Vers l'automatisation du classement des séquences candidates à la catégorie des prépositions complexes en français

Olivier Kraif[1,*], *Denis* Vigier[2] et *Ghayoung* Kahng[2]

[1] LIDILEM, Univ. Grenoble Alpes
[2] ICAR UMR 5191, Université Lumière Lyon II

[*]olivier.kraif@univ-grenoble-alpes.fr

Résumé. La présente contribution propose de nouvelles avancées dans le but de relever l'un des défis majeurs posé par la classe des prépositions complexes à la communauté des chercheurs en linguistique : la possibilité d'en dresser une liste. En vue du développement d'une méthode entièrement automatisée pour extraire des candidats appartenant à cette classe, nous proposons une étude expérimentale où sont croisées deux approches pour caractériser les prépositions complexes : d'une part, l'application d'une grille multicritère proposée par Vigier & Kahng (2022) (suite à Stosic & Fagard (2019)) qui nécessite de combiner des tests manuels avec des mesures statistiques extraites de corpus ; d'autre part l'extraction automatisée d'une série d'indices textométriques, dont certains sont originaux, comme le taux d'insertion ou une mesure de dispersion composite. Nos observations montrent que sur une liste de candidats comportant de nombreux intrus, quelques indices peuvent se révéler discriminants, tels que le taux d'insertion, la dispersion par année ou par fichier, ainsi que, dans une moindre mesure, des mesures d'association statistique comme le t-score et log rapport de vraisemblance. Mais nous montrons également, en étudiant une liste de candidats construits avec la préposition « en », que ces observations sont à nuancer et dépendent notamment du comportement syntaxique des prépositions mises en jeu.

1 Introduction

Depuis les années quatre-vingt-dix, un nombre important de travaux portant sur les prépositions simples a été publié, parmi lesquels Berthonneau & Cadiot (éds. 1991), Kupferman (éd., 1996), Cadiot (1997), de Mulder & Stosic (éds. 2009), Leeman (éd. 2006, 2007, 2008), Vigier (éd. 2013). Ces travaux ont conduit à un certain consensus parmi les linguistes quant aux caractéristiques constitutives de cette classe, que ce soit en termes de critères d'identification (critères syntaxiques, morphologiques, sémantiques, …) ou en termes de structures internes de la catégorie.

Parallèlement à ces recherches, les séquences plus ou moins figées telles que *quant à*, *en voie de*, *sous réserve de*, etc. comportant plusieurs unités lexicales dont au moins une préposition, et le plus souvent aptes à commuter avec des prépositions simples, ont aussi fait l'objet de publications - mais en nombre plus restreint. Parmi ces études on citera pour les « prépositions complexes » : Borillo (2001), Stosic & Fagard (2019), Vigier & Kahng (2022) …, les « prépositions composées » : Gross (1981), Borillo (1998) …, les « locutions prépositives » : Adler (2001)... Pour ce qui regarde notre travail, nous inspirant de Leeman (2006), nous avons choisi de retenir la notion de « préposition complexe » (désormais PrepComp) pour désigner aussi bien « des mots graphiquement séparés mais formant un tout entièrement figé » ("quant à", "à l'instar de" ...) » que « des groupes de mots formant une séquence non entièrement figée » comme "hors de proportion avec", "sous réserve de", "à l'insu de" … (ibid :10).

Stosic & Fagard (2019) ont récemment proposé une étude dont l'un des principaux mérites a été de relancer les travaux sur les prépositions complexes en linguistique française[1] en proposant une synthèse des études antérieures et en ouvrant de nouvelles perspectives. Soucieux de pointer les défis qui attendent encore d'être relevés en vue de faire progresser notre connaissance de cette sous-classe des prépositions, les deux auteurs soulignent la nécessité d'en dresser une liste[2]. Deux verrous expliquent selon eux que l'on n'y soit pas encore parvenu : la difficulté de définir de manière consensuelle la sous-classe des

PrepComp, d'une part ; le nombre important d'exemplaires qu'elle réunit, d'autre part : plusieurs centaines si l'on en croit Borillo (1991, 1997) et Le Pesant (2006) pour le français, Huddelston & Pullum (2002) pour l'anglais.

Le premier verrou signalé par les auteurs nous paraît aujourd'hui levé. Leur étude qui mobilise le cadre théorique de la théorie du prototype dans sa version standard (Kleiber 1990) constitue en effet une proposition convaincante de synthèse et de dépassement des études antérieures sur les prépositions complexes. Quant à l'étude critique de Vigier & Kahng (2022) qui l'a suivie, elle a permis d'associer à chaque critère proposé par Stosic & Fagard (2019) une série de tests formels permettant de le vérifier, et a formulé une grille multicritère plus ramassée au pouvoir discriminant plus grand entre les séquences candidates à la catégorie des PrepComp.

La présente étude souhaite donc se focaliser sur le second verrou qui entrave, actuellement, la possibilité de constituer une liste des PrepComp du français, dotée de la couverture la plus large possible. Ce verrou est en réalité double : non seulement le nombre des PrepComp est élevé en français, mais le nombre de critères et de tests associés auquel on doit recourir pour décider si une séquence est ou non une bonne candidate à cette catégorie demeure élevé : entre 15 et 21 selon les auteurs. Cette difficulté liée au nombre est en outre démultipliée par le fait que leur application essentiellement manuelle nécessite un temps très significatif, ce qui, *in fine*, tend à décourager toute tentative d'utilisation de ces grilles à grande échelle.

C'est la raison pour laquelle nous avons commencé à travailler sur une approche semi-automatisée[3] d'extraction et de classement des séquences candidates à la catégorie des PrepComp, fondée sur des critères textométriques. Comme dans la plupart des méthodes d'extraction de collocations (Seretan, 2011), nous procédons en deux temps : d'abord nous extrayons des candidats en nous appuyant sur un certain patron syntaxique puis nous élaborons des indices statistiques susceptibles d'ordonner les candidats par ordre de pertinence. Voilà pourquoi nous proposerons d'abord d'extraire automatiquement deux listes de 40 candidats en nous appuyant sur les critères de la fréquence et du patron syntaxique. La première liste est assez générique tandis que la seconde se focalise sur des constructions introduites par *en*, plus souvent figées. Ces 80 séquences appartiennent au même patron reconnu comme le plus productif en français. Dans un deuxième temps, nous classons ces candidats en fonction des critères et tests manuels d'une part, et en fonction d'indicateurs textométriques d'autre part : de la sorte, nous espérons déterminer les indicateurs calculables automatiquement qui seront le mieux corrélés aux tests manuels, dans les deux cas de figure considérés.

Dans la suite de cet article, après avoir présenté (section 2) la méthode et le corpus de travail que nous avons utilisés pour extraire deux listes contenant chacune 40 séquences candidates à la catégorie des PrepComp, nous présenterons la grille multicritère (Stosic & Fagard 2019, Vigier & Kahng 2022) que nous leur avons appliquée et les classements auxquels nous avons abouti. La deuxième partie de notre propos sera consacrée à la présentation des mesures textométriques choisies et à leur application aux séquences candidates à tester. Dans la troisième et dernière partie, nous exposerons les mesures statistiques de corrélation que nous avons utilisées en vue de déterminer quelles étaient les configurations de mesures textométriques qui s'approchaient au mieux des résultats des tests manuels accomplis sur les séquences. Ces analyses seront discutées avant de conclure sur les potentialités et les limites de la méthode mise en œuvre.

2 Extraction puis classement manuel et semi-automatique de listes de 40 séquences candidates à la catégorie des PrepComp.

2.1 Méthode et corpus

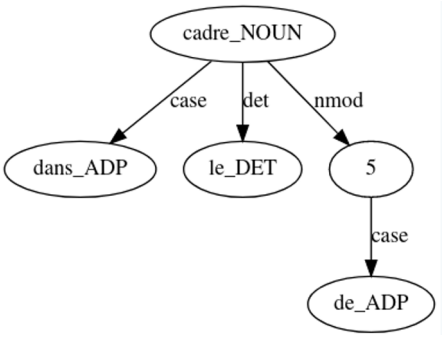
L'objectif de cette première phase de notre travail a consisté à extraire d'un corpus de travail, en nous fondant sur le seul critère de haute fréquence, deux listes de 40 séquences conformes à un patron

distributionnel spécifique. Ces séquences extraites, nous leur avons appliqué la grille multicritère de Vigier & Kahng (2022) afin de les répartir selon leur gradient de prototypicité.

Le choix du critère de haute fréquence pour l'extraction est lié au fait que, comme le font observer Stosic & Fagard (2019 : 20), les PrepComp « sont supposées avoir une fréquence d'emploi plus élevée que les structures libres relevant des mêmes patrons de construction » mais que « la fréquence d'une séquence donnée ne suffit pas pour évaluer son degré de figement ». En d'autres termes, extraire d'un corpus les séquences de mots i) conformes au patron le plus productif des PrepComp en français (Melis 2003, Stosic & Fagard 2019) et ii) dotées des fréquences les plus élevées[4], permet certes de retenir un certain nombre de PrepComp fortement figées, conformément au principe suivant lequel il existe un lien entre fréquence et figement (Zipf 1936). Mais on est aussi conduit à extraire des séquences peu ou pas figées, autrement dit de moins bonnes - voire de mauvaises - candidates, dont la haute fréquence d'utilisation s'explique pour des motifs autres que le figement (voir *infra*). Or c'est précisément cette variété de positionnement des séquences vis-à-vis du noyau prototypique des PrepComp que nous visons, afin de pouvoir disposer à priori d'un éventail assez large de candidates susceptibles d'être réparties ensuite, par l'application de tests manuels et semi-automatiques, tout au long d'une échelle de prototypicité allant de séquences dont l'agencement se révèle partiellement libre jusqu'à des séquences très figées appartenant au noyau prototypique de la catégorie.

2.1.1 Méthode

Dans le cadre de collocations binaires, comme l'ont montré Seretan (2011) puis Bartsh et Evert (2014), il est préférable de s'appuyer sur une conception syntaxique de la cooccurrence en s'adossant à des modèles de dépendance syntaxique. Ainsi, plutôt que de s'intéresser à des cooccurrents à l'intérieur d'une fenêtre de largeur constante (p.ex. 3 ou 5 mots précédant / suivant l'unité considérée), il est plus juste de s'intéresser à des unités qui entretiennent une relation syntaxique définie[5] : relation sujet entre un verbe et un nom, relation de détermination entre un nom et son déterminant, relation entre le nom tête d'un syntagme prépositionnel et sa préposition, etc. Comme nous mettons en œuvre ici des mesures d'association, nous nous situons dans ce paradigme de la « cooccurrence syntaxique » (Evert 2008).



<l=dans,c=ADP,#1>&&<l=le,c=DET,#2>&&
<l=cadre,c=NOUN,#3>&&<l=de,c=ADP,#4>:
:(case,3,1) (case,5,4) (det,3,2) (nmod,3,5)

Figure 1. Exemple de sous-arbre syntaxique correspondant à la PrepComp *dans le cadre de*. A droite, la requête correspondante, générée automatiquement.

Pour explorer des corpus analysés en dépendance, nous utilisons la plateforme du Lexicoscope 2.0 (Kraif, 2019). Celle-ci permet, pour une expression donnée, p.ex. "dans le cadre de", d'identifier dans le corpus le sous-arbre syntaxique lui correspondant, comme dans la figure 1.

Le modèle de dépendances utilisé ici est celui de Universal Dependency (de Marneffe et al., 2021)[6]. L'intérêt d'une telle démarche est qu'elle permet d'identifier d'éventuelles variations en surface de l'expression, telles que des insertions :

> (1) **Dans le cadre** innocent **d**'une partie, tu peux défier ton hôte ou ton invité. [Jaworski J.-P. (2013) Rois du monde 1 Même pas mort]

Pour cette étude, nous avons retenu un des patrons les plus productifs : *Prep (Det) Nom Prep*, la présence du déterminant étant facultative. Nous avons choisi de nous limiter aux 40 candidats les plus fréquents, dans notre corpus, correspondant à deux versions de ce patron : d'abord sans contrainte sur la première préposition (liste A), ensuite en nous focalisant sur la préposition *en*, en première position (liste B). Nous avons ainsi voulu distinguer deux niveaux de granularité au sein des occurrences du patron considéré : le premier étant le plus général, le second focalisant sur une préposition tête donnée. Le choix de « en » est lié à une thèse en cours[7]. Cette sélection de quatre-vingts séquences candidates est suffisante, nous le verrons, pour identifier des expressions que nous reconnaissons comme des PrepComp (p.ex. "à côté de"), mais aussi des intrus (p.ex. "du président de"), suffisamment nombreux pour voir si les critères étudiés lors de la deuxième étape sont discriminants (notre but étant d'identifier le pouvoir filtrant de ces critères).

2.1.2 Corpus d'étude

Pour la définition de notre corpus, trois critères sont à prendre en compte : la taille, la variété générique et la variété des domaines. En effet, afin de pouvoir nous appuyer sur des données statistiques significatives, il est important de pouvoir disposer d'un corpus suffisamment vaste. Comme le phénomène qui nous intéresse appartient à la langue générale, et non à tel ou tel discours spécialisé, la variété des genres textuels et la variété des domaines nous permettront de vérifier qu'une expression observée est suffisamment bien répartie à travers le corpus pour ne pas ressortir à un discours spécialisé. L'idéal serait de disposer d'un corpus de référence du français contemporain, mais celui-ci n'est pas encore disponible à l'heure actuelle.

Nous avons donc échantillonné notre corpus à partir de différentes sources présentes sur le Lexicoscope :

- Phraseorom_fr_fr : vaste corpus de romans contemporains réalisé dans le cadre du projet Phraseorom (Diwersy et al., 2021).

- ParaSHS_fr : corpus d'articles scientifiques en SHS téléchargés à partir du site d'Open Edition (Frérot et al., 2022)

- Reporterre_fr : articles de la revue Reporterre publiés entre 2008 et 2021 (Jackewicz et al., 2021)

- LMdiachro_fr : échantillon d'articles du Monde publiés entre 1980 à 2019.

Nous avons délibérément choisi une certaine profondeur diachronique (de 1950 à 2016 pour les romans et de 1980 à 2021 pour la presse) afin de garantir la diversité thématique. Ainsi constitué, le corpus couvre des domaines et des genres variés, et il permet d'extraire les différents niveaux de dispersion que nous avons retenus.

Le tableau ci-dessous résume la constitution du corpus :

Tableau 1. Constitution du corpus, représentant un total d'environ 370 millions de tokens

Nom du corpus	Années	Genre	Taille en tokens
Phraseorom_fr_fr	1950-2016	roman	103 809 358
Reporterre_fr	2008-2021	article de presse web	15 725 911
ParaSHS_fr	1990-2020	article scientifique	4 521 587
LMdiachro_fr	1980- 2019	article de presse quotidienne	265 500 470
Total			373 831 415

2.1.3 Choix du patron distributionnel et listes des séquences candidates extraites

Il existe un consensus (e.a. Melis (2003 : 107-108), Stosic & Fagard (2019 : 15-16)) pour considérer qu'en français le patron le plus productif est de type *Prep (Det) Nom Prep*: c'est donc celui que nous avons choisi. Pour l'extraction, nous avons d'abord cherché les occurrences d'un patron « à plat », c'est-à-dire en cherchant les suites d'étiquettes *ADP NOUN ADP* ou *ADP DET NOUN ADP* (étiquettes UD). Nous avons rapidement observé que ces suites correspondent à deux sous-arbres possibles dans le modèle UD, la relation de figement étant parfois rendue explicite *via* une relation « fixed »:

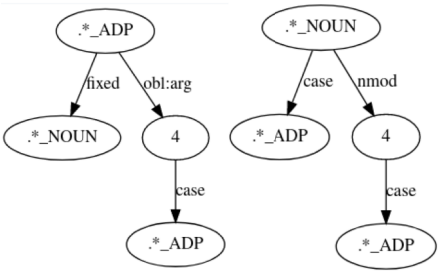


Figure 2. Sous-arbres syntaxiques correspondant aux expressions candidates selon le modèle UD[8]

Sur la base de ces deux requêtes, nous avons extrait deux listes d'expressions candidates en ne retenant à chaque fois que les 40 plus fréquentes.

Tableau 2. À gauche : liste A de 40 séquences conformes au patron distributionnel *Prep (Det) Nom Prep* retenues sur un critère de plus haute fréquence et classées par ordre alphabétique. À droite : Liste B de 40 séquences conformes au patron distributionnel *en (Det) Nom Prep*, retenues sur un critère de plus haute fréquence et classées par ordre alphabétique.

Prep (Det) Nom Prep		*en (Det) Nom Prep*	
à_ADP bord_NOUN de_ADP	de_ADP accord_NOUN avec_ADP	en_ADP accord_NOUN avec_ADP	en_ADP guise_NOUN de_ADP
à_ADP côté_NOUN de_ADP	de_ADP fait_NOUN de_ADP	en_ADP baisse_NOUN de_ADP	en_ADP hausse_NOUN de_ADP
à_ADP le_DET cours_NOUN de_ADP	de_ADP le_DET comité_NOUN de_ADP	en_ADP bas_NOUN de_ADP	en_ADP le_DET absence_NOUN de_ADP
à_ADP le_DET début_NOUN de_ADP	de_ADP le_DET conseil_NOUN de_ADP	en_ADP cas_NOUN de_ADP	en_ADP le_DET espace_NOUN de_ADP
à_ADP le_DET demande_NOUN de_ADP	de_ADP le_DET cour_NOUN de_ADP	en_ADP charge_NOUN de_ADP	en_ADP marge_NOUN de_ADP
à_ADP le_DET égard_NOUN de_ADP	de_ADP le_DET droit_NOUN de_ADP	en_ADP chef_NOUN de_ADP	en_ADP matière_NOUN de_ADP
à_ADP le_DET fin_NOUN de_ADP	de_ADP le_DET fédération_NOUN de_ADP	en_ADP collaboration_NOUN avec_ADP	en_ADP période_NOUN de_ADP
à_ADP le_DET intérieur_NOUN de_ADP	de_ADP le_DET loi_NOUN de_ADP	en_ADP compagnie_NOUN de_ADP	en_ADP position_NOUN de_ADP
à_ADP le_DET milieu_NOUN de_ADP	de_ADP le_DET ministère_NOUN de_ADP	en_ADP début_NOUN de_ADP	en_ADP présence_NOUN de_ADP
à_ADP le_DET niveau_NOUN de_ADP	de_ADP le_DET pays_NOUN de_ADP	en_ADP dehors_NOUN de_ADP	en_ADP provenance_NOUN de_ADP
à_ADP le_DET nom_NOUN de_ADP	de_ADP le_DET politique_NOUN de_ADP	en_ADP dépit_NOUN de_ADP	en_ADP quête_NOUN de_ADP
à_ADP le_DET occasion_NOUN de_ADP	de_ADP le_DET président_NOUN de_ADP	en_ADP direction_NOUN de_ADP	en_ADP raison_NOUN de_ADP
à_ADP le_DET sein_NOUN de_ADP	de_ADP le_DET prix_NOUN de_ADP	en_ADP échange_NOUN de_ADP	en_ADP remplacement_NOUN de_ADP
à_ADP le_DET suite_NOUN de_ADP	de_ADP le_DET service_NOUN de_ADP	en_ADP état_NOUN de_ADP	en_ADP réponse_NOUN à_ADP
à_ADP le_DET veille_NOUN de_ADP	de_ADP le_DET union_NOUN de_ADP	en_ADP face_NOUN de_ADP	en_ADP signe_NOUN de_ADP
à_ADP lieu_NOUN de_ADP	de_la_ADP part_NOUN de_ADP	en_ADP faveur_NOUN de_ADP	en_ADP situation_NOUN de_ADP
à_ADP proximité_NOUN de_ADP	en_ADP face_NOUN de_ADP	en_ADP fin_NOUN de_ADP	en_ADP temps_NOUN de_ADP
à_ADP sujet_NOUN de_ADP	en_ADP haut_NOUN de_ADP	en_ADP fonction_NOUN de_ADP	en_ADP terme_NOUN de_ADP
dans_ADP le_DET cadre_NOUN de_ADP	en_ADP tête_NOUN de_ADP	en_ADP forme_NOUN de_ADP	en_ADP tête_NOUN de_ADP
dans_ADP le_DET sens_NOUN de_ADP	sur_ADP le_DET marché_NOUN de_ADP	en_ADP garde_NOUN de_ADP	en_ADP voie_NOUN de_ADP

Les séquences présentes dans ces deux listes ne semblent pas toutes d'excellentes candidates à la catégorie des PrepComp. Ainsi, dans la liste A, "au bord de", "à l'égard de", "au lieu de" etc. sont à priori proches du noyau prototypique des PrepComp : de fait, on les trouve citées par les lexicographes (dans le *Dictionnaire de l'Académie* (DAc) 9è éd. par ex.). En revanche "du ministère de", "du président de", "sur le marché de" en semblent beaucoup plus éloignées.

Dans la liste B, si "en guise de", "en raison de" ou encore "en dépit de" sont a priori proches du pôle de prototypicité (on les trouve elles aussi répertoriées dans le DAc), "en chef de" ou "en garde à" en sont à priori très éloignées. Notons que dans cette liste B, nous n'avons pas retenu les très fréquentes

occurrences de “en_ADP œuvre_NOUN de_ADP”, “en_ADP place_NOUN de_ADP”, “en_ADP scène_NOUN de_ADP”, qui correspondent à un autre schéma syntaxique. Par exemple, pour “en_ADP scène_NOUN de_ADP”, on a le sous-arbre suivant (qui correspond le plus souvent à “mise en scène de...”). De même “en_ADP train_NOUN de_ADP” a été automatiquement écarté par la syntaxe, car le nom y est relié à une infinitive via une relation *xcomp* :

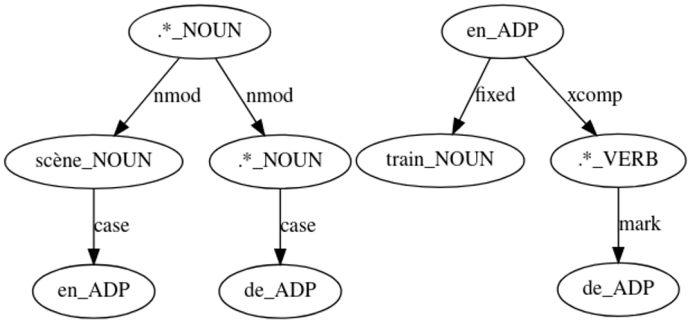


Figure 3. Les sous-arbres syntaxiques de “en scène de” et “en train de”

Dans la liste B, le recours à la syntaxe en dépendances permet donc de filtrer une grande partie du bruit, ce qui n’est pas le cas pour la liste A. En effet, pour cette liste, les mauvais candidats (voir scores présentés *infra*) tels que “du président de” ou “du ministère de” présentent une structure syntaxique non figée (de type [P][SN][de SN] : notation reprise de Borillo 1997 : 178) mais conforme au patron susceptible de conduire, par figement, aux prépositions complexes. Illustrons notre propos par un exemple : pour le patron choisi, la séquence « au lieu de » dont on verra qu’elle obtient le meilleur score des candidates à la catégorie des PrepComp dans la liste A, est issue d’un agencement syntaxique non figé (à l’image de « du président de ») qu’on peut encore trouver en français dans la séquence suivante par exemple:

> (2) Les graines flottantes s'ensevelissent **au lieu de** leur atterrissage. Il en naîtra des arbres pour 'ébénisterie. [Saint John Perse, *Amers*, 1957 - Frantext]

Si, dans (2), « **au lieu de** leur atterrissage » s’analyse comme suit : [P][SN][de SN], « **au lieu de** la soutane » dans (3) s’analysera comme suit: [P SN de] [SN]

> (3) (...) en dépit du port d'un vêtement laïque **au lieu de** la soutane (...) [H. Bianciotti, *Le Pas si lent de l’Amour,* 1995 - Frantext]

En revanche, pour la liste B, les « mauvais candidats » ne présentent souvent pas le même profil syntaxique, ce qui explique qu’ils aient été écartés par le filtrage de la syntaxe en dépendances. Pour reprendre les deux exemples de “en scène de” et “en train de” : la première séquence doit sa très haute fréquence à l’usage du nom polylexical “mise en scène” lui-même complété par un modifieur de type SP, de sorte que l’on a affaire à la structure syntaxique [N P N][de][SN]. Il ne peut donc s’agir d’une matrice possible pour une PrepComp. Quant à “en train de”, il s’agit d’une séquence dont la haute fréquence s’explique par son inclusion dans la structure plus vaste très figée “être en train de” qui entre dans les périphrases verbales de type “être en train de VInf.” D’où son filtrage par l’analyseur syntaxique utilisé.

Tous ces éléments doivent être maintenant validés ou invalidés par l’application de tests à même de vérifier si l’ensemble de ces séquences vérifient ou non les traits propres aux exemplaires prototypiques de la classe.

2.2 Présentation et application de la grille multicritère proposée par Vigier & Kahng (2022)

Faisant suite aux analyses formulées par Stosic & Fagard (2019), Vigier & Kahng (2022) ont proposé une grille alternative avec un nombre de tests plus réduit tout en permettant d’accroître son pouvoir de

discrimination entre les séquences candidates. Ce pouvoir discriminant renforcé tient essentiellement à trois éléments : la prime donnée aux critères syntaxiques sur les critères morphologiques ou sémantiques proposés initialement par Stosic & Fagard, le poids relatif accru donné aux critères fréquentiel (15% du poids total de la note contre 9% pour les patrons à noyau nominal) et à la mesure d'attraction réciproque (20% contre 3%), mesure elle-même adossée à une méthode d'analyse plus fine.

On rappellera de manière très cursive (pour plus de détail, nous renvoyons aux deux articles cités *supra*) les propriétés que teste la grille que nous avons appliquée aux séquences des listes A et B obéissant toutes au patron syntagmatique *Prep (Det) Nom Prep*.

Tests sur 11 propriétés syntaxiques : [P1] présence/absence d'un article devant le nom noyau ; [P2] possibilité/impossibilité d'anaphorisation du complément de la préposition par un déterminant possessif / [P3] par un déterminant démonstratif ; [P4] possibilité d'interroger le complément de la préposition à l'aide d'un déterminant interrogatif ; [P5] autres possibilités de variation du déterminant du nom noyau ; [P6] caractère référentiel/non-référentiel du SN ; [P7] possibilité /impossibilité d'insertion d'un modifieur adjectival ; [P8] possibilité /impossibilité de variation du modifieur adjectival; [P9] possibilité /impossibilité d'insertion d'un adverbe, d'un adverbial ou d'une incise à l'intérieur de la séquence candidate ; [P10] variation du modifieur adverbial ; [P11] possibilité /impossibilité de coordination avec une préposition simple ou une PrepComp prototypique ;

Test sur 2 propriétés sémantiques : [P12] extension/ absence d'extension sémantique de la séquence ; [P13] opacité / transparence sémantique de la séquence

Test sur 2 propriétés d'ordre quantitatif : [P14] fréquence relative de la séquence par millions de mots calculée dans FrTenTen (SketchEngine) ; [P15] score d'information réciproque calculé dans FrTenTen (SketchEngine).

Voici les scores obtenus par les 80 expressions. Rappelons que dans le calcul proposé, plus la séquence est figée, plus son score est faible.

Tableau 3. Récapitulatif des scores numériques (calculés en %) obtenus par les séquences de la liste A après application des tests de la grille multicritère proposée par Vigier & Kahng (2022)

Rang	Candidat	score synthétique	Rang	Candidat	score synthétique
1	à_ADP lieu_NOUN de_ADP	16	11	à_ADP le_DET niveau_NOUN de_ADP	55
2	à_ADP côté_NOUN de_ADP	21	11	à_ADP le_DET occasion_NOUN de_ADP	55
2	en_ADP tête_NOUN de_ADP	21	12	à_ADP le_DET fin_NOUN de_ADP	58
3	de_ADP accord_NOUN avec_ADP	24	12	à_ADP le_DET nom_NOUN de_ADP	58
4	à_ADP le_DET cours_NOUN de_ADP	26	13	dans_ADP le_DET sens_NOUN de_ADP	66
4	en_ADP face_NOUN de_ADP	26	14	à_ADP le_DET demande_NOUN de_ADP	68
4	en_ADP haut_NOUN de_ADP	26	15	de_ADP le_DET conseil_NOUN de_ADP	74
5	à_ADP le_DET veille_NOUN de_ADP	32	15	de_ADP le_DET comité_NOUN de_ADP	74
6	à_ADP le_DET égard_NOUN de_ADP	37	15	de_ADP le_DET ministère_NOUN de_ADP	74
6	à_ADP proximité_NOUN de_ADP	37	15	de_ADP le_DET service_NOUN de_ADP	74
7	à_ADP bord_NOUN de_ADP	39	15	de_ADP le_DET droit_NOUN de_ADP	74
7	à_ADP le_DET sein_NOUN de_ADP	39	16	de_ADP le_DET prix_NOUN de_ADP	79
7	à_ADP le_DET milieu_NOUN de_ADP	39	16	de_ADP le_DET politique_NOUN de_ADP	79
7	à_ADP le_DET suite_NOUN de_ADP	39	16	de_ADP le_DET président_NOUN de_ADP	79
7	dans_ADP le_DET cadre_NOUN de_ADP	39	16	de_ADP le_DET loi_NOUN de_ADP	79
8	de_la_ADP part_NOUN de_ADP	42	16	sur_ADP le_DET marché_NOUN de_ADP	79
8	à_ADP le_DET intérieur_NOUN de_ADP	42	17	de_ADP le_DET cour_NOUN de_ADP	89
9	de_ADP fait_NOUN de_ADP	47	18	de_ADP le_DET fédération_NOUN de_ADP	95
10	à_ADP sujet_NOUN de_ADP	53	18	de_ADP le_DET pays_NOUN de_ADP	95
10	à_ADP le_DET début_NOUN de_ADP	53	18	de_ADP le_DET union_NOUN de_ADP	95

Tableau 4. Récapitulatif des scores numériques (calculés en %) obtenus par les séquences de la liste B après application des tests de la grille multicritère proposée par Vigier & Kahng (2022)

Rang	Candidat	Score synthétique	Rang	Candidat	Score synthétique
1	en_ADP tête_NOUN de_ADP	16	21	en_ADP matière_NOUN de_ADP	39
2	en_ADP fonction_NOUN de_ADP	16	22	en_ADP direction_NOUN de_ADP	39
3	en_ADP dépit_NOUN de_ADP	17	23	en_ADP face_NOUN de_ADP	39
4	en_ADP dehors_NOUN de_ADP	17	24	en_ADP situation_NOUN de_ADP	39
5	en_ADP faveur_NOUN de_ADP	22	25	en_ADP quête_NOUN de_ADP	42
6	en_ADP raison_NOUN de_ADP	24	26	en_ADP réponse_NOUN à_ADP	42
7	en_ADP fin_NOUN de_ADP	25	27	en_ADP accord_NOUN avec_ADP	42
8	en_ADP charge_NOUN de_ADP	26	28	en_ADP cas_NOUN de_ADP	45
9	en_ADP terme_NOUN de_ADP	26	29	en_ADP remplacement_NOUN de_ADP	47
10	en_ADP guise_NOUN de_ADP	28	30	en_ADP temps_NOUN de_ADP	50
11	en_ADP marge_NOUN de_ADP	28	31	en_ADP compagnie_NOUN de_ADP	52
12	en_ADP collaboration_NOUN avec_ADP	32	32	en_ADP hausse_NOUN de_ADP	53
13	en_ADP bas_NOUN de_ADP	33	33	en_ADP signe_NOUN de_ADP	55
14	en_ADP présence_NOUN de_ADP	34	34	en_ADP état_NOUN de_ADP	58
15	en_ADP provenance_NOUN de_ADP	34	35	en_ADP le_DET espace_NOUN de_ADP	58
16	en_ADP début_NOUN de_ADP	37	36	en_ADP baisse_NOUN de_ADP	58
17	en_ADP forme_NOUN de_ADP	37	37	en_ADP position_NOUN de_ADP	60
18	en_ADP échange_NOUN de_ADP	37	38	en_ADP période_NOUN de_ADP	68
19	en_ADP le_DET absence_NOUN de_ADP	37	39	en_ADP chef_NOUN de_ADP	100
20	en_ADP voie_NOUN de_ADP	37	40	en_ADP garde_NOUN à_ADP	100

Pour les séquences de la liste A, on observe la présence d'excellentes candidates telles que "au lieu de", "à côté de", "en tête de" du côté gauche, et de très mauvaises telles que "de la fédération de", "du pays de", "de l'union de" à droite. Entre ces deux extrêmes se répartissent les 34 autres séquences.

Le second tableau montre une situation différente : la distribution des scores des 38 premières séquences est beaucoup plus resserrée que pour la liste A (16% -> 68% *versus* 16% -> 95%). Deux expressions obtiennent un score maximal : "en chef de" et "en garde à", ce qui confirme leur non-appartenance à la classe. Une vérification manuelle de leurs occurrences révèle deux cas de figure un peu différents : "en chef de" correspond à des expressions comme "rédacteur en chef de + Nom Propre", "économiste en chef de + Nom Propre", etc. Or on constate que l'analyseur a systématiquement mal analysé ces occurrences en rattachant le deuxième syntagme prépositionnel ("de + Nom Propre") à "chef". Si l'analyse avait été correcte, cet intrus aurait été filtré par la syntaxe, tout comme pour les expressions telles que "mise en œuvre de". Pour "en garde à", la situation est différente : il s'agit simplement d'un fragment de l'expression "en garde à vue". C'est cet aspect fragmentaire et non autonome (au plan sémantique) de la séquence qui est responsable de la non-applicabilité des tests manuels.

3 Vers un classement entièrement automatisé des séquences candidates à la catégorie des PrepComp

Comme dit dans notre introduction, notre objectif consiste à proposer de nouvelles pistes en vue d'automatiser entièrement le classement d'un nombre très significatif de séquences extraites sur des critères distributionnels et susceptibles d'appartenir à la catégorie des prépositions complexes.

Pour ce faire nous proposons la méthode suivante :

- Définir a priori un ensemble de mesures textométriques susceptibles d'être pertinentes pour la détection des séquences fortement figées.
- Appliquer ces mesures aux deux listes A et B présentées *supra*.
- Mesurer le degré de corrélation entre le score synthétique calculé a moyen de la grille multicritère présentée sous § 2.2. et ces critères, pris indépendamment ou combinés.

Nous retenons les critères textométriques suivants :

- la fréquence : le nombre d'occurrences du candidat dans le corpus.

- La dispersion : dans leur analyse sur l'extraction de mots clés spécifiques à un document, Egbert & Biber (2019) ont montré qu'il était très utile de combiner la mesure de spécificité à une mesure de dispersion, afin d'identifier des mots « jouant le rôle de mot-clé dans une large proportion des textes du corpus » (notre traduction, Egbert and Biber 2019 : 92). Dans leur expérience, ils s'appuient sur une mesure simple de la dispersion, à savoir la fréquence de document (le nombre de documents dans lequel un mot apparaît). Comme le note Gries (2021: 4), « les travaux récents en linguistique de corpus ont commencé à souligner l'importance jouée par la dispersion pour ce type d'analyse et pour d'autres types. » Ici, nous ne cherchons pas à caractériser des mots spécifiques à tel ou tel corpus, mais au contraire des unités générales, quasi grammaticales, et de fait, en tant que telles, susceptibles d'être uniformément réparties dans l'ensemble de notre corpus. En fonction des métadonnées dont nous disposons pour notre corpus, nous testerons différentes dimensions de la dispersion :

- disp1000 : le nombre de blocs de 1000 phrases successives où l'expression apparaît ;

- dispDoc : la fréquence de document ;

- dispAuth : le nombre d'auteurs différents utilisant l'expression ;

- dispYear : le nombre d'années différentes où l'expression est attestée dans le corpus ;

- dispGenre: le nombre de genre de textes différents où l'expression apparaît ;

- dispFile : le nombre de fichiers où l'expression apparaît.

Cette dernière mesure est composite, car le découpage en fichiers intègre le découpage par genre, et pour chaque genre, correspond à un découpage différent : par décades pour les articles du *Monde*, par année pour les articles de *Reporterre*, par discipline pour les articles scientifiques, par auteur pour les romans.

- Le taux d'insertion : nombre moyen de mots s'insérant à l'intérieur de la séquence. Ce type d'indice a déjà été combiné à des mesures d'association par Daille (1994). Par exemple, pour "au milieu de", nous obtenons un taux d'insertion de 0,5 du fait des occurrences fréquentes de "au beau milieu de" ou "au milieu même de". Nous présupposons que les vraies PrepComp sont plus figées syntaxiquement que les autres réalisations du patron et qu'elles acceptent donc moins fréquemment des insertions.

- La mesure d'association interne : les mesures d'associations statistiques (information mutuelle spécifique, log rapport de vraisemblance, t-score, z-score, Evert 2008, Seretan 2011) sont couramment employées pour identifier le degré d'attraction mutuelle de deux formes dans une collocation. Les mêmes mesures peuvent s'employer au-delà de la collocation binaire, p.ex. quand on a des collocations de collocation (Seretan 2011) ou des Arbres lexicosyntaxiques récurrents (Kraif, Tutin, Diwersy, 2014). Bien que le Lexicoscope puisse calculer, pour un arbre donné, la moyenne des mesures d'association entre chaque feuille et le reste de l'arbre, nous avons opté pour un calcul plus simple, car le modèle de dépendance utilisé, UD, ne se prête pas à ce calcul. En effet, si l'on considère les deux sous-arbres de la figure 2, on constate que les prépositions initiales sont tantôt analysées comme feuilles, et tantôt comme têtes (dans une relation de figement, on prend arbitrairement la première forme comme tête), ce qui rendrait difficilement comparable le calcul des mesures d'association au niveau des feuilles.

Nous avons donc calculé la mesure d'association interne entre la première préposition et le reste de l'expression, et entre le dernière préposition et ce qui précède : assoc("à côté de") = *moyenne(assoc("à",*

"côté de"), assoc("à côté", "de")). Pour cette mesure d'association, nous avons utilisé 3 calculs différents : le t-score, l'information mutuelle spécifique (notée PMI) et le log rapport de vraisemblance (noté LLR).

Une fois obtenus ces différents indices, nous pouvons mesurer le degré de corrélation entre le score synthétique et ces critères, pris indépendamment ou combinés.

4 Résultats et discussion

4.1 Liste A

Nous avons d'abord concentré nos efforts sur la liste A, qui présente l'intérêt de comporter de nombreuses expressions apparaissant intuitivement comme des intrus. Comme le montre le tableau ci-dessous, on peut, en fonction des scores attribués manuellement, découper la liste en trois sous-groupes (avec des seuils à 40 et à 70).

Tableau 5. Scores associés aux 40 séquences candidates (liste A), réparties en 3 groupes (vert, orange, rouge).

Rang	Candidat	score synthétique	Rang	Candidat	score synthétique
1	à_ADP lieu_NOUN de_ADP	16	11	à_ADP le_DET niveau_NOUN de_ADP	55
2	à_ADP côté_NOUN de_ADP	21	11	à_ADP le_DET occasion_NOUN de_ADP	55
2	en_ADP tête_NOUN de_ADP	21	12	à_ADP le_DET fin_NOUN de_ADP	58
3	de_ADP accord_NOUN avec_ADP	24	12	à_ADP le_DET nom_NOUN de_ADP	58
4	à_ADP le_DET cours_NOUN de_ADP	26	13	dans_ADP le_DET sens_NOUN de_ADP	66
4	en_ADP face_NOUN de_ADP	26	14	à_ADP le_DET demande_NOUN de_ADP	68
4	en_ADP haut_NOUN de_ADP	26	15	de_ADP le_DET conseil_NOUN de_ADP	74
5	à_ADP le_DET veille_NOUN de_ADP	32	15	de_ADP le_DET comité_NOUN de_ADP	74
6	à_ADP le_DET égard_NOUN de_ADP	37	15	de_ADP le_DET ministère_NOUN de_ADP	74
6	à_ADP proximité_NOUN de_ADP	37	15	de_ADP le_DET service_NOUN de_ADP	74
7	à_ADP bord_NOUN de_ADP	39	15	de_ADP le_DET droit_NOUN de_ADP	74
7	à_ADP le_DET sein_NOUN de_ADP	39	16	de_ADP le_DET prix_NOUN de_ADP	79
7	à_ADP le_DET milieu_NOUN de_ADP	39	16	de_ADP le_DET politique_NOUN de_ADP	79
7	à_ADP le_DET suite_NOUN de_ADP	39	16	de_ADP le_DET président_NOUN de_ADP	79
7	dans_ADP le_DET cadre_NOUN de_ADP	39	16	de_ADP le_DET loi_NOUN de_ADP	79
8	de_la_ADP part_NOUN de_ADP	42	16	sur_ADP le_DET marché_NOUN de_ADP	79
8	à_ADP le_DET intérieur_NOUN de_ADP	42	17	de_ADP le_DET cour_NOUN de_ADP	89
9	de_ADP fait_NOUN de_ADP	47	18	de_ADP le_DET fédération_NOUN de_ADP	95
10	à_ADP sujet_NOUN de_ADP	53	18	de_ADP le_DET pays_NOUN de_ADP	95
10	à_ADP le_DET début_NOUN de_ADP	53	18	de_ADP le_DET union_NOUN de_ADP	95

On voit que le groupe vert correspond à des prépositions complexes canoniques, alors que le groupe rouge correspond à des mauvais candidats, tandis que le groupe orange constitue un entre-deux intéressant (le phénomène étant scalaire il nous semble plus pertinent de dégager une zone « grise » intermédiaire que d'opposer deux groupes avec une frontière nette).

Si l'on se tourne maintenant vers les mesures de corrélation calculées entre les scores synthétiques obtenus par l'application (essentiellement manuelle) de la grille multicritère d'une part, les critères textométriques pris individuellement ou diversement combinés d'autre part, on observe que les meilleurs critères sont : le taux d'insertion, la dispersion par année, la dispersion par fichier et le t-score. Le LLR obtient une corrélation linéaire modeste de -0.31 ($p < 0,05$).

Tableau 6. Meilleurs indices, ordonnés par corrélation décroissante des rangs (les seuils de significativité statistiques sont exprimés entre parenthèses).

Indice	Corrélation des rangs (coeff. de Spearman)	Corrélation linéaire (coeff. de Pearson)
insertion	+ 0,724 ($p < 0,001$)	+ 0,556 ($p < 0,001$)
log(insertion)	+ 0,724 ($p < 0,001$)	+ 0,737 ($p < 0,001$)
dispYear	- 0,578 ($p < 0,001$)	- 0,511 ($p < 0,001$)
dispFile	- 0,549 ($p. < 0,001$)	- 0,565 ($p < 0,001$)
t-score	- 0,458 ($p. < 0,01$)	- 0,497 ($p < 0,01$)

L'indice le plus fiable pour ce premier jeu de données est le taux moyen d'insertion, qui obtient de très bonnes corrélations en rang, ainsi qu'une corrélation linéaire forte en échelle logarithmique. La dispersion apparaît également comme un indice intéressant, qu'elle soit prise au niveau temporel (*dispYear*) ou au niveau composite des fichiers (*dispFile*) qui mêle plusieurs types de dispersion (genre, année, auteurs). Enfin, au plan des mesures d'association, seul le *t-score* présente une assez bonne corrélation, ainsi que le LLR dans une moindre mesure. On retrouve, par le jeu des indices automatiques, les critères déjà présents dans les tests manuels : le figement plus ou moins fort (taux d'insertion faible) et l'association statistiques entre les éléments qui la composent. A ceux-ci s'ajoute le critère de dispersion : en tant qu'expression en voie de grammaticalisation, les PrepComp apparaissent comme non spécialisées, et donc plutôt uniformément réparties dans le corpus.

Nous avons tenté de combiner linéairement ces différents indices, pour identifier une éventuelle complémentarité, mais nous n'avons obtenu que des augmentations trop faibles pour être significatives.

4.2 Liste B

Concernant cette deuxième liste, il apparaît beaucoup plus difficile de déterminer des seuils pertinents de score synthétique pour distinguer les PrepComp canoniques des mauvaises candidates : les intrus les plus évidents ("en train de", "en œuvre de", "en scène de", "en garde de", etc.) ont en effet déjà été éliminés par le filtre des requêtes syntaxiques. Seules "en chef de" et "en garde à" obtiennent des scores maximaux du fait de l'inapplicabilité de certains tests manuels, qui les disqualifient d'emblée.

Cette difficulté à discriminer les bons candidats des mauvais au sein de la liste obtenue se retrouve au niveau des corrélations statistiques. Cette fois on obtient des corrélations plutôt faibles avec les scores synthétiques : seule la dispersion par année obtient une corrélation linéaire de -0,476 ($p < 0,005$), la dispersion par fichier de -0,387 ($p < 0,05$), et le LLR une corrélation des rangs de -0,356 ($p < 0,05$). Le taux d'insertion ne montre pas de corrélation significative (0,265 pour les deux), ce qui reflète le fait que toutes les expressions retenues dans la liste A sont également figées, ce qui corrobore notre hypothèse sur le fait que les expressions avec "en" sont plus souvent figées. On pourra noter que les deux intrus les plus évidents, "en chef de" et "en garde à", sont en queue de liste pour les mesures de dispersion par année (resp. aux rangs 35 et 38) et par fichier (resp. aux rangs 35 et 37), ce qui confirme le caractère opératoire de la dispersion pour identifier les moins bons candidats.

5 Conclusion

Nous avons cherché, dans cette étude expérimentale, à croiser deux approches pour caractériser les prépositions complexes : d'une part, l'application de la grille multicritère proposée par Vigier & Kahng (2022) qui nécessite de combiner des tests manuels avec des mesures statistiques extraites de corpus ; d'autre part l'extraction complètement automatisée d'une série d'indices textométriques, dont certains sont originaux, comme le taux d'insertion ou une mesure de dispersion composite. Le score synthétique

obtenu de façon manuelle et semi-automatique jouant le rôle d'étalon de référence, notre objectif était d'identifier quels indices semblent être les plus discriminants pour départager les mauvaises expressions candidates des bonnes prépositions complexes.

Nos observations ont montré que sur une liste de candidats comportant de nombreux intrus, quelques indices pouvaient se révéler très discriminants : tout d'abord, le taux d'insertion, qui indique le nombre moyen de mots pouvant s'intercaler à l'intérieur des réalisations des expressions, est apparu très fortement corrélé avec le score synthétique (avec des coefficients de Pearson et de Spearman aux alentours de +0,73). Ensuite, ce sont des mesures de dispersion qui se sont révélées utiles, montrant que les prépositions complexes sont caractérisées par une certaine homogénéité de leurs répartitions au sein d'un corpus hétérogène. Parmi les mesures d'association statistiques, enfin, c'est le t-score qui s'est révélé plus discriminant, suivi par le log rapport de vraisemblance. Alors que nous pouvions supposer que ces mesures d'association seraient parmi les indices les plus fiables pour identifier le caractère plus ou moins figé des prépositions complexes, il est apparu qu'elles jouaient un rôle secondaire, les intrus pouvant également obtenir de très bons scores d'association du fait de leur grande récurrence dans certaines parties du corpus.

Une deuxième série d'observations sur des candidats utilisant la préposition "en" ont montré que le recours à des requêtes s'appuyant sur la syntaxe en dépendances permettait d'évacuer l'essentiel du bruit, par rapport à une approche basée sur l'extraction de simples patrons "à plat". Sur cette deuxième liste, les indices statistiques sont apparus beaucoup moins discriminants, sans doute parce qu'ils avaient moins de bruit à filtrer. La dispersion est néanmoins apparue comme relativement corrélée au score synthétique. En vue d'une extraction complètement automatisée des prépositions complexes, il parait donc judicieux de combiner une présélection s'appuyant sur des requêtes syntaxiques et l'application de filtres fondés sur le calcul du taux d'insertion et de la dispersion. Concernant la dispersion, il faut noter le caractère assez grossier de son mode de calcul : par exemple, pour la dispersion par genre, les valeurs s'échelonnent entre 1 et 4, et presque toutes les expressions obtiennent 4, car le calcul ne permet pas de tenir compte des fréquences respectives dans chaque sous-genre, et d'éventuels déséquilibres dans la répartition. Par ailleurs, les différents sous-corpus n'ayant pas la même structure, le calcul ici présenté peut impliquer des biais : par exemple, le corpus littéraire s'échelonne sur une période de presque 70 ans tandis que le corpus journalistique du *Monde* sur seulement 40 ans. De ce fait, des expressions plus spécifiques à la presse pourraient avoir une dispersion par année plus faible que celles qui sont employées dans les romans : ce qui fait qu'*in fine*, la dispersion par année ne serait qu'un reflet de la dispersion par sous-genre qu'un calcul trop grossier n'a pas permis de refléter correctement. Dans des travaux ultérieurs, nous prévoyons de raffiner le calcul de dispersion, par exemple en utilisant la divergence de Kullback Leibler proposée par Gries (2021) afin de mieux caractériser les dimensions pertinentes de celle-ci.

Enfin, il nous parait important de signaler une des limites des méthodes mises en œuvre : aucun des indices textométriques étudiés n'est conçu pour tenir compte du caractère fragmentaire d'une expression, comme dans le cas de "en garde à". Or l'expression complète "en garde à vue" est elle-même très figée, et obtient naturellement des scores élevés en termes de mesure d'association et très faibles en termes de taux d'insertion. Un test statistique simple permettrait d'identifier ce cas de figure, par exemple une mesure d'association orientée, qui permettrait d'identifier que la probabilité conditionnelle de trouver le cooccurrent "vue" est très élevée. Nous prévoyons d'intégrer cet indicateur dans de futures expérimentations.

Références bibliographiques

Adler, S. (2001). Les locutions prépositives : questions de méthodologie et de définition. *Travaux de linguistique*, 42-43,1, p. 157-170.

Bartsch, S., Evert, S. (2014). *Towards a Firthian notion of collocation*. In A. Abel and L. Lemnitzer (eds.), V*ernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in*

Internetwörterbüchern, number 2/2014 in OPAL - Online publizierte Arbeiten zur Linguistik, p. 48-61, Institut für Deutsche Sprache, Mannheim.

Berthonneau, A.-M. & Cadiot, P. (éds.) (1991). *Préposition, représentation, référence, Langue Française*, 91.

Borillo, A. (1991). Le lexique de l'espace : prépositions et locutions prépositionnelles de lieu en français. In Tasmowski, L. & A. Zribi-Hertz (éds.), *Hommage à N. Ruwet*. *Communication & Cognition*, Gand, p. 176-190.

Borillo, A. (1997). Aide à l'identification des prépositions composées de temps et de lieu. *Faits de langue*, 9, p. 173-184.

Borillo, A. (1998). *L'espace et son expression en français*, Paris : Ophrys.

Borillo, A. (2001). Il y a prépositions et prépositions, *Travaux de linguistique*, 42-43 (1-2), p. 141-155.

Cadiot, P. (1997). *Les prépositions abstraites en français*, Paris : Armand Colin.

De Marneffe M.-C., Manning, C., Nivre, J.,Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, ISSN 1530-9312, vol. 47, 2, p. 255-308.

De Mulder, W. & Stosic, D. (2009) (éds). *Approches récentes de la préposition*, *Langages,* 173.

Egbert, J. Biber,D. (2019). Incorporating text dispersion into keyword analyses. *Corpora* 14,1, p. 77-104.

Daille, B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, Nonantum Inn, Kennebunkport, Maine : Association for Computational Linguistics, p. 29-36.

Diwersy, S., Gonon, L., Goossens, V., Kraif, O., Novakova, I., Sorba, J & Vidotto, I. (2021). La phraséologie du roman contemporain dans les corpus et les applications de la PhraseoBase. *Corpus*, 22, p. 1-22.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, Berlin : Mouton de Gruyter, p. 1212-1248.

Frérot, C. Hartwell, L. M., Kraif, O. (2024). Corpus Approaches to Parallel Concordancing. In M. Mahlberg & G. Brookes (éds.), *Bloomsbury Handbook of Corpus Linguistics*, London: Bloomsbury.

Gries, S. Th. (2021). A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics* 9, 2, p. 1-33.

Gross, G. (1981). Les prépositions composées. In C. Schwarze (éd.), *Analyse des prépositions, 3e colloque franco-allemand de linguistique théorique*, p. 29-39.

Huddleston, R. & G. K. Pullum. (2002). *The Cambridge Grammar of the English Language*. Cambridge : Cambridge University Press. https://doi.org/10.1017/9781316423530

Jackiewicz, A., Kraif, O., Ding, Y. (2021). Réalités émergentes, possibles, désirables : comment les nommons-nous ? *Journées SEH Humanités environnementales*, Montpellier 5-7 sept. 2021

Kleiber, G. (1990). *La sémantique du prototype*. *Catégories et sens lexical*. Paris : PUF.

Kraif, O., Tutin A., Diwersy S. (2014). Extraction de pivots complexes pour l'exploration de la combinatoire du lexique : une étude dans le champ des noms d'affect, *Actes du Congrès Mondial de Linguistique Française 2014*, 19-23 juillet 2014, Berlin, p. 2663-2674.

Kraif, O. (2019). Explorer la combinatoire lexico-syntaxique des mots et expressions avec le Lexicoscope. In Max Silberstein (éd.), *Langue française* 203, p. 67-82.

Leeman, D. (2006, 2007) (éd.). *La préposition en français,* I, II, III, *Modèles linguistiques* 53, 54, 55, tomes XXVII-I & II et XXVIII-I.

Leeman, D. (2008) (éd.). *Énigmatiques prépositions*. *Langue Française*, 157.

Le Pesant, D. (2006). Classification à partir des propriétés syntaxiques. *Modèles linguistiques*, 53, p. 51-74.

Melis, L. (2003). *La préposition en français*. Paris : Ophrys.

Seretan,V. (2011). *Syntax-Based Collocation Extraction. Text, Speech and Language Technology*, Springer.

Stosic, D. & Fagard, B. (2019). Les prépositions complexes en français. Pour une méthode d'identification multicritère. *Revue romane*, 54, 1, p. 8-38.

Vigier, D. (2013)(éd.). *La préposition en*. *Langue Française*, 178.

Vigier, D., Kahng, G. (2022). Catégoriser les prépositions complexes en français. 8ème *Congrès Mondial de Linguistique Française*, Orléans, France.

Zipf, G. (1936). *The Psychobiology of Language*. London : Routledge.

[1] Voir par exemple le colloque international PrepComp2021 (https://blogs.univ-tlse2.fr/prepcomp2021/tag/prepcomp/)

[2] « Contre toute attente et en dépit de nombreux travaux sur la question, le premier défi à relever pour un linguiste travaillant sur les prépositions complexes est d'en dresser la liste. » (2019 : 8)

[3] Notre objectif étant de mettre au point une approche entièrement automatisée.

[4] De ce fait nous nous focalisons ici surtout sur des PrepComp fréquentes. Pour les expressions figées plus rares (par ex. « à l'insu de », « à l'instar de » …), d'autres méthodes devront être envisagées. On peut penser néanmoins que certaines de ces séquences à fréquence faible, très figées et apparues dans des états de langue plus anciens, sont bien connues et répertoriées par les lexicographes. Celles-ci ne posent donc pas de difficultés d'identification.

[5] Comme on le verra plus loin (§ 2.1.3.), ce filtrage syntaxique permet d'écarter des séquences comme « en_ADP œuvre_NOUN de_ADP », « en_ADP place_NOUN de_ADP », « en_ADP scène_NOUN de_ADP », qui, très fréquentes, correspondent néanmoins à un autre schéma syntaxique que celui recherché.

[6] Les analyses ont été obtenues avec Stanza (https://stanfordnlp.github.io/stanza/) avec le modèle entrainé sur le corpus GSD.

[7] G. Kahng, « Les prépositions complexes construites avec la préposition tête *en* », Université Lumière Lyon 2.

[8] ADP et NOUN désignent respectivement des prépositions et des noms. La relation *nmod* désigne un complément du nom, et *case* la relation rattachant une préposition à la tête de son syntagme. La relation *obl:arg* désigne un complément oblique, argument d'un prédicat. Quant à *fixed*, cette relation est liée au fait que dans la *treebank* de référence, les annotateurs ont parfois identifié le figement comme étant suffisamment fort pour être marqué explicitement de cette manière : mais pour ne pas biaiser nos analyses, il nous fallait considérer cette relation comme une relation quelconque, correspondant à un des patrons possibles.