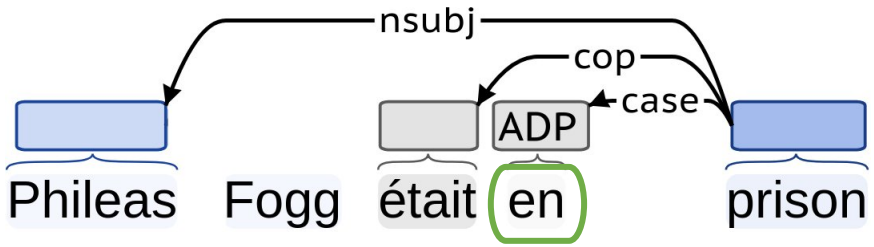
- CLB = réponse typique à la question « Où est X (par rapport à Y) ? »

« Phileas Fogg *était* en prison »

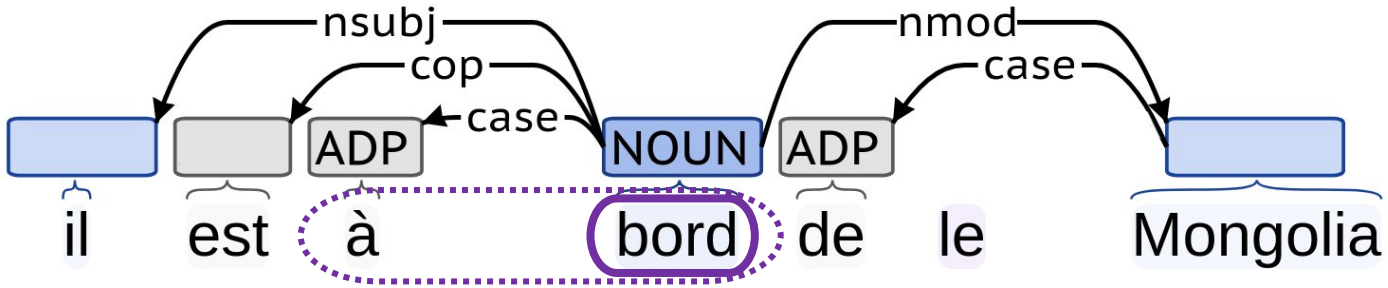
« Il *était* à bord du Mongolia »

issus de (Lecuit et al., 2011, corpus EN80jours)

Les Constructions Locatives de Base (CLB)

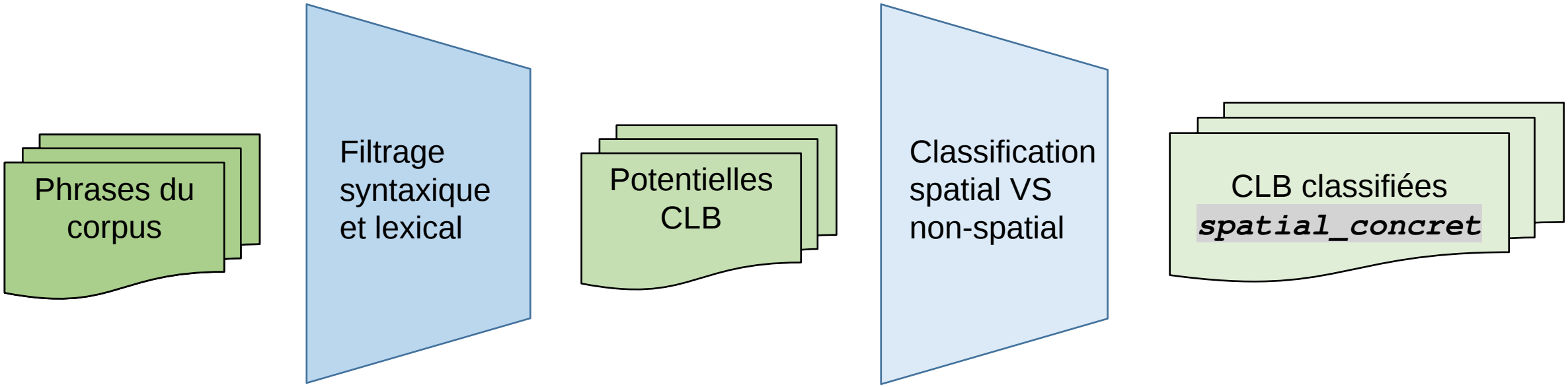


« simple » préposition

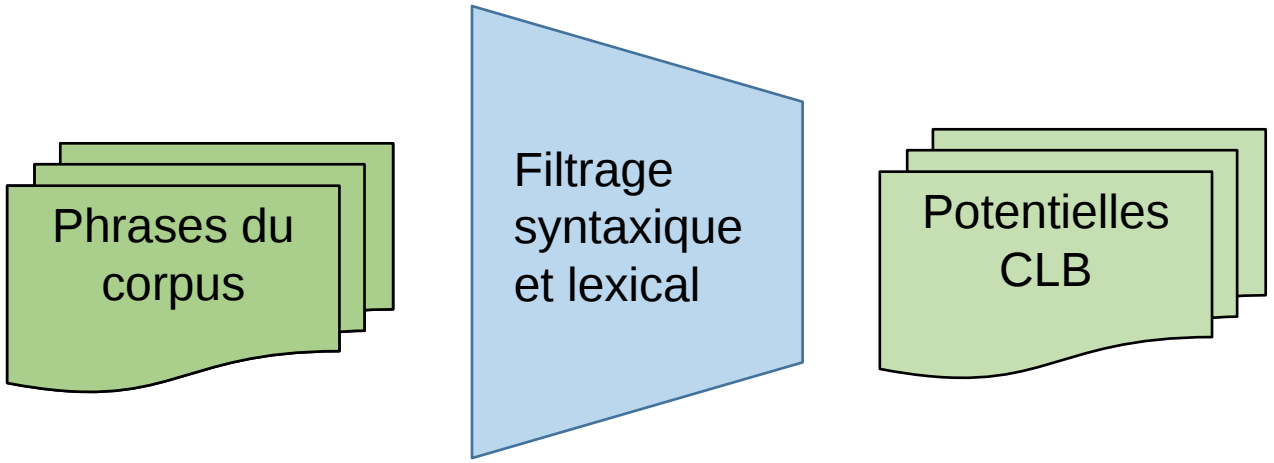


Nom de Localisation Interne (NLI)

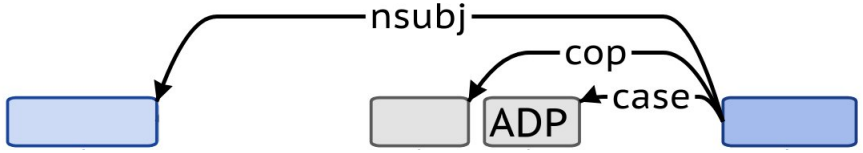
Extraction des CLB (Viechnicki 2024)



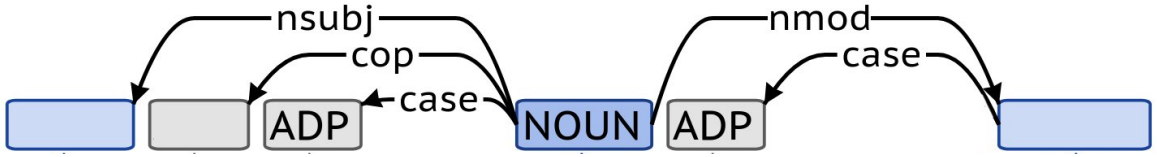
Filtrage syntaxique et lexical



Motif 1 :



Motif 2 :

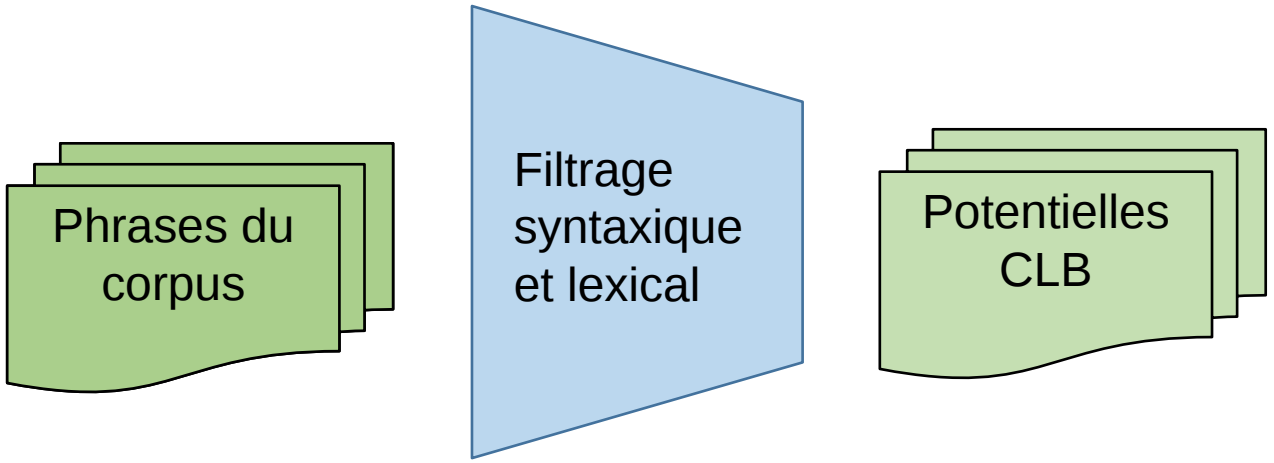


Langage SEMGREX : Stanford CoreNLP (Manning et al., 2014)

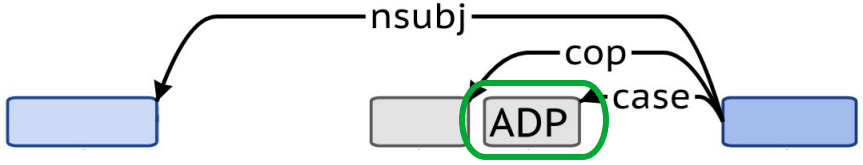
```
{}=site
>nsubj {}=cible
>cop {pos:/VERB|AUX/}=verbe
>case {pos:/ADP|ADV/}=prep
```

```
{tag:/NOUN|PROPN/}=ILN
>/^nmod.*/ {}=site
>nsubj {}=cible
>cop {tag:/VERB|AUX/}=verbe
>case {tag:/ADP/}=prep
```

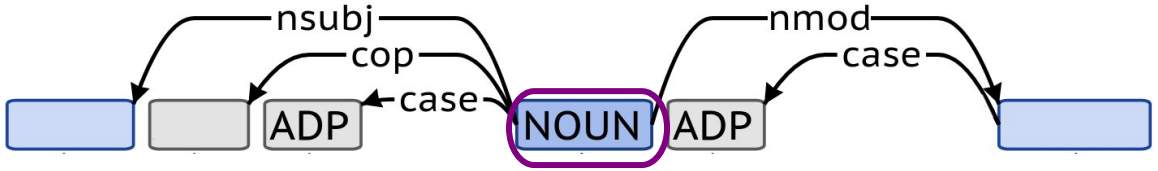
Filtrage syntaxique et lexical



Motif 1 :



Motif 2 :



Liste fermée des prépositions « simples »

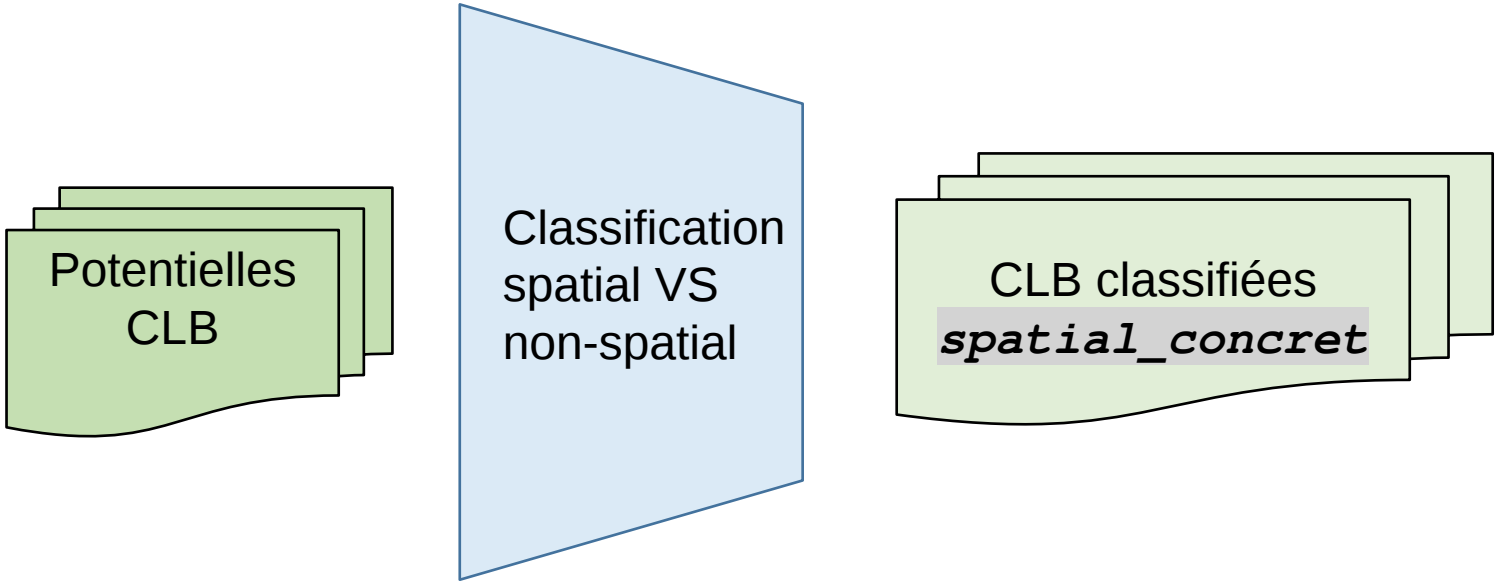
au-dessus, chez, contre, dans, derriere, devant, hors, par-derriere, proche, sous, sur ...

Liste ouverte des NLIs

avant, bord, couchant, dessus, est, levant, nord, ouest, occident, orient, sud ...

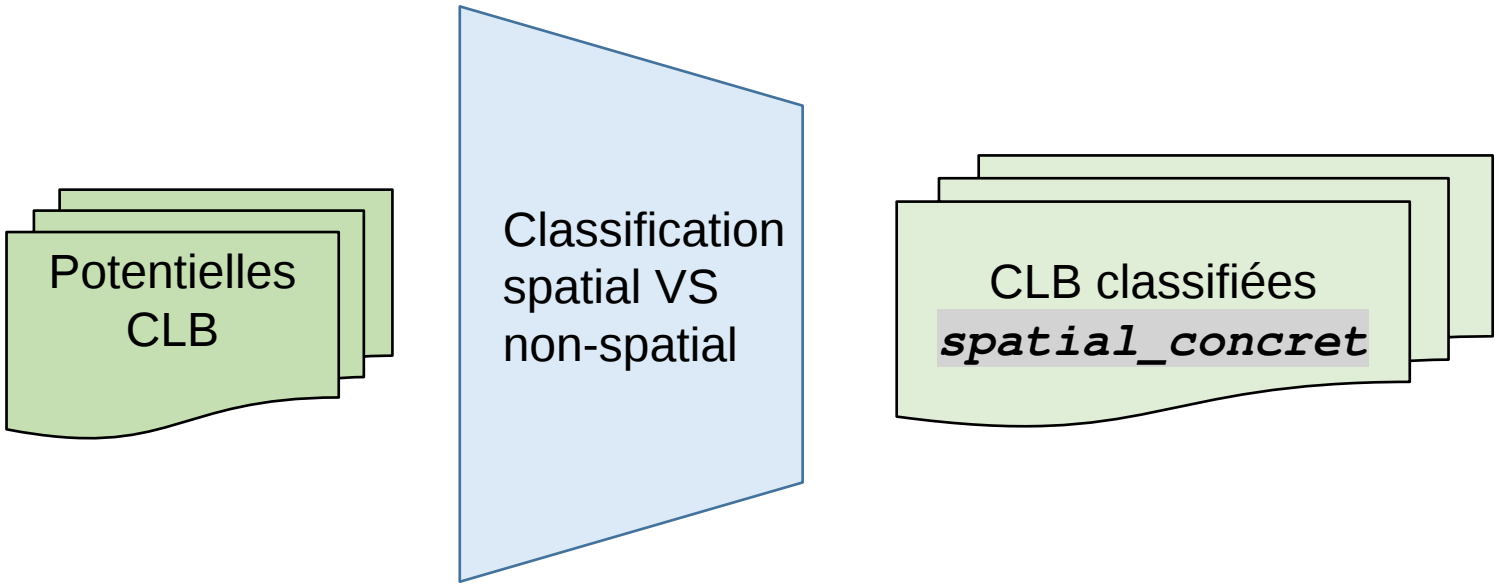
Classification ***spatial_concret*** VS ***non-spatial***

- Classification binaire
- *Few-shot* avec le modèle de langage pré-entraîné ***llama3 :70b***
- Prompt rédigé en français
- ***spatial_concret*** : X et Y sont des entités concrètes
  non-spatial : X et/ou Y sont des entités abstraites



Classification ***spatial_concret*** VS ***non-spatial***

Spatial	Non-spatial
✓ X et Y sont des entités concrètes du monde physique. ✓ Je peux les toucher ou les pointer du doigts. ✓ La phrase répond à la question « Où est X par rapport à Y »	✓ X et/ou Y sont des entités abstraites ✓ La phrase ne contient pas d'information spatiale ✓ La phrase ne te permet pas de localiser X par rapport à Y



Classification ***spatial_concret*** VS ***non-spatial***

```
Exemples :
1. "Le chat s'est assis sur le piano"
Tu retournes donc uniquement le JSON suivant :

{
  "cible":"chat",
  "site":"piano",
  "marqueur_spatial":"sur",
  "méthode": "La cible (chat) et le site (piano) sont des entités
  ↪  concrètes, du monde physique : tu peux les pointer du doigt, ou
  ↪  les toucher. La phrase permet de localiser le chat par rapport au
  ↪  piano. La phrase répond donc bien à la question <Où est situé le
  ↪  chat par rapport au piano ?>",
  "tag": "spatial_concret"
}
```

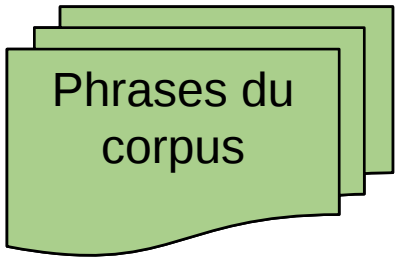
Classification ***spatial_concret*** VS ***non-spatial***

```
2. "Il travaille sur le dossier"
Tu retournes donc uniquement le JSON suivant :

{
  "cible":"Il",
  "site":"dossier",
  "marqueur_spatial":"sur",
  "méthode": "La cible (Il, renvoie probablement à une personne) et le
  ↪  site (dossier) sont des entités concrètes, du monde physique : tu
  ↪  peux les pointer du doigt, ou les toucher. MAIS la phrase ne
  ↪  contient pas de relation spatiale entre la cible et le site.",
  "tag": "non_spatial"
}
```

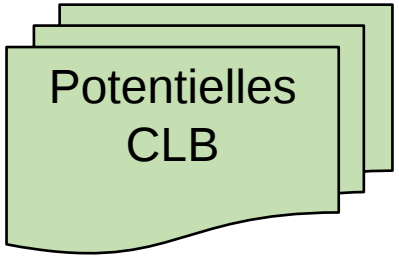
Évaluation

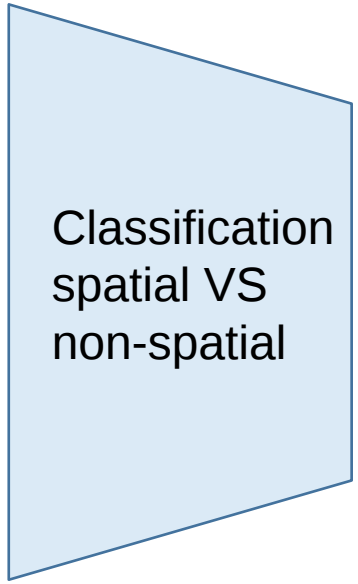
4390 phrases



recall inconnu

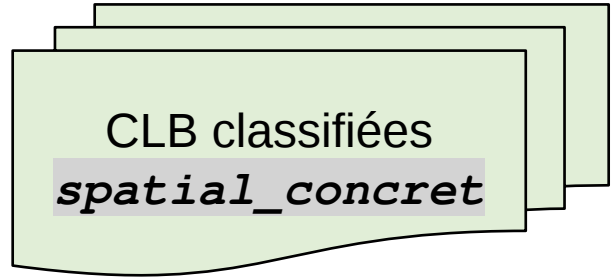
46 phrases ont le motif lexico-syntaxique





précision = 0.69
recall = 0.90
F-score = 0.78

spatial_concret **:**
tp: 9
fp: 4
non-spatial:
tn: 32
fn: 1



Analyse des Constructions Locatives de Base

- *à* (4 occurrences) est traduit en *at, on* ou *in* pour ce qui est de l'anglais, et *an* ou *in* en allemand
- *dans* (2 occurrences) est traduit en *in* ou *on* en anglais, et *in* ou *bei* en allemand
- Les CLB en français ne sont pas systématiquement traduites par une CLB en EN ou DE

a. Et comme cela, nous sommes **à** Suez? (n614)
'*So this is Suez?*' [EN]

b. ce grand triangle renversé dont la base est **au** nord [...] (n831)
'*with its base **in** the north [...]*' [EN]
'*dessen Grundlinie **im** Norden [...] liegt*' [DE]

c. tous quatre étaient **à** bord. (n3845)
'*befanden sich alle vier **an** Bord.*' [DE]

Conclusions

- Extraction automatique des Constructions Locatives de Base
- Filtrage syntaxique + LLM prompting

Avantages	Inconvénients
• Basé sur des règles • Ajout de règles dans le pipeline simple • Outils *off-the-shelf :* analyse en dépendance via CoreNLP, recherche Semgrex, *few-shot prompting*	• la précision est garantie par le filtrage lexico-syntaxique : ce n'est pas un pipeline d'extraction d'information spatiale • Établir le catalogue des motifs syntaxiques demande d'effectuer des itérations entre notre corpus et *l'output* du processus

- Application de la méthode à un plus grand corpus et extension du jeu de données « gold »
- Ajout de règles lexicales « exclusives » ou « inclusives » en parallèle de la classification automatique
- Vers l'inclusion des verbes autres qu'à valeur copulative, l'extraction de la dynamique, des adverbes de cadrage, etc

Bibliographie

- Aurnague M., Boulanouar K., Nespoulous J.-L., Borillo A., Borillo M. (2000). Spatial semantics: the processing of internal localization nouns. Cahiers de Psychologie Cognitive-Current Psychology of Cognition
- Landau B. (2024). Are spatial terms rooted in geometry or force-dynamics? yes. Cognitive Processing, vol. 25, p. 85–90.
- Lecuit E., Maurel D., Vitas D. (2011). En80jours [corpus]. https://hdl.handle.net/11403/en80jours/v1.
- Levinson S. C., Wilkins D. P. (2006). Grammars of space: Explorations in cognitive diversity (vol. 6). Cambridge University Press.
- Manning C. D., Surdeanu M., Bauer J., Finkel J. R., Bethard S., McClosky D. (2014) The stanford corenlp natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, p. 55–60.
- Viechnicki P., Duh K., Kostacos A., Landau B. (2024). Large-scale bitext corpora provide new evidence for cognitive representations of spatial terms. In Y. Graham, M. Purver (Eds.), Proceedings of the 18th conference of the european chapter of the association for computational linguistics, p. 1089–1099. ACL

Vers une extraction automatique de structures spatiales statiques pour le français

Application au corpus parallèle EN80jours

Antoine TARONI

antoine.taroni@insa-lyon.fr