Des motifs séquentiels aux motifs hiérarchiques : l'apport des arbres lexico-syntaxiques récurrents pour le repérage des routines discursives

Olivier Kraif, Agnès Tutin

Des motifs séquentiels aux motifs hiérarchiques : l'apport des arbres lexico-syntaxiques récurrents pour le repérage des routines discursives

From sequential to hierarchical motifs : what can bring Recurrent Lexico-syntactic Trees to the identification of discursive routines.

Olivier Kraif, Agnès Tutin
LIDILEM – Univ. Grenoble Alpes

Résumé : Cet article propose une réflexion à la fois théorique et méthodologique sur les objets de la phraséologie étendue, qui s'intéresse à des unités préfabriquées du discours au-delà des critères de figement. Plus précisément, nous tentons de clarifier le concept général de motif, ainsi que celui, plus spécifique, de routine discursive. Nous proposons ensuite de comparer deux approches méthodologiques différentes pour l'identification des routines en corpus : une méthode hiérarchique, basé sur le repérage d'arbres lexico-syntaxiques récurrents (ALR), et la méthode séquentielle classique des segments répétés (SR) ou n-grams. Nous montrons, au travers d'une étude sur corpus, que la méthode des ALR présente un réel intérêt pour le repérage des routines et des collocations, mais que les SR semblent plus adaptés et plus simples à mettre en œuvre pour des locutions figées ou des constructions syntaxiques impliquant des lexèmes grammaticaux – le modèle syntaxique des ALR nécessitant une adaptation pour pouvoir identifier ces cas.

Mots-clés : phraséologie étendue, motif, routine discursive, arbres lexico-syntaxiques récurrents, segments répétés

Abstract : This article proposes a theoretical and methodological reflection in the field of extended phraseology, which focuses on prefabricated units of discourse. More precisely, we try to clarify the concepts of *motif* and discursive *routine*. We propose to compare two different methodological approaches for the identification of routines in corpora: a hierarchical method based on the identification of Recurrent Lexico-syntactic Trees (RLT) and the classical sequential n-gram method. We show, through a corpus study, that the ALR method has a real interest in spotting routines and collocations, but that the n-grams seem more adapted and easier to implement for frozen locutions or syntactic constructions. The underlying syntactic model of ALR would require some adaptation to be able to identify these latter cases.

Keywords: extended phraseology, motif, discursive routine, recurrent lexico-syntactic trees, n-grams

Introduction

L'extension du domaine de la phraséologie, de la lexicologie au discours (*cf*. Legallois & Tutin, 2013), conduit à s'intéresser désormais à des segments préfabriqués de la langue, particulièrement remarquables dans certains genres institués (Maingueneau, 2004), comme dans les écrits scientifiques ou les rapports d'éducateurs (Sitri & Tutin, 2016). L'intérêt pour la préfabrication en discours est étroitement lié au développement de méthodes informatiques d'exploration qui permettent de repérer ces séquences préconstruites, diversement dénommées « motifs » (Longrée & Mellet, 2013), « motifs séquentiels » (Quiniou *et al.,* 2012), « routines discursives » (Née, Sitri, Veniard 2014), « segments répétés » (Salem, 1987). Ce foisonnement

terminologique s'accompagne également d'une diversité de définitions, des plus discursives (la notion de « routine discursive» chez Née *et al.* 2014) aux plus statistiques (« segments répétés » chez Salem (1987) ou « paquets lexicaux » chez Biber *et al.* (2004)).

Dans cette contribution, nous souhaitons participer à ce débat, à la fois au plan des notions linguistiques, mais aussi au plan méthodologique. Nous souhaitons ainsi proposer quelques définitions pour essayer d'y voir plus clair dans le maquis terminologique des « motifs » et autres « formules » en distinguant trois notions clés : celles de « motif », de « routine » et d' « arbre lexico-syntaxique récurrent » (ou ALR). Par ailleurs, nous désirons observer dans quelle mesure la technique des ALR, basée sur des corpus analysés syntaxiquement, constitue une méthode particulièrement adaptée pour mettre en évidence des motifs correspondant à des routines discursives. Notre champ d'application sera celui des écrits scientifiques, en particulier autour des routines liées aux verbes de communication.

1. Une mise au point terminologique et notionnelle : Motifs, Routines, Arbres lexico-syntaxiques récurrents (ALR)

À notre connaissance, la première utilisation du terme de motif pour désigner des structures linguistiques récurrentes remonte à 2001, avec les travaux de Jean-Gabriel Ganascia en traitement automatique des langues (Ganascia, 2001). Ce dernier propose une méthode d'identification de structures lexico-syntaxiques répétées, regroupées sous-forme de classes, grâce à un calcul de similarité. La méthode est illustrée par l'extraction de motifs caractéristiques de certains écrits de Madame de Lafayette, comparés à ceux d'autres auteurs. Par la suite, la notion de « motif » a été formalisée dans un cadre plus général en analyse de discours et textométrie par Dominique Longrée et Sylvie Mellet pour la description topologique des textes, plus précisément pour « l'étude de la

structuration interne des textes et à [leur] caractérisation au sein d'un corpus contrastif» (Longrée, Luong & Mellet, 2008, p. 734). Dans une première version (Longrée, Luong & Mellet, 2008), le motif est avant tout une association récurrente d'éléments de nature variée (mots, lemmes, traits catégoriels morphosyntaxiques ou autres) dans une structure linéaire multidimensionnelle, pouvant comprendre des éléments facultatifs. Dans une deuxième version (Longrée & Mellet, 2013), la notion de « motif textuel » est précisée, ce type de motif étant défini par une fonction structurante ou discursive, par exemple le motif « *quae cum ita sint* (« les choses étant ce qu'elles sont », « étant donné la situation », « dans ces conditions »). La notion de motif est reprise chez Quiniou, Cellier, Charnois & Legallois (2012), avec la formalisation de la notion d'*itemset*, les unités répétées n'étant pas de simples items éventuellement hétérogènes (p.ex. des mots et des catégories grammaticales), mais des combinaisons de plusieurs traits issus de dimensions diverses (p.ex. un lemme assorti de sa classe morphosyntaxique et d'une classe sémantique). La méthode appliquée sur de simples *items* (p.ex. des mots) impose de considérer des trous (ou « *gaps* ») plus ou moins larges (jusqu'à 5 mots) pour trouver des séquences récurrentes assez générales : *le/la/l' * qui * et * qui*, qui apparaît par exemple dans « la nuit qui m'oppresse et qui trouble mes yeux ». En revanche, les *itemsets*, sans prendre en compte d'éventuels trous qui induisent une forte augmentation de la complexité algorithmique (et donc du temps de calcul) ont ceci d'intéressant qu'ils permettent de généraliser certaines parties du motif en ne retenant que des traits catégoriels : par exemple le motif *des N plus ADJ que* qui apparaît dans « il a *des morsures plus venimeuses que* celles de ta bouche ». Les motifs spécifiques d'un sous-ensemble du corpus, comme c'est le cas de ces exemples qui ont été tirés d'un sous corpus d'œuvres poétiques, sont dits « émergents ». Dans une perspective similaire, Legallois, Charnois et Poibeau (2016) étudient les clichés typiques des romans sentimentaux en extrayant des motifs sous la forme de séquences récurrentes mêlant lexèmes, catégories syntaxiques, catégories

sémantiques : par exemple le patron *PRES le NC, il V-PS* trouvé dans des occurrences telles que « Rouvrant les yeux, elle retint un cri ». Dans cette étude, toutefois, le remplacement de certaines catégories de mots par leurs étiquettes est effectué *a priori*, en amont de l'algorithme de recherche, et les séquences sont contiguës (sans « gap ») : la méthode est donc intermédiaire entre celles des *items* et celles des *itemsets* de Quiniou *et al.* (2012). À travers tous ces usages du terme de motif, il semble qu'on puisse trouver des traits définitoires relativement stables :

- la récurrence : les motifs sont avant tout des structures récurrentes, caractérisées par une fréquence d'occurrence dont la pertinence peut être éventuellement quantifiée par des mesures d'association statistique (information mutuelle, rapport de vraisemblance, spécificité[1], *etc*.) ;
- la séquentialité : les motifs sont des séquences d'unités éventuellement discontinues, pouvant franchir, éventuellement, la frontière de la phrase. L'ordre d'apparition des constituants du motif est fixé : même les « motifs syntaxiques » identifiés sous forme d'arbres par Ganascia (2001) sont linéairement définis, les arbres étant ordonnés ;
- la multidimensionnalité : les motifs peuvent combiner des éléments hétérogènes : lemmes, classes morphosyntaxiques, traits sémantiques ;
- la fonction : la récurrence formelle d'une structure en tant que telle ne suffit pas. Pour être reconnue comme motif, celle-ci doit assumer une certaine fonction sur le plan linguistique, discursif ou textuel, ce que Legallois *et al.* (2016) formulent ainsi :

1 Ces mesures d'association permettent d'évaluer l'écart entre le nombre de cooccurrences observées et le nombre de cooccurrences attendues par le simple jeu du hasard (hypothèse nulle). Ces mesures indiquent si cet écart est significatif, c'est-à-dire hautement improbable, en fonction du seuil de rejet de l'hypothèse nulle qui est retenu.

> Quelle que soit la méthode, sont qualifiés de motifs les seuls segments présentant une régularité d'ordre lexical (un même paradigme lexical est employé dans le patron — par exemple, les lexèmes relatifs au corps), et / ou fonctionnel : un motif possède une fonction sémantique, pragmatique, rhétorique discernable, voire une fonction d'organisation des plans textuels (…) (Legallois et al., 2016)

Longrée et Mellet (2013) identifient quant à eux principalement deux fonctions : une fonction « structurante », qui peut être liée par exemple à des aspects stylistiques, et une fonction « caractérisante », en rapport avec le registre de langue ou le genre textuel.

En ce qui nous concerne, même si nous retenons cette définition générale du motif, nous nous affranchissons partiellement de la contrainte de linéarité : certains motifs sont d'après nous plutôt hiérarchiques que séquentiels (Kraif *et al*., 2016), car ils sont basés sur des structures prédicatives, ce que nous tenterons de démontrer plus loin par l'étude autour des arbres lexico-syntaxiques récurrents (désormais ALR). Un motif est donc, au sens large, une configuration lexico-syntaxique récurrente assumant une certaine fonction, pouvant correspondre à une unité phraséologique (composition de phrasèmes ou de semi-phrasèmes dans le cas des collocations) ou non, caractéristiques d'un texte ou d'un genre textuel.

Plus précisément, nous nommons *routines* un sous-ensemble de motifs discursifs, particulièrement présents dans certains genres institués, où on observe des façons de dire et d'écrire spécifiques. En particulier, les **routines sémantico-rhétoriques** (*cf*. Tutin & Kraif, 2016) présentent un ensemble de caractéristiques :

- elles ont une fonction rhétorique et discursive, propre au genre considéré, et qui ne peut pas nécessairement être déduite du seul contenu lexical. Par exemple, dans les écrits scientifiques, certaines routines associées aux verbes de constat (Ex : (*comme*) *on l'a vu/constaté/observé dans/sur*) ont comme objectif d'apporter une preuve dans l'argumentation.

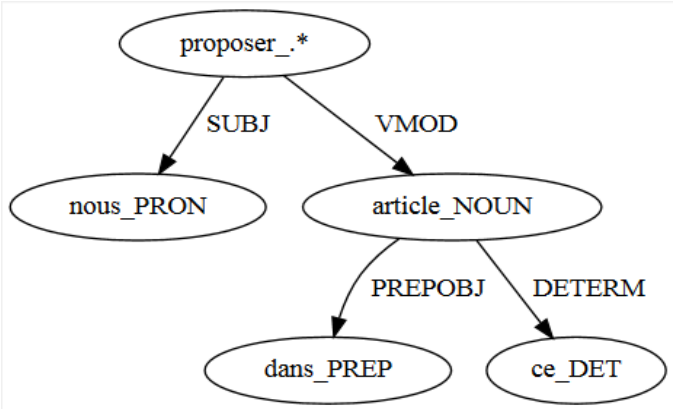
- la routine met en œuvre une configuration lexico-syntaxique, mais les éléments de la configuration sont variables sur le plan lexical. Ils appartiennent à des classes paradigmatiques (*cf*. Bolly, 2011). Par exemple, la routine précédente met en œuvre non seulement le verbe *voir* mais aussi la classe des verbes de constat. Par ailleurs, l'ordre des mots est variable du moment que les relations sémantiques sont identiques[2]. Par exemple, les expressions *Nous reprenons la définition de Duschmoll* ou *la définition adoptée ici est celle de Duschmoll* mettront en jeu les mêmes grands types d'éléments : un agent (*nous, ici*), un objet scientifique (*définition*), un prédicat d'emprunt (*reprenons, adoptée*), une source (*Duschmoll*)
- la routine est généralement une proposition qui est actualisée dans le discours et renvoie à des référents clairement identifiés ; par exemple pour l'écrit scientifique, les auteurs du texte, les référents évoqués dans le texte. Cela différencie clairement pour nous les routines des collocations (*cf*. partie 3).

Enfin, nous distinguons les objets linguistiques, motifs ou routines, des **objets textométriques** relatifs aux outils de traitement de corpus, qui se rapportent à des procédures techniques d'extraction, et dont les sorties, comme on le verra dans la partie 3, peuvent comporter de nombreux éléments sans réel intérêt (ce qu'on peut nommer « du bruit »). Ainsi, la méthode des *segments répétés* (Salem, 1987), également appelés *n-grams* ou *clusters* (Granget & Paquot, 2008 ; Scott, 2004), permet d'identifier la récurrence de séquences contiguës de formes ou de lemmes en surface (p.ex. a- « il aller sans dire que » ou b-« le avoir dire, ». Parmi ces récurrences, certaines correspondent à des motifs (exemple a), d'autres non (exemple b) – ces objets sont les résultats d'une méthode d'identification mais n'ont pas de statut linguistique en tant que tel. Il en va de même pour les ALR, qui sont d'une certaine façon le pendant hiérarchique des segments répétés, mais

2 Ce point est déjà signalé chez plusieurs auteurs comme Sinclair (1991) et Mellet & Longrée (2012) mais n'est pas systématisé dans les représentations qui restent avant tout linéaires.

en s'appuyant sur des relations syntaxiques entre des éléments qui ne sont pas nécessairement contigus et dont l'ordre n'est pas déterminé. L'arbre de la figure 1 représente un ALR, dont une représentation linéaire serait <nous+proposer+dans+ce+article>[3], et qui correspond aux occurrences suivantes tirées de notre corpus (*cf.* partie 3) :

> à partir d'une revue de littérature sur les vingt dernières années, **nous proposons dans cet article** de clarifier les articulations entre ces concepts (...) (Sciences de l'information).
>
> L'étude que **nous proposons dans cet article** met en œuvre un test utilisateur où le sujet , en position debout , est évalué (...) (Psychologie)
>
> **Dans cet article, nous proposons** d'analyser la formation endogène d' une ville monocentrique . (Economie)



3 Nous noterons désormais entre < > la liste des lemmes constituant un ALR. Pour alléger la lecture, les relations hiérarchiques ne sont pas représentées dans cette notation condensée, mais l'on intercale un signe + entre chaque lemme afin d'indiquer que l'ordre n'est pas défini. Les lemmes sont donnés dans un certain ordre à titre illustratif, mais il faut bien garder à l'esprit que l'ordre des unités n'est pas spécifié dans la structure de l'ALR.

Figure 1 : L'ALR correspondant à
<nous+proposer+dans+ce+article>[4]

La section suivante détaille cette approche, dont nous faisons l'hypothèse qu'elle est particulièrement adaptée, sur un plan méthodologique, au repérage des routines.

2. La méthode des ALR

À partir d'un corpus arboré, on peut s'intéresser à l'identification de sous-arbres récurrents. De même qu'il est possible, en parcourant la surface du texte, d'extraire toutes les séquences de 2 à *n* mots (pour une valeur donnée de *n*, p.ex. 8), assorties de leurs fréquences, il est possible d'énumérer, pour un corpus arboré, tous les sous-arbres comportant 2 à *n* nœuds. Mais la combinatoire est beaucoup plus importante dans le cas des arbres : théoriquement, pour un arbre de *t* nœuds, on peut avoir jusqu'à $\sum_{k=2}^{n} \binom{t-1}{k}$ sous arbres comportant 2 à *n* nœuds (Corman, 2012). A titre d'exemple, pour une phrase de 20 tokens[5] on obtient en tout 54 segments

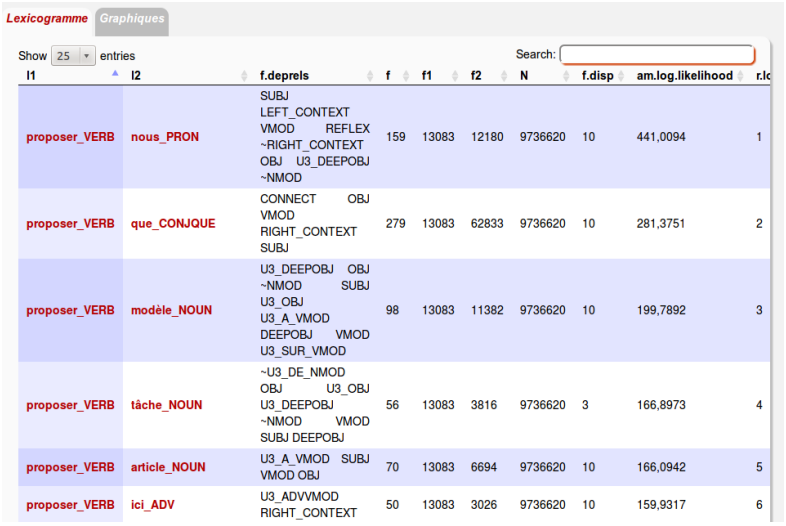
4 Cet arbre illustre le modèle d'analyse en dépendances utilisé par le parseur XIP. Les analyses en dépendances s'inspirent des stemmas de Tesnière (*Éléments de syntaxe structurale*, Klincksieck, Paris 1959), qui schématisent les relations fonctionnelles à l'intérieur de la phrase par des « connexions » – ou liens de dépendance – binaires entre un élément régissant (p.ex. le verbe) et ses éléments subordonnés (p. ex. sujet et objet direct). Ces dépendances relient donc directement des lexèmes entre eux, à différents niveaux de profondeur, dans l'arbre de dépendances, et non des syntagmes complets. L'élément racine, qui régit toute la phrase, est souvent un verbe. Dans cette figure, les relations sont typées par un jeu d'étiquettes déterminé par XIP : SUBJ pour sujet ou premier actant, OBJ pour objet direct ou deuxième actant, VMOD pour circonstant, PREPOBJ pour la relation entre la tête du groupe prépositionnel et la préposition, DETERM pour la détermination du nom.

5 Par « token », on désigne l'ensemble des unités graphiques issues de la segmentation d'une phrase : mots, ponctuations, nombres, *etc*.

répétés (désormais SR) contigus de longueur 2 à 4, contre 704 sous-arbres de 2 à 4 nœuds (*ibid*.).

Pour résoudre le problème de calculabilité lié à cette explosion combinatoire, nous le simplifions en nous intéressant aux cooccurrences binaires entre nœuds reliés par des relations syntaxiques (en l'occurrence des relations de dépendances). Ainsi, la méthode des ALR a été développée au sein d'une architecture logicielle centrée sur la notion de « cooccurrence syntaxique », pour reprendre les termes de Evert (2007), qui caractérise une association statistique significative reliant deux mots par une relation syntaxique, par exemple (*jouer* →OBJ→ *rôle*). Nous nous sommes appuyés sur un outil baptisé Lexicoscope (Kraif & Diwersy, 2012 ; 2014), qui permet d'extraire, pour un pivot (ou mot-pôle) donné, un tableau enregistrant, dans un espace de relation syntaxique déterminé (qui peut couvrir toutes les relations ou seulement un sous-ensemble de celles-ci), ses collocatifs syntaxiques les plus significatifs. Ce tableau est nommé lexicogramme, et présente les cooccurrents significatifs d'une manière analogue au Word Sketches du Sketch Engine (Kilgariff & Tugwell, 2001), sauf que les relations concernées ne sont pas séparées dans différents tableaux en fonction de leur type. Par exemple, pour *proposer*, on obtient les collocatifs syntaxiques suivants, triés par degré d'association décroissant (avec la mesure du rapport de vraisemblance notée *loglike*[6], *cf.* Dunning, 1993) :

6 On trouve de nombreuses mesures dans la littérature, comme l'information mutuelle spécifique, le *t-score,* le Dice, ou le χ^2. Le *loglike*, proche du χ^2, présente l'avantage de ne pas surpondérer les événements de basse fréquence, comme l'information mutuelle spécifique. C'est une mesure fiable dans de nombreux cas de figure (Evert, 2007). Par défaut, on retient les cooccurrences qui obtiennent un score supérieur à 10,83, ce qui correspond à une probabilité inférieure à 1/1000 d'obtenir le tableau de contingence par le seul jeu du hasard.



Lexicogramme | Graphiques

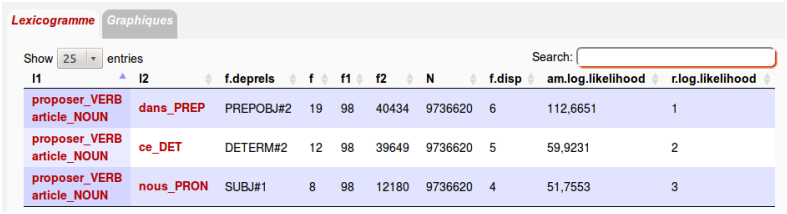
Show 25 entries — Search:

l1	l2	f.deprels	f	f1	f2	N	f.disp	am.log.likelihood	r.lo
proposer_VERB	nous_PRON	SUBJ LEFT_CONTEXT VMOD REFLEX ~RIGHT_CONTEXT OBJ U3_DEEPOBJ ~NMOD	159	13083	12180	9736620	10	441,0094	1
proposer_VERB	que_CONJQUE	CONNECT OBJ VMOD RIGHT_CONTEXT SUBJ	279	13083	62833	9736620	10	281,3751	2
proposer_VERB	modèle_NOUN	U3_DEEPOBJ OBJ ~NMOD SUBJ U3_OBJ U3_A_VMOD DEEPOBJ VMOD U3_SUR_VMOD	98	13083	11382	9736620	10	199,7892	3
proposer_VERB	tâche_NOUN	~U3_DE_NMOD OBJ U3_OBJ U3_DEEPOBJ ~NMOD VMOD SUBJ DEEPOBJ	56	13083	3816	9736620	3	166,8973	4
proposer_VERB	article_NOUN	U3_A_VMOD SUBJ VMOD OBJ	70	13083	6694	9736620	10	166,0942	5
proposer_VERB	ici_ADV	U3_ADVVMOD RIGHT_CONTEXT	50	13083	3026	9736620	10	159,9317	6

Figure 2 : Un lexicogramme pour le pivot *proposer*.

Outre les statistiques fréquentielles et les mesures d'associations choisies, ce lexicogramme contient des informations sur les relations syntaxiques mises en jeu, ainsi que sur la *dispersion*, qui indique le nombre de sous-corpus où la cooccurrence a été identifiée. Cette donnée est utile pour cibler des phénomènes généraux, partagés par l'ensemble des sous-corpus étudiés, certaines récurrences pouvant être très saillantes dans une petite partie du corpus (voire un seul document) sans pour autant avoir de portée générale.

L'architecture du Lexicoscope permet d'étudier les collocatifs pour des pivots simples, mais aussi pour des arbres, nommés *pivots complexes*, comparables à ce que Rainsford & Heiden (2014) nomment des *keynodes*. Ainsi pour le sous-arbre <proposer+article> on peut extraire les collocatifs de la figure 3 :



Lexicogramme | Graphiques

Show 25 entries — Search:

l1	l2	f.deprels	f	f1	f2	N	f.disp	am.log.likelihood	r.log.likelihood
proposer_VERB article_NOUN	dans_PREP	PREPOBJ#2	19	98	40434	9736620	6	112,6651	1
proposer_VERB article_NOUN	ce_DET	DETERM#2	12	98	39649	9736620	5	59,9231	2
proposer_VERB article_NOUN	nous_PRON	SUBJ#1	8	98	12180	9736620	4	51,7553	3

Figure 3 : Extraction d'un lexicogramme pour un sous-arbre

On voit que ces cooccurrences, une fois composées entre elles deux-à-deux, permettent de reconstituer l'arbre complet de la figure 1, correspondant à la routine ‘on/nous + proposer + dans + cet + article‘[7]. À partir de ces outils, fondés sur la cooccurrence binaire entre un sous-arbre et un pivot simple, nous avons développé une méthode d'extraction itérative des arbres récurrents. Cette méthode, entièrement automatisée, fonctionne de la manière suivante :

1. on part d'un pivot initial (mot simple ou arbre) ;
2. on en extrait le lexicogramme ;
3. tous les collocatifs dépassant un certain seuil de cooccurrence et de mesure d'association (ici le *loglike*) sont rattachés au pivot, avec la relation concernée, pour former des arbres augmentés ;
4. on réitère l'étape 2 en reprenant ces nouveaux arbres comme pivot ; le processus est répété tant que l'on obtient, pour augmenter les arbres, des collocatifs dépassant les seuils de significativité, et que les arbres extraits n'ont pas dépassé une certaine longueur (paramétrable : dans la suite, la longueur sera fixée à 8 éléments).

---

7 Afin d'éviter la confusion entre une instance précise, telle que *Dans cet article, nous avons proposé…* et la routine sous-jacente, qui n'impose pas d'ordre particulier sur l'ordre des éléments et qui peut intégrer des éléments plus génériques (des paradigmes ou classes lexicales), nous noterons désormais les routines entre ‘ ‘, et nous intercalerons des + entre les éléments. Pour nommer telle ou telle instance, après lemmatisation, nous utiliserons le terme générique de *séquence*.

Les arbres ainsi obtenus sont nommés ALR, pour arbres lexico-syntaxiques récurrents. Ces étapes sont illustrées sur la figure 4, pour l'ALR correspondant à <proposer+dans+ce+article> :

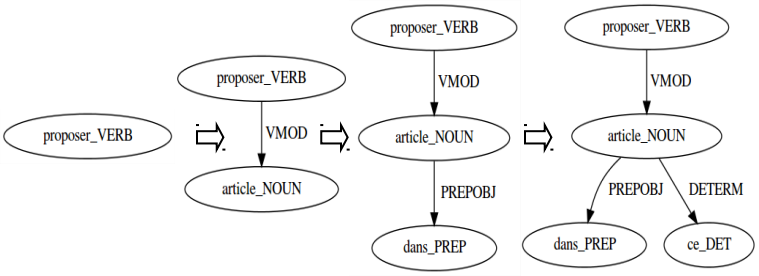


Figure 4 : Extraction de l'ALR <proposer+dans+ce+article>

Cette méthode repose sur l'hypothèse que la plupart des arbres récurrents intéressants en tant que motifs comportent au moins deux nœuds adjacents fortement associés[8], ce qui permet d'amorcer le processus itératif. Une fois les deux premiers nœuds regroupés en un seul arbre, les mesures d'associations avec les autres nœuds sont habituellement élevées, même si les mesures d'association des mots simples pris deux à deux sont initialement faibles (car la fréquence du sous-arbre constitué des deux premiers nœuds est en général beaucoup plus faible que la fréquence des mots pris isolément). L'analyse des résultats sur corpus permettra de juger si cette hypothèse est valide, et d'évaluer le silence dans les résultats, c'est-à-dire d'analyser, au moins qualitativement, les expressions intéressantes non identifiées.

8 Par cette approche heuristique, nous sommes conscients néanmoins de laisser de côté certaines expressions intéressantes. La comparaison des résultats avec les SR permettra de mieux évaluer l'intérêt et les limites de la méthode.

3. Une étude de corpus : comparaison des segments répétés et des ALR des verbes de communication dans les écrits scientifiques

3.1 Objectifs de l'étude

Cette étude de corpus a pour objectif de comparer, à travers des exemples concrets, les différents types de segments extraits par la méthode des ALR présentée en section 2 et une méthode classique très utilisée en phraséologie et en stylistique, la méthode des *segments répétés* (ou n-grammes) qui repère les suites récurrentes de mots, lemmes ou ponctuations contigus (Salem 1987). On s'intéressera en particulier aux segments récurrents associés aux verbes de communication, qui ont été traités au sein du lexique scientifique transdisciplinaire développé dans le cadre du projet Termith[9] (Hatier *et al.* 2016). Les 25 verbes de communication retenus relèvent de plusieurs sous-champs sémantiques[10] : 'mention', 'discussion', 'faire savoir', 'formulation', 'insistance', 'proposition'. Autour de ces verbes, on peut extraire des segments tels que *comme on l'a dit*, *nous l'avons souligné* ou des segments non pertinents comme *article se propose de*, car le segment est ici incomplet (le déterminant n'est pas présent et le motif n'est donc pas « actualisée »[11]). Parmi ces segments, les routines associées à des fonctions rhétoriques et discursives dans l'écriture scientifique nous intéressent particulièrement.

Le corpus utilisé pour l'expérimentation est le corpus Termith, un corpus arboré contenant 500 articles scientifiques d'environ 5

9 Voir sur le site : http://www.atilf.fr/ressources/termith/.

10 La liste des verbes est la suivante : *citer, débattre, dire, diffuser, discuter, émettre, énoncer, évoquer, exposer, exprimer, formuler, indiquer, insister, intervenir, invoquer, médiatiser, mentionner, préciser, proposer, résumer, révéler, signaler, souligner, suggérer, véhiculer*.

11 En revanche, des segments (lemmatisés) comme *ce article proposer de*, ou *le article propose de* correspondent pour nous à des routines.

millions de mots, rassemblés dans le cadre des projets Scientext et Termith (Hatier *et al.,* 2014), et annoté syntaxiquement à l'aide du logiciel XIP (Aït-Mokhtar *et al.,* 2002). Ce corpus est partitionné en 10 domaines : anthropologie, économie, histoire, géographie, linguistique, psychologie, sciences de l'éducation, sciences de l'information, sciences politiques, sociologie. Le calcul de dispersion est basé sur cette partition. À partir d'une typologie linguistique des segments extraits, nous évaluerons d'un point de vue quantitatif et qualitatif les points forts de chacune des méthodes utilisées.

3.2 Méthodes d'extraction et typologie linguistique des segments extraits

Les méthodes d'extraction utilisent le corpus sous forme lemmatisée. Les segments répétés (SR) ont été extraits à partir d'un script développé par nous, repérant les suites contiguës de mots et ponctuations faibles (essentiellement, virgules) apparaissant au moins 8 fois dans au moins trois disciplines. De la même façon, ont été extraits les ALR apparaissant au moins 8 fois à chaque itération (avec une mesure de rapport de vraisemblance supérieure à 10,81) dans au moins 3 disciplines[12]. Aucune mesure de spécificité n'a été employée, notre étude ne portant que sur les régularités endogènes au corpus scientifique : la comparaison avec un corpus exogène fera l'objet d'une étude ultérieure. En revanche, la mesure de dispersion s'est révélée utile pour cibler les expressions transdisciplinaires, et donc les routines spécifiques au genre article scientifique sans être spécifique à une seule discipline.

Dans un second temps, nous avons caractérisé les segments extraits, en nous appuyant sur une typologie linguistique, de façon à mieux comprendre la complémentarité des méthodes. Nous avons bien entendu prêté une attention particulière aux routines, mais

12 Ces seuils sont déterminés de façon empirique à partir de nos observations préliminaires et dépendent étroitement de la taille et des caractéristiques du corpus.

d'autres types de segments se sont également dégagés : collocations, constructions syntaxiques spécifiques, expressions complètement figées ainsi que des expressions non pertinentes du point de vue phraséologique et/ou fonctionnel (ce qu'on pourrait qualifier comme « bruit »). Un retour aux exemples du texte a souvent été nécessaire, pour caractériser les segments avec plus de précision. Nous présentons ci-dessous le détail de cette typologie.

a. Routines

Les routines, comme présentées dans la partie 1, sont des motifs, qui correspondent généralement à des propositions, qui sont actualisés (souvent par un pronom, une détermination ou un adverbe à fonction déictique) et qui remplissent une fonction dans l'écrit scientifique, telle qu'effectuer une démonstration, apporter une preuve, renvoyer à un autre segment du texte (fonction métatextuelle), etc. Les segments suivants, qui sont lemmatisés lors de l'extraction, correspondent à des routines (ou des instances de routines) ainsi que nous les avons définies :

ce résultat suggérer que
comme le avoir souligner
comme nous le avoir souligner
il falloir dire que
il falloir souligner que

b. Collocations

À la différence des routines, les collocations sont avant tout des associations lexicales binaires de mots pleins, mais elles peuvent, bien entendu, être utilisées dans des routines (Hausmann, 1989 ; Tutin, 2013 ; Tutin & Kraif, 2016). Les segments suivants correspondent pour nous à des collocations.

formuler le hypothèse
insister sur le caractère
proposer un analyse

a. Constructions syntaxiques spécifiques

Nous avons également intégré le type ‘construction’ pour les constructions syntaxiques, principalement les alternances, qui sont caractéristiques de certains verbes : il peut s’agir du passif, du passif impersonnel, du passif pronominal ou de constructions à l’aide de modaux ou de négations, certaines de ces constructions apparaissant tout à fait caractéristiques du genre et de certaines classes de verbes (Fifielska, 2015). Nous n’avons pas intégré ici les sous-catégorisations (les « colligations ») des verbes, c’est-à-dire la cooccurrence entre un lexème et une configuration syntaxique.

avoir être souligner
avoir être soumettre
être évoquer par
permettre de préciser
pouvoir se résumer

b. Locutions figées

Ce type intègre les expressions polylexicales figées, dont le sens n’est pas compositionnel, contrairement aux collocations, par exemple :

c’est-à-dire
qui vouloir dire
aller sans dire

c. Non pertinent

Le type Non pertinent est utilisé pour des segments qui ne correspondent à aucun des types précédents. Les segments de ce type peuvent être intégrés dans d’autres segments pertinents ou être plus larges que les segments des types précédents. En voici quelques exemples :

, il proposer
, proposer par
, qui marquer

avoir dire que il
celui proposer par
dire que ce être

3.3 Comparaison quantitative

Les extractions effectuées sur les segments répétés aboutissent à de nombreuses séquences. Pour limiter le bruit de manière simple, nous avons effectué un premier filtrage, en supprimant les SR qui comportent des erreurs d'étiquetage ainsi que les SR qui finissent par un déterminant, comme dans le cas suivant :

,_PUNCT|comme_CONJ|le_PRON|indiquer_VERB|le_DET

En effet, nous avons remarqué que, dans la majorité écrasante des cas, ces expressions sont redondantes avec des SR plus pertinents, qui soit n'incluent pas le DET, soit incluent un N pertinent à la suite du DET, par exemple : insister_VERB| sur_PREP|le_DET|nécessité_NOUN. Après filtrage, il reste un total de 435 SR à examiner.

Les ALR extraits sont quant à eux beaucoup moins nombreux, puisqu'ils ne comptent que 276 éléments, soit un peu plus de la moitié des SR. Cette parcimonie des ALR est liée, nous semble-t-il, au seuil élevé de la mesure d'association (*loglike* > 10,83) qui filtre les cooccurrences syntaxiques. Notamment, de très nombreux SR incluant des signes de ponctuation n'ont pas été couverts par des ALR, la mesure d'association étant plutôt faible quand elle met en jeu des formes aussi fréquentes. Parmi les SR extraits, 124

segments sont couverts par des ALR[13]. 45% des segments repérés par les ALR sont donc également extraites par les SR.

Pour évaluer l'intérêt des méthodes employées, il faut maintenant examiner la pertinence des séquences extraites, et leur intérêt pour les études linguistiques envisagées. On se demandera en particulier dans quelle mesure les routines sémantico-rhétoriques du corpus ont été repérées.

Nous avons classé, d'une part, les séquences extraites par les SR, et d'autre part les ALR obtenus, en reprenant la typologie décrite ci-dessus. La figure 5 indique les résultats de cette analyse, en utilisant les résultats bruts, alors que la figure 6 indique la répartition relative pour chaque méthode.

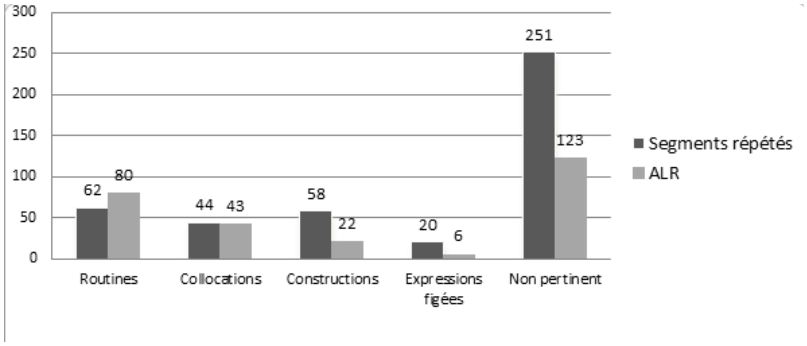


Figure 5 : Comparaison des résultats par type (résultats bruts)

13 À strictement parler, les SR et les ALR ne sont pas tout à fait comparables. Par exemple, l'ALR <on+pouvoir+dire> correspond en fait à la structure hiérarchique [subj(pouvoir,on),obj(pouvoir,dire),subj(dire,on)] qui permet de repérer des segments non continus par exemple *on peut [ainsi] dire* ou adoptant un ordre différent comme *ainsi, peut-on dire*... Ainsi, un même ALR peut couvrir plusieurs SR à la fois, et concerner plus d'occurrences dans le texte que tous les SR qu'il couvre.

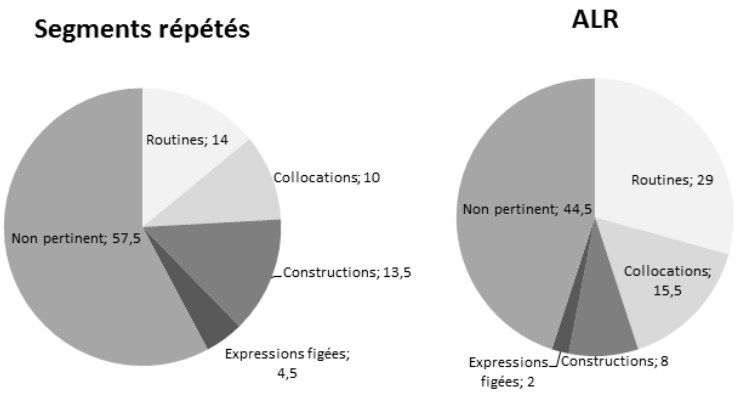


Figure 6 : Comparaison des répartitions relatives par type

De manière générale, les résultats confirment assez bien nos attentes. Au niveau des résultats bruts, les ALR extraient moins d'expressions que les SR[14] mais davantage de routines et un nombre comparable de collocations. En revanche, pour les expressions figées et les constructions, qui sont davantage des phénomènes surfaciques, le rappel de la technique des SR apparaît supérieur. Le contraste entre les deux approches est encore plus flagrant si l'on examine la répartition des types en termes de pourcentage (*cf.* Figure 6). La technique des ALR produit sans conteste des résultats plus satisfaisants pour les phénomènes phraséologiques « étendus » qui nous intéressent (routines et collocations), puisque près de la moitié des résultats entrent dans ces deux catégories, mais elle est décevante pour les expressions figées et les constructions. La précision globale des ALR est de 55,5%, de 13 points supérieure à celle des SR qui est de 42,5%, mais vu que la méthode est beaucoup plus sophistiquée, on s'attendait à une précision supérieure (le bruit des résultats non pertinents reste considérable à près de 44,5%).

L'utilisation conjointe des deux méthodes permet-elle d'obtenir de meilleurs résultats ? Pour observer cela, nous avons

14 N'oublions pas aussi que nous avons éliminé les SR terminant par un déterminant.

également observé la répartition des séquences extraites par les deux méthodes à la fois (*cf.* Figure 7).

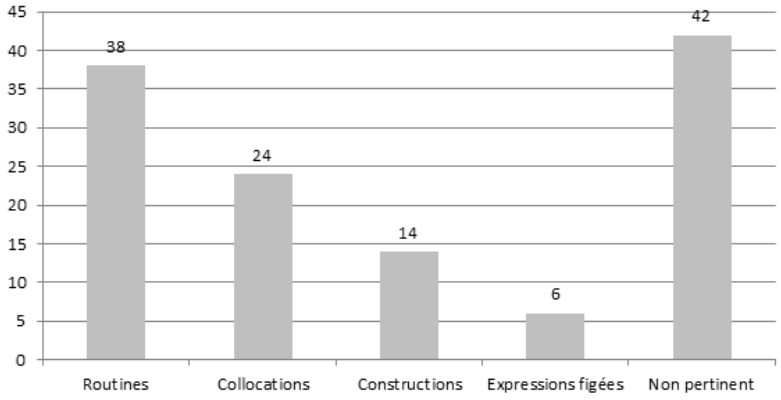


Figure 7 : Répartition des types de séquences extraites par les deux méthodes (sur le nombre total de séquences)

L'utilisation conjointe des deux méthodes semble donner des résultats intéressants en ce qui concerne la précision, car c'est la méthode qui extrait le moins de résultats non pertinents (la précision est cette fois de 66 % environ). Le rappel reste toutefois décevant pour les phénomènes qui nous intéressent (38 routines contre 80 avec la technique des ALR), ce qui la rend peu adaptée pour les études linguistiques envisagées.

L'analyse qualitative dans la section suivante permettra d'explorer plus en finesse les résultats.

3.4 Comparaison qualitative

La comparaison quantitative n'est pas suffisante pour analyser la complémentarité des méthodes et la spécificité des approches utilisées. Pour cela, nous observerons les types de séquences extraites.

a) Les routines

L'observation des routines[15] extraites à la fois par les deux méthodes montre qu'il s'agit d'expressions contenant des éléments généralement contigus, facilement repérables par les deux techniques, comme on peut en observer quelques-unes dans le tableau 1. De manière générale, les fréquences observées pour les SR sont moins importantes que pour les ALR, ce qui peut facilement s'expliquer par l'insertion possible de certains modifieurs, comme des adverbes, ou une modification de l'ordre (*cf.* Tableau 1).

	Fréquence d'occurrence	
Routine	SR	ALR
'ça + vouloir + dire'	12	16
'ce + article + proposer'	23	40
'ce + article + se + proposer'	15	15
'ce + que + indiquer'	8	9
'ce + que + souligner'	14	12
'ce + que + suggérer'	12	13
'ce + qui + vouloir + dire'	8	15
'ce + résultat + indiquer'	10	16
'ce + résultat + suggérer'	13	12
'cela + vouloir + dire'	9	30
'comme + le + avoir + souligner'	14	39

15 Le terme de « routine » est en quelque sorte ici un raccourci. Il s'agit plutôt d'élément appartenant à une routine (puisque la routine est une structure abstraite).

‘comme + le + dire’	11	15
‘comme + le + indiquer’	45	60
‘comme + le + préciser’	8	8
‘comme + le + révéler’	9	9
‘comme + le + signaler’	12	14
‘comme + le + souligner’	95	97
‘comme + le + suggérer’	31	34
‘comme + nous + le + avoir + dire’	9	11

Tableau 1 : Routines extraites à la fois par les SR et les ALR

Cela explique aussi que certaines routines d’assez faible fréquence qui échappent aux SR soient repérées tout de même par les ALR, par exemple : ‘cela + revenir + à + dire‘ (freq = 8), ‘comme + le + avoir + mentionner‘ (fréq. = 8), ‘il + falloir + insister‘ (fréq. = 10), ‘il+falloir+mentionner’ (fréq. = 8) … Par ailleurs, parmi les routines uniquement repérées par les ALR, on observe bien entendu des routines dont les éléments sont souvent éloignés ou dans des positions variables, comme ‘on + émettre + hypothèse’, ‘on + pouvoir + émettre + hypothèse’ où le déterminant est variable.

Inversement, quelques routines sont mieux repérées par les SR que les ALR. Il s’agit principalement de segments comme ‘ce + article + se + proposer + de‘, ‘ce + résultat + indiquer + que‘, ‘ce + résultat + suggérer + que‘, ‘en + insister + sur + le + fait + que‘ qui finissent par des prépositions ou des conjonctions. L’omission des prépositions et des conjonctions avec les ALR est principalement due au modèle syntaxique de dépendance utilisé, puisque ces mots grammaticaux ne sont pas directement rattachés aux verbes et noms

qui les introduisent, mais à l'élément régi[16]. Cette information pourrait toutefois être intégrée avec les ALR, mais supposerait un post-traitement.

En bref, les ALR sont dans l'ensemble plus pertinents pour l'extraction des routines, même si l'on ne repère que peu de routines longues, du fait de la non contiguïté de certains éléments (et du seuil élevé de fréquence que nous avons fixé à 8). L'intégration des mots grammaticaux doit cependant encore être réfléchie dans le cadre de cette méthode.

b) Les collocations

Parmi les collocations communes repérées, beaucoup sont de la structure V Det N ou Det N V. Les SR fonctionnent ici bien lorsque le déterminant est fixe (par exemple, dans 'souligner+le+importance','souligner+le+difficulté'). En revanche, les ALR apparaissent bien évidemment plus efficaces quand les déterminants sont variables (par exemple, 'insister+sur+aspect','insister+sur+dimension' qui renvoient à des exemples comme *nous insistons sur cet aspect* ou *il faut insister sur cette dimension*) ou des insertions possibles (Ex : *'proposer+de+étudier'* qui correspond à des instances textuelles comme *nous proposons ici d'étudier* ou *nous proposons d'en étudier*). Enfin, parmi les collocations repérées uniquement par les SR, on relève de nombreuses associations Verbe-Adverbe mal repérées par les ALR comme *cité plus haut*, *dire plus haut, se révéler ainsi...*

Dans les deux premier cas, il s'agit d'un problème d'analyse syntaxique : le syntagme *plus haut* n'est tout simplement pas rattaché au reste de la phrase, pour des raisons inexpliquées. Pour le troisième, on constate que l'analyseur a tendance à rattacher *ainsi* à l'adjectif qui suit la séquence plutôt qu'au verbe, ce qui explique qu'elle ne soit pas rattachée à l'ALR.

---

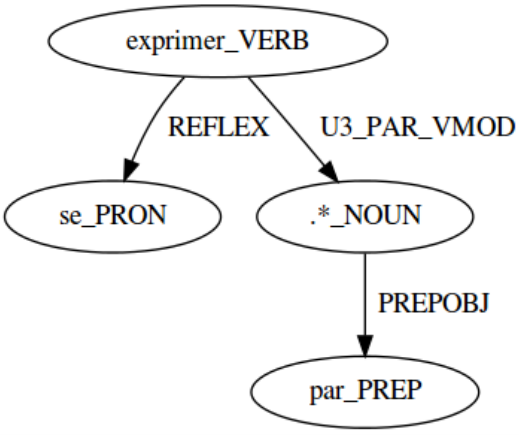
16 Ce phénomène est facile à observer sur la figure 1. Dans l'analyse de dépendance de *nous proposons dans cet article* effectuée par XIP*, dans* dépend non de *proposer* mais de *article.*

Les deux méthodes apparaissent donc assez complémentaires pour ce type d'expression, même si les ALR, comme on l'a vu plus haut, ont un meilleur rappel, en particulier pour les routines, car ils sont plus à mêmes de capturer certaines variations[17].

c) Les constructions

Parmi les 68 constructions repérées, 14 seulement sont communes (par exemple, 'pouvoir+être+proposer','avoir+être+formuler', 'permettre+de+souligner'). De manière générale, les SR extraient des structures plus diversifiées que les ALR, qui peinent à extraire les constructions négatives ou impliquant des prépositions.

Là encore, cela est souvent dû au modèle syntaxique employé par le parseur. Par exemple, dans la construction 'se+exprimer+par', la préposition se retrouve subordonnée au nom, comme le montre l'arbre de la figure 8 :



17 Rappelons également que nous n'extrayons ici que des expressions de longueur 3 au moins.

Figure 8 : Exemple de non-rattachement direct de la préposition au verbe dans le modèle syntaxique du parseur

Pour les constructions impliquant la négation (p.ex. *ne pas intervenir*), on observe que les ALR identifient une partie de l'expression (<pas+intervenir>), mais sans l'adverbe *ne*, qui n'est jamais rattaché à la phrase par le parseur.

En bref, si ici les SR se révèlent plus adaptés au repérage des constructions, c'est essentiellement dû aux lacunes du modèle d'annotation en dépendance, qu'il faudrait modifier et corriger, peut-être avec un post-traitement. Notons toutefois que ce problème n'est pas spécifiquement lié à XIP : la plupart des parseurs en dépendances rejettent les éléments grammaticaux (préposition, négation, pronom réfléchi, conjonction de coordination, ...) dans les feuilles de l'arbre, ce qui rend difficile leur rattachement dans nos ALR.

d) Les expressions figées

Les expressions figées sont dans l'ensemble peu nombreuses (ex : *c'est-à-dire, cela va sans dire, cela n'est pas sans évoquer*), puisqu'on n'en relève que 22. Pour ces expressions de surface, les SR apparaissent dans l'ensemble légèrement plus efficaces que les ALR, ce qui était prévisible vu leur quasi-invariabilité.

4. Conclusion

Nous avons mené une étude empirique destinée à comparer deux méthodes d'identification de motifs textuels, les motifs étant entendus comme des expressions polylexicales récurrentes, éventuellement multidimensionnelles, et assurant une certaine fonction, caractérisante ou structurante, sur les plans discursifs et

textuels[18]. En comparant la méthode des segments répétés et celle des arbres lexico-syntaxiques récurrents, nous avons mis de côté la multidimensionnalité, en interrogeant seulement le caractère séquentiel ou hiérarchique de certains motifs. Parmi les expressions identifiées, nous avons trouvé un certain nombre de routines sémantico-rhétoriques dont le caractère plus hiérarchique que séquentiel est apparu par une meilleure couverture des ALR par rapport au SR – les routines étant susceptibles de se réaliser de manière variable tant au niveau des insertions possibles que de l'ordre des unités.

Les ALR semblent bien adaptés pour un autre type d'expressions polylexicales récurrentes, les collocations, elles-mêmes étant assez flexibles dans leurs réalisations. En revanche, pour les deux autres types d'expressions que nous avons caractérisés, à savoir les constructions et les expressions figées, la méthode des SR est apparue plus directe et plus efficace, le rattachement syntaxique de certaines unités (prépositions, adverbes de négation, conjonction de subordination, *etc*.) apparaissant comme étant problématique avec le modèle utilisé par notre parseur.

Notons que pour ces trois derniers types d'expression, le statut de motif n'est pas systématique : pour s'assurer de leur fonction éventuellement caractérisante, il faudrait calculer leur spécificité en comparant le corpus d'étude à un corpus de référence, ce qui pourra faire l'objet d'études complémentaires. Quant à la fonction structurante, elle doit être déterminée par un retour au texte : ces outils fournissent des pistes et des guides pour la découverte des motifs, au fil du texte, mais *in fine*, seule l'interprétation du linguiste permet d'en déterminer les fonctions du point de vue global de la textualité. Enfin, dans de futurs travaux, nous

18 Rappelons ici que les méthodes en tant que telles, si elles permettent d'extraire des objets textométriques pertinents, ne permettent en aucun cas dans leur forme actuelle de caractériser les éléments extraits.

aimerions développer l'aspect multidimensionnel des ALR, en développant des techniques permettant d'identifier automatiquement, au niveau de certains nœuds, des catégories morphosyntaxiques plutôt que des unités lexicales. Le potentiel structurant de ce type de représentation hiérarchique permet en effet de substituer les lemmes à des classes syntaxiques ou sémantiques plus générales, mieux à même de rendre compte de la structuration abstraite de certains motifs.

Références

Aït-Mokhtar S., Chanod J.-P., Roux C. (2002). "Robustness beyond shallowness: incremental dependency parsing", *Special issue of the NLE Journal*, 8(2/3): 121-144.

Biber D., Conrad S., Cortes V. (2004). "*If you look at*…: Lexical bundles in university teaching and textbooks", *Applied linguistics*, *25*(3) : 371-405.

Bolly C. (2011). *Phraséologie et collocations*. *Approche sur corpus en français L1 et L2*. Bruxelles : Peter Lang.

Corman J. (2012). *Extraction d'expressions polylexicales sur corpus arboré*, Mémoire de master recherche Industries de la langue, Université Stendhal Grenoble 3.

Dunning T. (1993). "Accurate Methods for the Statistic of Surprise and Coincidence", *Computational Linguistics*, 19(1) : 61-76.

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

Fifielska E. (2015). *Les constructions syntaxiques de l'écrit scientifique: exploration et analyses de corpus*. Mémoire de master 2 Recherche, Université de Grenoble 3.

Ganascia J.-G. (2001). « Extraction automatique de motifs syntaxiques », *Actes de TALN 2001*, Tours, 2-5 juillet 2001.

Granger S., Paquot M. (2008). Disentangling the phraseological web. in S. Granger and F. Meunier (dir.), *Phraseology: An Interdisciplinary Perspective*. Amsterdam: John Benjamins, 27-49.

Hatier S., Augustyn M., Tran T. T. H., Yan R., Tutin A., Jacques M.-P. (2016). "French cross-disciplinary scientific lexicon: extraction and linguistic analysis", *Proceedings of Euralex*, Tbilissi, 6-10 september 2016, 355-366.

Hausmann F. J. (1989). « Le dictionnaire de collocations », *in* Hausmann, F.J., Reichmann, O., Wiegand, H.E., Zgusta, L. (eds), *Wörterbücher : ein internationales Handbuch zur Lexicographie. Dictionaries. Dictionnaires*. Berlin : De Gruyter, 1010-1019.

Kilgariff A.,Tugwell D. (2001). "WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography", *Proc ACL workshop on COLLOCATION Computational Extraction Analysis and Exploitation*, Toulouse July 2001, 32-38.

Kraif O., Diwersy S. (2012). « Le Lexicoscope : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques », *Actes de la conférence TALN 2012*, Grenoble, 399-406.

Kraif, O., Diwersy, S. (2014). "Exploring combinatorial profiles using lexicograms on a parsed corpus: a case study in the lexical field of emotions", *in* Blumenthal P., Novakova I., Siepmann D.(éd). *Les émotions dans le discours. Emotions in discourse.* Bern *:* Peter Lang, 381-394.

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 », Dans (Vol. 8, p. 2663-2674), *Actes de la 4e édition du Congrès Mondial de Linguistique Française (CMLF)*, Berlin, 19-23 juillet 2014, 2663-2674

Kraif O., Novakova I., Sorba J. (2016). « Constructions lexico-syntaxiques spécifiques dans le roman policier et la science-fiction », *Lidil* 53 : 143-159.

Legallois D., Tutin A. (2013). « Présentation: Vers une extension du domaine de la phraséologie », *Langages*, (1) : 3-25.

Legallois D., Charnois T., & Poibeau T. (2016). « Repérer les clichés dans les romans sentimentaux grâce à la méthode des "motifs" », *Lidil. Revue de linguistique et de didactique des langues*, (53) : 95-117.

Longrée D., Mellet S. (2013). « Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours ». *Langages*, *189,*(1) : 65-79.

Maingueneau D. (2004). « Retour sur une catégorie : le genre », *in* J.-M. Adam, J.-B. Grize et Magid Ali Bouacha (eds), *Texte et discours : catégories pour l'analyse*. Dijon : Editions Universitaires de Dijon, 107-118.

Mellet S., Longrée D. (2012). « Légitimité d'une unité textométrique: le motif », *in* A. Dister, D. Longrée & G. Purnelle (éds)*, JADT* : 716-728.

Née E., Sitri F., Veniard, M. (2014). « Pour une approche des routines discursives dans les écrits professionnels », *Actes de la 4e édition du Congrès Mondial de Linguistique Française (CMLF)*, Berlin, 19-23 juillet 2014, 2113-2124.

Quiniou, S., Cellier, P., Charnois, T., Legallois, D. (2012). « Fouille de données pour la stylistique: cas des motifs séquentiels émergents », In *Proceedings of the 11th International Conference on the Statistical Analysis of Textual Data, Liege,* 821-833.

Rainsford Th., Heiden S. (2014). « Key Node in Context (KNIC) Concordances: Improving Usability of an Old French Treebank", *in Actes de la 4e édition du Congrès Mondial de Linguistique Française* (CMLF). Berlin, 19-23 juillet 2014, Vol. 8, 2707-2718.

Salem A. (1987). *Pratique des segments répétés*. Paris : Klincksieck.

Scott M. (2004-6), *WordSmith Tools. Version 4.0. Manual*. Oxford: Oxford University Press.

Sinclair J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sitri F., Tutin A. (eds) (2016). « Phraséologie et genres de discours: Patrons, Motifs et Routines », *Lidil* 53.

Tesnière L. (1959). *Eléments de syntaxe structurale*. Paris : Librairie C. Klincksieck.

Tutin A.(2013). « Les collocations lexicales: une relation essentiellement binaire définie par la relation prédicat-argument », *Langages* (1) : 47-63.

Tutin A., Kraif O.. (2016). "From binary collocations to grammatically extended collocations: Some insights in the semantic field of emotions in French", *Mémoires de la Société néophilologique de Helsinki*, Helsinki : Société néophilologique de Helsinki, 245-266.

Tutin, A., Kraif, O. (2016). « Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines: l'apport des arbres lexico-syntaxiques récurrents ». *Lidil. Revue de linguistique et de didactique des langues*, (53), 119-141.