Vers une extraction automatique de structures spatiales statiques pour le français

Application au corpus parallèle EN80jours

Antoine TARONI[1], **Ludovic MONCLA**[1], **Frédérique LAFOREST**[1]

INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205, 69621 Villeurbanne

{prenom}.{nom}@liris.cnrs.fr

ABSTRACT. *Basic Locative Structures (BLCs) have long served as a controlled framework for investigating space in language and cognition. Recently, (Viechnicki* et al.*, 2024) proposed an automatic BLC extraction method from parallel corpora using English as the pivot language. In this study, we adapt and apply this methodology to the EN80Jours parallel corpus, with French serving as the pivot language. By combining syntactic and lexical pattern searches with few-shot learning using Pretrained Language Models to filter for spatial expressions, we achieve precise retrieval of a refined set of locative constructions. Finally, we make initial observations of the inter-linguistic variability of spatial markers within BLCs by aligning French spatial markers with their German and English counterparts. This preliminary analysis highlights challenges in aligning BLCs across languages and suggests the need for further exploration of linguistic variations and the integration of lexical rules to refine classification.*

MOTS-CLÉS : *sémantique spatiale, construction locative de base, prépositions, détection d'expression locative*

KEYWORDS: *spatial semantics, basic locative constructions, prepositions, location phrase detection*

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

1. Introduction

La sémantique spatiale reflète la relation entre notre système cognitif et l'espace. En effet, notre perception de celui-ci, et donc notre capacité à y naviguer, façonnent nos représentations non-linguistiques et linguistiques. (Regier, Zheng, 2007 ; Landau, 2024). Les Constructions Locatives de Base (CLB, *Basic Locative Structures* en anglais) (Levinson, Wilkins, 2006) permettent de situer une entité cible, généralement petite et mobile, relativement à une entité site, bien souvent plus grande, immobile, et dont la position connue est propice à l'ancrage du système de référence. Elles permettent une analyse restreinte au noyau sémantique de divers marqueurs spatiaux (Aurnague, Vieu, 2013). Du point de vue syntaxique, une CLB comporte en français, dans l'ordre : la cible, le verbe à valeur copulative (typiquement *être*), et le syntagme prépositionnel (SP). Le SP est introduit par un marqueur spatial, et se termine par le site. Le marqueur spatial peut être de plusieurs natures ; une simple préposition (*dans, sur, derrière*), ou bien une locution prépositionnelle (*à droite, au levant*), et présente alors un Nom de Localisation Interne (NLI) (respectivement *droite* et *levant* dans les exemples précédents) (Aurnague *et al.*, 2000). Si les prépositions constituent en français une classe fermée, les locutions prépositionnelles forment une classe ouverte, ce qui complexifie l'automatisation de leur repérage. Les exemples (1a) et (1b) montrent deux CLB issues du corpus parallèle `EN80jours` (Lecuit *et al.*, 2011) basé sur le roman de Jules Verne *Le Tour du monde en quatre-vingt jours* (1872). Les marqueurs spatiaux sont en gras et l'identifiant de la phrase est donné entre parenthèse.

(1) a. Phileas Fogg était **en** prison. (n4038)
b. [...] s'il est **à bord** du Mongolia. (n474)

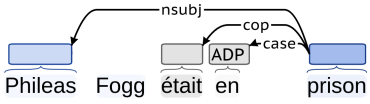
Nous présentons dans ce travail préliminaire un processus d'extraction automatique des CLB, et l'appliquons à un corpus aligné, avec l'objectif de comparer l'expression de la localisation dans trois langues.

2. Méthodologie pour l'extraction des CLB

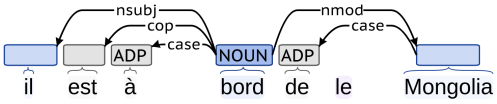
Dans cette étude, nous nous inspirons de (Viechnicki *et al.*, 2024), travail dans lequel les auteurs introduisent une méthodologie pour le repérage automatique du motif syntaxique basé sur une simple préposition (Figure 1a), pour l'anglais. Nous adaptons ce processus de deux étapes au français, et ajoutons un second patron dédié aux motifs comportant une locution prépositionnelle (Figure 1b).

La première étape consiste en un filtrage syntaxique et lexical. Ce module vise à identifier toutes les occurrences des motifs définis aux Figure 1a et 1b. Afin de simplifier notre travail, nous nous limitons dans cette contribution aux

motifs impliquant le verbe copulatif *être* et imposons aux CLB potentielles de faire suivre la copule directement par le marqueur spatial. Enfin, nous nous limitons aux motifs affichant une préposition figurant dans la liste en Annexe A, ou bien une locution prépositionnelle centrée sur un NLI (voir Annexe A).



(a) comportant une simple préposition



(b) comportant une locution prépositionnelle

Figure 1. Motif de dépendance syntaxique des CLB en français

La seconde étape vise à classifier les motifs selon qu'ils soient effectivement des CLB, ou bien ne contiennent pas de relation spatiale. Ce module repose sur un modèle de langage génératif, employé en *few-shot* pour ladite tâche de classification. Nous considérons donc deux classes : `spatial_concret` et `non_spatial`. En effet, l'Exemple 2a répond à la question "Où est le Proctor (cible) par rapport au train (site) ?". Il s'agit donc d'une CLB et est annoté `spatial_concret`. Mais les constructions n'impliquant pas d'entités physiques concrètes, ou bien n'introduisant pas de relation spatiale comme l'Exemple 2b, sont annotées `non-spatial`.

(2) a. Ce Proctor est **dans** le train ! [...] (n3137)
b. Le train y sera **dans** une heure. (n3355)

Le prompt donné en entrée du modèle de langage contient une description de la tâche ainsi que des exemples, comme ceux fournis en Annexe B, suivi de la CLB potentielle à classifier (augmentée de deux tokens en amont et aval).

3. Expérimentations

Nos premières expérimentations se font sur le corpus parallèle `EN80jours`. Nous effectuons une analyse automatique en dépendance sur l'ensemble du corpus avec l'outil `Stanford CoreNLP` (Manning *et al.*, 2014). L'étape de repérage des motifs syntaxiques des CLB retourne ensuite 46 occurrences. Une analyse manuelle de ces occurrences nous permet de confirmer qu'elles sont toutes correctes, l'analyse en dépendance automatique de `CoreNLP` est donc ici

satisfaisante en terme de précision. Cependant, nous n'avons pas étudié son rappel. Nous n'avons par ailleurs pas trouvé de mesure du taux d'occurrence des CLB en français dans la littérature.

Nous avons utilisé le modèle de langage *open weights* `llama3:70b` pour la classification `spatial_concret` / `non_spatial`. Les 46 occurrences ont été annotées manuellement afin de pouvoir évaluer les performances. Il en ressort que seulement 10 sont des constructions locatives de base. Le module de classification obtient une précision de 0.69, un recall de 0.90 et un F-score de 0.78. Nous remarquons que tous les faux positifs (au nombre de quatre) ont pour sujet syntaxique le pronom démonstratif *ce*, comme l'illustre l'Exemple 3. Ce pronom ne renvoyant pas à une entité concrète, ces occurrences doivent être annotées `non_spatial`. L'application d'une simple règle lexicale "exclusive" permet ainsi, dans notre cas d'étude, d'atteindre une précision de 1. Cela ouvre la réflexion sur l'intégration de règles lexicales ciblant le sujet syntaxique pour affiner la classification.

(3) C' était **sur** cette contrée [...] (n954)

Enfin, nous nous sommes intéressés à la variabilité inter-linguistique des marqueurs spatiaux de CLB. Nous avons pour cela aligné les marqueurs spatiaux du français vers l'allemand et l'anglais. Ainsi, nous observons que *à* (4 occurrences) est traduit en *at, on* ou *in* pour ce qui est de l'anglais, et *an* ou *in* en allemand. De la même manière, *dans* (2 occurrences) est traduit en *in* ou *on* en anglais, et *in* ou *bei* dans la version allemande du corpus. Nous remarquons également que dans deux cas sur 10, si une CLB spatiale a été détectée depuis la langue pivot, la traduction anglaise comporte des structures d'une autre nature (Exemples 4a et 4b). La méthodologie de (Viechnicki *et al.*, 2024) ne garantit en effet pas un alignement des CLB de la langue pivot sur des CLB dans toutes les langues cibles.

(4) 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]

Par ailleurs, dans trois alignements avec l'allemand, on retrouve deux fois un verbe à valeur copulative autre que *sein* (*sich befinden*, Exemple 4c), et un verbe de position (*liegen*, Exemple 4b). Cela met ainsi en lumière une limite de

la seule recherche des CLB comportant le verbe *être*, car d'autres verbes comme *se trouver* ou *se situer* semblent avoir une valeur copulative en français.

4. Conclusion

Nous avons présenté un procédé d'extraction simple mais strict des CLB en langue française combinant filtrage syntaxique et classification *few-shot*. Ces structures fournissent un cadre d'analyse contrôlé du noyau sémantique des marqueurs spatiaux. Toutefois, leur extraction du corpus `EN80jours` semble mettre en évidence un faible taux d'occurrences. Une mesure du recall de la méthode est toutefois nécessaire pour le confirmer. Par ailleurs, l'analyse des occurrences identifiées montre une fréquente intégration de ces motifs dans des phrases plus complexes non limitées aux CLB seules. Enfin, le volume des CLB extraites du corpus étant trop faible pour quantifier la variabilité de l'alignement des marqueurs spatiaux vers l'anglais et l'allemand, il ne s'agit ici que d'un premier pas vers une reproduction de (Viechnicki *et al.*, 2024) avec le français comme langue pivot. Ces résultats préliminaires ouvrent la porte à une analyse de leur apport à l'échelle phrastique, puis discursive.

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*, vol. 19, n° 1, p. 69–110.

Aurnague M., Vieu L. (2013). Retour aux arguments: pour un traitement «relationnel» des prépositions spatiales. *Faits de Langues*, vol. 42, n° 1, p. 17–38.

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.

Regier T., Zheng M. (2007). Attention to endpoints: A cross-linguistic constraint on spatial meaning. *Cognitive Science*, vol. 31, n° 4, p. 705–719.

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.

Annexe A. Marqueurs spatiaux employés

Liste de prépositions : à, après, au-dedans, au-delà, au-dessous, au-dessus, au-devant, auprès, autour, avant, avec, chez, ci-contre, ci-devant, contre, dans, dedans, dehors, depuis, derrière, devant, en, entre, face à, hors, jusqu'à, loin, par, par-delà, par-derrière, par-dessus, par-devant, parmi, proche, près, sous, sur, vers

Sélection de quelques Noms de Localisation Interne: alentours, arrière, avant, bas, bord, centre, confins, contact, contrebas, couchant, côté, dedans, dehors, delà, dessous, dessus, deçà, direction, droite, environs, est, extérieur, face, fond, gauche, haut, intérieur, levant, long, milieu, nord, occident, orient, ouest, proximité, périphérie, sommet, sud, surplomb, travers, écart

Annexe B. Échantillon du contexte donné dans le prompt

```
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"
}

2. "Il travaille sur le dossier"
tTu 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"
}
```