Segments phraséologiques et séquences textuelles

Questions méthodologiques et approches topologiques

Marion Bendinelli

Édition électronique

URL : <http://journals.openedition.org/corpus/2844>
ISSN : 1765-3126

Éditeur

Bases ; corpus et langage - UMR 6039

Édition imprimée

Date de publication : 15 janvier 2017
ISBN : 1638-9808
ISSN : 1638-9808

Référence électronique

Marion Bendinelli, « Segments phraséologiques et séquences textuelles », *Corpus* [En ligne], 17 | 2017, mis en ligne le 15 janvier 2018, consulté le 08 septembre 2020. URL : <http://journals.openedition.org/corpus/2844>

Ce document a été généré automatiquement le 8 septembre 2020.

© Tous droits réservés

Segments phraséologiques et séquences textuelles

Questions méthodologiques et approches topologiques

Marion Bendinelli

- 1 Cette livraison¹ de *Corpus* propose de s'intéresser aux faits phraséologiques, c'est-à-dire à cet ensemble de « prêts-à-dire », de « déjà-là », de formules « clef-en-main » aux fonctions communicationnelles variées : zones de confort pour les uns, les faits phraséologiques peuvent faciliter la prise de parole ou les tâches rédactionnelles (Née, Sitri, Veniard 2014 ; Cislaru, Sitri, Pugnière-Saavedra 2013), contribuer à marquer une identité sociale ou professionnelle (Römer 2010 ; Née, Sitri, Veniard 2014), véhiculer une posture idéologique (Cheng 2007). Cauchemars pour d'autres, ils posent de véritables problèmes d'apprentissage, y compris dans le cas de langues génétiquement apparentées, ou témoignent de la dérive de certains discours, notamment dans le cas des discours politico-médiatiques où les expressions toutes faites, répétées à l'envi, paraissent tels des « tunnels linguistiques dans lesquels la créativité du locuteur recule au profit d'une forme de récitation [jusqu'à former un] discours creux où les formules les plus lourdes sont les moins chargées de sens » (Mayaffre 2007 : 9-10).
- 2 L'étiquette de « faits phraséologiques » permet d'englober plus de soixante catégories opératoires (Wray 2002, *in* Whal, Gries à par.) dont les désignations, diverses, reflètent la variété des courants théoriques, périodes et propriétés définitoires retenues. En guise d'inventaire non exhaustif, on retiendra : unités phraséologiques, phraséologies ou phrasèmes, *phrases* (au sens de « manières de dire, expressions ») semi-préconstruites (Sinclair 1991), expressions figées, collocations (de nature lexicale ou grammaticale - Firth 1957, Halliday 1961 notamment), *lexical extended units* (unités lexicales étendues - Sinclair 2004), *formulaic expressions* (formules, séquences conventionnelles - Biber 2009), discours figé (Branca-Rosoff 1990), *lexical bundles* (paquets lexicaux - Biber, Johansson, Leech Conrad, Finegan 1999 ; Biber, Conrad, Cortes 2004), cadres collocationnels (Renouf, Sinclair 1991), *phrase-frame/p-frame* (Römer 2010), colligations (Firth 1968 dans son acception originale ; Hoey 2005 pour une acception plus moderne, distinguant les types lexical et textuel), (*grammar*) *patterns*

(ou (grammaire des) patrons - Hunston & Francis 2000), constructions (Fillmore 1988 ; Bouveret & Legallois 2012 pour le français), collostructions (Gries 2009), n-grams/n-grammes, segments répétés (Salem 1986) et quasi-segments répétés (Bécue, Peiro 1993), séquences textuelles (Salem 2006), motifs (Longrée, Mellet 2012), routines discursives (Née, Sitri, Veniard 2014), matrices lexicales (Anscombe 2011), constructions pré-formées (Schmale 2013), prosodie sémantique (Sinclair 1991), *multi-word expressions* (expressions polylexicales - Wahl, Gries 2008, à par.), polycooccurents (Martinez 2012 ; Fleury, Lefeuvre, Pirès 2012).

- 3 Ces dénominations² non-synonymiques ont en commun de postuler l'existence de structures pré-formées combinant deux (ou plus) unités. Notons ici que, si la longueur est théoriquement infinie, il apparaît que, dans la pratique, les chercheur-e-s se concentrent sur les unités comprenant entre deux et cinq constituants. André Salem favorise quant à lui le « recensement des segments répétés composés d'au moins 4 formes [afin de sélectionner] des séquences susceptibles de correspondre à des constructions syntaxiques un peu complexes éliminant du même coup des segments très fréquents » (Salem 2006 : 843) qu'il estime, pour sa part, moins intéressants en raison d'une perspective de recherche intrinsèquement discursive et politique.
- 4 Les items constitutifs de faits phraséologiques ont pour caractéristique commune d'être unis par une relation orientée : en effet, il existe une hiérarchie entre les unités « qui fait qu'en formulant, le locuteur progresse en partant de la base pour arriver au collocatif et non inversement » (Hausmann, Blumenthal 2006), les auteurs de mentionner la relation unissant *endurci* à *célibataire* ; si l'on formule ceci en termes mathématiques, on dira, à la suite de Stephen Th. Gries que la probabilité « $p(\text{mot}_1 | \text{mot}_2)$ n'est pas la même que $p(\text{mot}_2 | \text{mot}_1)$ » (2015b : 139)³, postulat illustré par la relation unissant *of* à *in spite*. Cette propriété distingue les faits phraséologiques des simples cooccurrences récurrentes.
- 5 En revanche, les différents items se différencient notamment par le fait
  - (i) d'être constitués d'unités appartenant ou non au même niveau linguistique (lexical, syntaxique, grammatical, sémantique, discursif, prosodique, multimodal) ;
  - (ii) d'être pourvus ou non d'une autonomie sémantique, syntaxique, référentielle ;
  - (iii) d'être détectables par application de critères linguistiques, de mesures de fréquence ou de tests statistiques⁴ ; dans ces deux derniers cas, la cooccurrence (statistique) d'items participe de la définition des faits phraséologiques et contribuent à les éclairer⁵ ;
  - (iv) de tolérer ou non la variation en leur sein (choix étendu ou restreint d'items, possibilité de non-réalisation d'un ou plusieurs items).
- 6 Le sens des structures pré-formées peut varier le long d'un continuum allant de la non-compositionnalité (i. e., le sens n'est pas déductible de ses constituants) à la compositionnalité (laquelle n'est cependant pas synonyme de transparence) et leur actualisation peut être sujette à variations lexicales et/ou morphosyntaxiques. Le caractère préconstruit de ces faits phraséologiques repose sur des mécanismes inscrits en langue (aux niveaux phonologique, prosodique, sémantique, morphosyntaxique, énonciatif) et/ou en discours : enchaînements d'ordre discursif, pragmatique, voire structurations informationnelle (des énoncés précédents) et cohésive (du texte). Les faits phraséologiques peuvent ainsi être classés le long d'un continuum allant des unités les plus langagières (donc inscrites en langue), comme les *patterns* syntaxiques

(ex. *phrasal verbs*) ou les réalisations lexicales figées (expressions, locutions...), aux plus discursives, comme les motifs et routines⁶.

- 7 Ce sont ainsi six paramètres que Gries (2008 : 4, 2015b : 136-137) identifie pour définir les faits phraséologiques :

- la nature des items	- la distance admise entre les constituants
- le nombre d'items	- le degré de flexibilité lexicale et syntaxique dont bénéficie l'unité considérée comme phraséologique
- le nombre de fois où une unité doit être employée pour être considérée comme phraséologique	- le degré de transparence / compositionnalité sémantique de l'unité phraséologique par rapport à ses constituants

- 8 Nous choisissons pour notre part de nous intéresser aux faits phraséologiques (i) pourvus d'un signifié, (ii) définis et/ou repérés au moyen d'un ou plusieurs calculs statistiques (iii) au sein d'empans textuels non limités au cotexte immédiat et à la phrase ; ces faits existent en corpus, caractérisent un genre discursif et/ou un type de textes et présentent une fréquence remarquable, car inattendue (sur- ou sous-employée par rapport à une distribution aléatoire). Nous nommons ces faits « segments » ou « segments phraséologiques ».

1. Étude des faits phraséologiques : « actualité » de la recherche

- 9 La recherche sur les faits phraséologiques est actuellement particulièrement féconde en France et à l'étranger, et se présente comme « un objectif de plus en plus important pour les chercheurs issus de diverses disciplines » (Wahl, Gries à par. : 17)⁷ ; un rapide – et non exhaustif – recensement des initiatives menées ou en cours depuis 2015 en témoigne : pas moins de neuf événements scientifiques (colloques, journées d'étude, école d'été, *workshop*) et six publications lui sont consacrés (liste en annexe 2).
- 10 Si la recherche phraséologique se révèle donc d'actualité, il ne s'agit pour autant pas d'un objet nouveau⁸ : Ferdinand de Saussure en son temps note que « [d]ans la règle, nous ne parlons pas par signes isolés, mais par groupes de signes, par masses organisées qui sont elles-mêmes des signes » (cité par Hausmann, Blumenthal 2006 : 10) ; dans les **années 1920**, Charles Bally met en évidence l'existence de séquences nommées séries phraséologiques, quand, dans les **années 1960**, Bernard Pottier identifie des lexies, Émile Benveniste parle de synapsies, André Martinet de synthèmes et Eugène Coseriu de clichés lexicaux. Les distributionnalistes et contextualistes anglo-saxons inaugurent dans les **années 1950 et 1960**, par le décloisonnement du lexique, de la syntaxe et de la sémantique d'une part, et l'avènement d'une vision probabiliste sur la langue d'autre part, la recherche sur les collocations, ouvrant alors la voie aux travaux de John Sinclair sur le principe idiomatique (*idiom principle*) et de Maurice Gross sur le lexique-grammaire (néanmoins sans perspective statistique ici). La linguistique cognitive, qui se dessine à la suite de Ronald Langacker dans **années 1980**, fait sienne le rejet d'une séparation stricte entre lexique et syntaxe, et annonce les

travaux portant sur les *patterns* puis les constructions, notion amenée à englober celle d'unité phraséologique. Du côté des approches outillées, dans les **années 1970 et 1980**, l'école de Saint Cloud et la lexicométrie mettent au cœur de leurs analyses des textes politiques les phénomènes de répétition et de récurrence, lesquelles sont au fondement de la recherche sur les cooccurences, associations privilégiées ou poly-cooccurences. Le traitement automatique des langues, enfin, se saisit dès ses origines dans les **années 1960**, de questions d'ordre phraséologique en s'interrogeant sur la description de groupes de formes graphiques alors identifiés comme des unités lexicales complexes.

- 11 Au même titre que les faits de langue et les faits de discours, les faits phraséologiques sont une unité fondamentale pour l'analyse de productions empiriques, en ce qu'ils participent à la **construction du sens**⁹. Frédérique Sitri et Agnès Tutin rappellent que

l'analyse de discours est attentive depuis ses origines aux phénomènes de répétition et de redondance car ils constituent un mode d'accès à des éléments préconstruits, « déjà-là », qu'elle cherche à mettre en relation avec plusieurs niveaux de détermination du discours. Ainsi [le Laboratoire de Saint-Cloud dans les années 1980 met] en évidence, à partir des inventaires de segments répétés et des concordances, plusieurs niveaux de figement : figements en langue, mais aussi figements propres au genre, et figements propres au domaine (ou secteur d'activité pour reprendre le terme de Bakhtine), comme le discours syndical [...]. (Sitri, Tutin 2016 : 11-12)

- 12 Pour l'école contextualiste, dont les figures tutélaires sont John R. Firth, Zellig S. Harris, et leurs élèves et héritiers John Sinclair et Michael A.K. Halliday, les unités phraséologiques seraient d'ailleurs la seule véritable unité utile à la description du sens tel qu'il se développe dans un **texte**...

le sens est normalement porté par une expression et non par un mot ; le mot n'est qu'un cas particulier d'expression, et n'a pas d'autre statut dans la description du sens [...]. L'un des grands atouts de l'approche phraséologique est de préserver l'intégrité du texte, bien plus que ne le font d'autres approches descriptives, et cela assure en retour la préservation du sens. (Sinclair 2008 in Römer 2010 : 96)¹⁰

- 13 ... et dans une **langue**, ou plus précisément *a restricted language* (Firth 1959)¹¹, c'est-à-dire un sous-ensemble d'une langue. En effet, les moyens expressifs variant entre genres discursifs et types de textes (cf. rappel ci-après des travaux sur la dimension phraséologique des discours et des genres), il apparaît comme particulièrement difficile, sinon impossible et dans tous les cas non pertinent, de lister l'ensemble des possibilités lexico-grammaticales d'un idiome. L'école contextualiste s'intéresse ainsi très tôt à la description de pratiques sociales et discursives situées, rapprochant les *restricted languages* des registres et discours. Un tel postulat justifie pleinement la place centrale qu'occupe la recherche phraséologique dans les champs de la **didactique des langues**, de la **traduction**, de la **terminologie**, en particulier lorsque ces champs sont appliqués à des pratiques professionnelles ou des discours de spécialité. On trouvera des exemples de cette approche en didactique de la phraséologie dans le dernier numéro des *Cahiers de lexicologie* (n° 108) coordonné par Olivier Kraif et Agnès Tutin, et en particulier aux contributions traitant des discours scientifique et économique. S'inscrivent également dans cette perspective, les publications s'interrogeant sur la structure argumentative de débats politiques (Bendinelli 2014, Rouveyrol 2014), sur l'établissement de vocabulaires fondamentaux (Benigno, Kraif, Grossmann, Velez 2016) ou encore sur l'évaluation de la qualité de travaux d'étudiants dans un contexte universitaire (Römer, O'Donnell 2009).

- 14 Différents travaux ont montré enfin que les faits phraséologiques participent de l'**intertextualité**, par la circulation de structures figées entre discours et formations sociales – voir à ce sujet la notion de formule développée par Alice Krieg-Planque, ou bien les travaux sur les collocations de Jean-Michel Adam pour qui
- [I]es associations lexicales établies dans un texte peuvent être tellement fortes qu'elles se stabilisent mémoriellement et deviennent les supports de connexions intertextuelles. [...] Ce type de figement de collocations lexicales qui circulent dans l'interdiscours de l'état donné d'une formation sociale est très productif. (Adam 2011b : 102)
- 15 Une voie récemment explorée dans la sphère francophone mais déjà largement établie est le rôle caractérisant des faits phraséologiques dans la **définition et l'identification des genres discursifs**¹². On trouvera des exemples de ces travaux dans le numéro 108 des *Cahiers de lexicologie* susmentionné, ainsi que dans le dernier numéro de la revue *Lidil* (n° 53) coordonné par Frédérique Sitri et Agnès Tutin. Ainsi, diverses contributions envisagent les unités phraséologiques (parmi lesquelles « patrons, motifs, routines » pour citer le sous-titre du dossier thématique), comme des marques de genres discursifs – discours politique : allocutions (Pražuch 2016), vœux présidentiels (Leblanc 2016) ; discours professionnels : compte-rendu de bilans orthophonistes (Brin-Henry, Knittel 2016), écrits professionnels (Née, Sitri, Veniard 2016) ; discours littéraire : roman contemporain (Siepmann 2016), romans sentimentaux (Legallois, Charnois, Poibeu 2016), correspondances privées (Branca-Rosoff 1990) – pour ne citer que quelques exemples. Douglas Biber avait déjà tracé cette voie dès la fin des années 1980 dans le contexte anglo-saxon, identifiant pour divers textes et registres (écrits académiques, écrits journalistiques, conversations ordinaires, écrits de fiction...), des configurations de catégories et de traits lexico-syntaxiques et sémantiques, des *phrases* et des blocs lexicaux (*lexical bundles*) statistiquement pertinents (Biber 1988, 2006 ; Biber, Johansson, Leech, Conrad, Finegan 1999 ; Biber, Conrad, Cortes 2004).
- 16 Une nouvelle voie semble se dessiner, ou s'affirmer, dans la continuité des travaux de Née, Sitri, Veniard (2014, 2016) où la question des routines est envisagée d'un point de vue sociolinguistique, faisant de ces unités des **marqueurs d'identité sociale, professionnelle ou discursive** : c'est en tant qu'unité opératoire de l'analyse du discours que ces auteures entendent positionner la notion de routine. Une question serait de savoir si tous les schèmes lexico-grammaticaux se prêtent à une telle fonction, ou si la définition en soi de routines implique cet usage.
- 17 Les faits phraséologiques, on l'aura compris, couvrent des phénomènes intéressants de nombreux champs de recherche : linguistique théorique et descriptive, psycholinguistique, linguistique cognitive, linguistique contrastive, lexicographie, didactique des langues, analyse du discours, linguistique de corpus, linguistique textuelle. Ce sont ces deux dernières perspectives que nous souhaitons mettre en lumière dans le présent numéro de *Corpus*. En adoptant ainsi une perspective explicitement textuelle, nous interrogerons également les implications théoriques et méthodologiques que la recherche phraséologique peut avoir pour la constitution et l'exploration de corpus. Le texte, que François Rastier définit comme une « suite linguistique empirique attestée, produite dans une pratique sociale déterminée, et fixée sur un support quelconque » (1996 [en ligne]), est le produit d'une structuration linéaire (plan de texte et séquences) et non linéaire (réticulaire : réseaux de cooccurrents, isotopies ; configuration : de dimensions thématique et illocutoire)¹³. Le texte, parce qu'il « se rattache à la langue par un discours et à un discours par la

médiation d'un genre » (Rastier 2001 : 230) nous donne à observer les unités phraséologiques telles qu'elles se réalisent sous l'influence d'une pratique socio-discursive et d'un dispositif communicationnel, dans un cadre interactionnel particulier : moins que les principes sous-jacents situés en langue et rendant possibles ces réalisations, c'est une **perspective textuelle** que nous adoptons donc ici, laquelle nécessite d'inscrire l'étude **en corpus**. Quelques commentaires s'imposent : c'est l'objet de la section qui suit.

2. Faits phraséologiques : rapport(s) au corpus et au texte

2.1 Unités phraséologiques et corpus textuels

- 18 À la suite de Rastier (2005 : 31), nous retenons que « [l]e texte est pour une linguistique évoluée l'unité minimale, et le corpus l'ensemble dans lequel cette unité prend son sens » ; à ce titre, c'est bien en corpus que vont pouvoir s'observer, se repérer, s'analyser les faits phraséologiques. Ces corpus recueilleront, dans le respect des principes formulés en linguistique de corpus (homogénéité, hétérogénéité, clôture, réflexivité, représentativité), des productions authentiques que les contributions ici rassemblées cherchent à décrire sur les plans microtextuel – description locale de la matérialité langagière – ou macrotextuel – description de la structuration linéaire et non-linéaire de cette matérialité dans ses rapports au genre et au discours dont les textes sont représentatifs. John Sinclair estime que les faits phraséologiques sont « le point de contact idéal entre un corpus et une description, car ils se satisfont des phénomènes de surface, ce qui est, à l'origine, ce qu'un corpus fournit : aucun prétraitement n'est requis, aucune abstraction, aucune information telle que les parties du discours n'est ajoutée » (2008, *in* Römer 2010 : 96)¹⁴. Le linguiste britannique limite ici l'unité phraséologique aux collocations, suite ordonnée et contiguë de deux formes lexicales ou plus, envisagées sous leurs seules formes graphiques, que l'on pourrait également désigner sous le nom de paquets lexicaux (Biber *et al.* 1999), n-grams ou segments répétés (Salem 1987) avec, dans ces deux derniers cas, un repérage par « une fréquence remarquable, évaluée en probabilité » (Viprey, Mayaffre 2012 [en ligne]). Mais ces collocations, n-grams ou segments répétés présentent l'inconvénient d'être figés, donc de ne pas tolérer l'ajout ou la suppression d'éléments lexicaux et de nier la variation morpho-syntaxique¹⁵ (Longrée, Mellet 2015 ; Leblanc 2016). Les études phraséologiques de cette dernière décennie se sont affranchies de ces contraintes strictement lexicale et syntaxique pour faire place à des unités qui, à divers degrés, se présentent comme multi-dimensionnelles (association d'items lexicaux, grammaticaux, prosodiques...) et à réalisations souples (admettant donc une certaine variation formelle et l'existence d'items facultatifs) : en ce sens, on trouvera les collostructions (Gries 2009), motifs (Longrée, Mellet 2013, 2015) ou arbres lexico-syntaxiques récurrents (Kraif, Tutin 2016 ; Kraif 2016 ; Kraif, Novakova, Sorba 2016). Le repérage de ces unités phraséologiques impose d'interroger le corpus par moissonnage automatisé ou par des requêtes alliant items lexicaux, items ou traits grammaticaux, catégories ou traits sémantiques et/ou relations syntaxiques : les corpus explorés sont, on le comprend, nécessairement annotés, enrichis d'informations morphosyntaxiques, parfois également sémantiques. La finesse de l'annotation dépend toutefois des objectifs de recherche, des outils d'exploration utilisés, de l'unité phraséologique

recherchée ; elle pourra se faire manuellement, semi-automatiquement ou automatiquement, en ayant recours à des étiqueteurs éprouvés¹⁶.

- 19 Par ailleurs, les publications les plus récentes, dans les deux espaces scientifiques, témoignent d'une réflexion grandissante et d'un intérêt pour la description de méthodologies intégralement *corpus-driven*, c'est-à-dire ascendantes, inductives¹⁷ : c'est par le traitement statistique des données réunies en corpus que les segments phraséologiques doivent émerger, sans *inputs* préalablement fournis à l'analyse, sans sélection préalable des items autour desquels construire les requêtes. Quiniou, Cellier, Charnois, Legallois (2012) rappellent que les travaux de Biber (2006) articulant relevés de paquets lexicaux et cadres collocationnels sont à cet égard les premiers à avoir adopté une approche pleinement inductive, bien qu'elle impose une phase de tri manuelle servant à distinguer les items facultatifs des items obligatoires ; l'approche développée par Née, Sitri, Veniard (2014, 2016) ou par Kraif (2016) s'inscrit dans cette filiation. L'expertise acquise en traitement automatique des langues et en *data mining* semble devancer, chronologiquement. Ainsi, l'approche par *patterns* séquentiels fréquents présentée par Quiniou, Cellier, Charnois, Legallois (2012) et formulée originellement au milieu des années 1990¹⁸ permet d'extraire des unités similaires aux cadres collocationnels sans aucune intervention manuelle.

2.2 Unités phraséologiques, textualité et topologie textuelle

- 20 Notre objectif est de mettre en lumière **le(s) rôle(s) caractérisant(s) et/ou structurant(s)** des segments au sein des textes, en envisageant notamment leur contribution éventuelle à la construction de séquences textuelles, définies par Jean-Michel Adam comme

des unités textuelles complexes, composées d'un nombre limité de paquets de propositions-énoncés : les macro-propositions. Ces macro-propositions sont des sortes de périodes dont la propriété principale est d'être des unités liées à d'autres macro-propositions, occupant des positions précises au sein du tout ordonné de la séquence. Chaque macro-proposition prend son sens par rapport aux autres, dans l'unité hiérarchique complexe de la séquence. [...] À la différence des simples périodes, les macro-propositions qui entrent dans la composition d'une séquence relèvent d'agencements pré-formatés de propositions. Ces différents agencements sont dits « narratif », « argumentatif », « explicatif », « dialogal » et « descriptif ». (Adam 2011b : 131)

- 21 Sans que la question du lien, voire d'une éventuelle corrélation entre segments, ou plus largement faits phraséologiques, et séquences textuelles n'ait été ouvertement posée, plusieurs études mentionnent dans leurs analyses avoir repéré des *patterns*, des propositions, des configurations de traits lexico-morphosyntaxiques, jouant une fonction textuelle-discursive. Citons à cet égard les travaux de Jean-Michel Adam pour qui la description des macro-propositions et de leurs enchaînements au sein de séquences peut se présenter comme une mise au jour non explicitée de patrons lexico-syntaxiques ordonnés. Ainsi, le linguiste s'est-il par exemple aventuré, sans pour autant le formuler explicitement, vers la reconnaissance de schèmes prototypiques au fondement de la **séquence explicative** : « [l']explication peut apparaître dans de courts segments. Ainsi la forme de phrase périodique au présent qui combine SI (introduceur d'une proposition qui pose problème) avec C'EST QUE ou C'EST PARCE QUE, introduceurs d'une explication » (Adam 2008 : 153 ; voir également 2011b : 153-158

pour une complexification de ce patron) ; ou d'un **genre discursif comme l'insulte rituelle et l'histoire drôle** fondé sur les propositions *p* et *q* telles que :

Type 1
Proposition <i>p</i> : X être/avoir... SI/TELLEMENT [a] >> QUE >> proposition <i>q</i> <i>Proposition p comportant un intensif [a] >> corrélation de conséquence >> proposition q</i>

(Adam 2011 : 118)

- 22 C'est également le cas des travaux de Fleury, Lefevre, Pirès (2012) qui, décrivant le mot « quoi » dans le Corpus du français parlé parisien des années 2000, identifient le « patron syntaxique régulier [...] c'est + vraiment + attribut + quoi » (2012 : 9) et lui attribuent une **fonction de (re)formulation**. Les travaux de Née, Sitri, Veniard (2014) s'inscrivent dans une démarche similaire, repérant « le patron « être dans », qui [participe] de la **visée descriptive/évaluative** des rapports éducatifs » (2016 : sp) et constatant qu'un fonctionnement discursif peut « s'actualiser plus particulièrement dans une « partie » du texte, participant ainsi à la construction du texte » (Née, Sitri, Veniard 2016 : sp).
- 23 Les études littéraires fournissent également des indices sérieux quant à la forte contribution des segments à l'organisation textuelle : ainsi Dirk Siepmann constate-t-il après étude d'un corpus de romans contemporains en français et en anglais,

[p]lus un segment répété est long et plus il correspondra à une unité fonctionnelle (un patron narratif) sous-jacent. Un patron narratif peut-être défini, dans une première approche, comme une co-occurrence (collocation, séquence figée, motif...) récurrente dans un corpus de textes narratifs et ayant une fonction narrative (et/ou descriptive) spécifique. (2016 : 28)
- 24 Et D. Siepmann de citer les **patrons narratifs** permettant d'exprimer l'espace, le temps, la rencontre entre deux personnes, l'acquiescement. Dans le même ordre d'idée, on pourra se référer aux travaux de Legallois, Charnois, Poibeu (2016 : np) où l'« expression *la première fois*, statistiquement sur-employée dans le sous-genre [des romans sentimentaux], joue le rôle « narratif », dans des configurations syntaxiques différentes, d'exprimer une époque révolue, et l'avènement d'une nouvelle situation importante dans l'histoire ».
- 25 Les **séquences descriptives, narratives et argumentatives** font également l'objet de commentaires de la part de Laurent Rouveyrol. L'auteur constate, à partir de l'étude d'un corpus de débats politiques médiatisés britanniques, et sans pour autant inscrire son travail dans le champ de la phraséologie, que les premières « sont souvent introduites par des verbes de perception comme « I saw », « we've seen » ou bien des expressions introduisant un exemple, une illustration (« for instance ») ; que les deuxièmes s'organisent autour d'« un glissement temporel [opéré] après l'embrayeur *now* et trouvent leur place au sein d'une séquence argumentative plus large marquée par l'alternance *présent / prétérit / présent* « correspond[a]nt exactement au découpage [...] (conclusion 1 / narration / conclusion 2) » (2014 [en ligne]). Née, Sitri, Veniard (2016 : sp – nous soulignons) identifient quant à elles, à partir de leur exploration outillée d'un corpus d'écrits professionnels, le patron « JUGEMENT D'ÉTAT – CAPACITÉ et/ou ÉVÉNEMENT NOTABLE [dont l]a principale fonction [...] est la preuve par l'exemple, ce qui fait de lui une **routine argumentative** ».

- 26 Du côté anglo-saxon, on pourra citer les travaux de Michael Hoey et sa théorie du *priming* lexical qui considère que les unités lexicales, outre leur préférence co-textuelle lexico-syntaxique (collocations, cooccurents, construction, collostruction, colligation, prosodie sémantique...), ont également une préférence pour certaines positions dans les textes (début ou fin de textes, de paragraphes, de phrases), selon le principe suivant : « chaque mot fait l'objet d'une amorce le portant à apparaître ou à éviter certaines positions du discours » (Hoey 2005 : 13 in O'Donnell, Scott, Mahlberg, Hoey 2012 : 73-4)¹⁹. C'est ce que M. Hoey nomme colligation textuelle, et dont on trouvera une illustration et une application chez Römer (2010) pour l'étude de la structure des **critiques d'ouvrages scientifiques**, dans Römer, O'Donnell (2009) pour la **structure des productions écrites d'étudiants**, ou encore dans O'Donnell, Scott, Mahlberg, Hoey (2012) autour de la question de la **structure des articles de presse** et des **routines journalistiques**. Cet axe de recherche se retrouve en France chez Gledhill (2000) et Ho-Dac, Péry-Woodley (2008) pour l'étude de la **structure des articles scientifiques**, ou encore chez Carter-Thomas (2014)²⁰ pour celle d'un corpus d'**articles de presse économique**, étude publiée dans le numéro 13 de la présente revue. Ces trois dernières contributions ont en commun de se réclamer de la linguistique hallidayenne, et plus généralement de la linguistique de corpus ; à titre d'exemple, Ho-Dac et Péry-Woodley mettent au jour de manière semi-automatique des configurations d'items lexico-grammaticaux en des positions textuelles précises, concluant ainsi à leur fonction de signaux de l'organisation discursive : « les expressions sujets ou les adverbiaux détachés en position initiale de phrase expriment ce à propos de quoi on parle [et l']organisent en participant à la répartition des contenus dans des segments de texte » (2008 : 739). Parmi les travaux cités ici, deux approches de la structuration des textes sont adoptées. Les travaux anglo-saxons déterminent en amont de l'extraction phraséologique la structure de leurs textes, et encodent ces empan. Ceux-ci peuvent être définis sur la base de critères de taille, de fonctions textuelles préalablement identifiées, d'organisation visuelle du texte et du paratexte (titre, sous-titre, paragraphe initial ou non...)²¹. L'expérience menée par Ho-Dac et Péry-Woodley se fonde, elle, sur le repérage de configurations lexico-grammaticales récurrentes pour saisir l'organisation textuelle. Leur approche est donc inductive, les faits phraséologiques servent l'identification de la structure du texte, tandis que les travaux anglo-saxons s'inscrivent dans une démarche déductive, les faits phraséologiques éclairant et/ou confirmant l'existence de structures préalablement annotées.
- 27 Il ressort de ces quelques exemples d'études publiées ces dernières années un intérêt grandissant, et justifié, pour la structuration séquentielle des textes et sa saisie au moyen de phénomènes phraséologiques de nature lexicale, morphosyntaxique, sémantique ou hybride. C'est bien par le repérage de co-apparitions statistiquement remarquables d'items linguistiques dans un empan textuel donné que l'on peut faire émerger du corpus, outre des segments phraséologiques, des empan porteurs d'une fonction textuelle et/ou participant à l'organisation séquentielle du texte. La question de la position des items dans l'espace textuel est donc doublement mis sous les projecteurs : non seulement les chercheurs (ré)interrogent la question de l'empan dans lequel observer les phénomènes de cooccurrence, mais ils inaugurent un usage renouvelé des outils permettant la visualisation des positions des items les uns par rapport aux autres d'une part, et au sein du texte d'autre part. **C'est la notion de topologie textuelle qui s'en trouve ainsi (ré)ouverte.**

- 28 Rappelons que la topologie textuelle rend compte du fait que tout texte prend place au sein d'une chaîne (spatiale, temporelle, thématique...) et se définit lui-même comme une trame, c'est-à-dire un enchaînement d'unités linguistiques dépendantes, imbriquées (Mellet, Barthélémy 2007 : 3-4). Cela revient donc à considérer le texte comme une structure ordonnée, « un espace ou un plan sur lequel [d]es données s'enchaînent (plus que s'additionnent) et s'organisent » (Mayaffre 2007 : 3). Empruntant aux mathématiques une conception et une définition de l'espace et des ensembles (permettant notamment d'aborder la question du voisinage, chère également aux études en linguistique de corpus), la topologie textuelle développe dès sa première formulation des outils de visualisation désignés sous le nom de topographie (ou cartographie) textuelle. Cette dernière a notamment été travaillée et exposée dans Lamalle, Salem (2002) et Salem (2006) dans le cadre du développement et de l'utilisation du logiciel *Lexico3*, appliquée dans les deux cas à l'étude du discours syndical et à la distribution de types²² et de segments répétés dans les sections²³ du texte. C'est cette méthodologie qui est exploitée dans le cadre des études de corpus multilingues de Zimina (2005, 2014) et de Fleury, Zimina (2004) utilisant le logiciel *Le Trameur* : les *patterns* lexico-grammaticaux mis en relation par la comparaison de leur distribution cartographique permet, dans le cas de cartographies proches voire similaires, de les considérer comme des équivalents de traduction, et dans le cas contraire de rejeter cette hypothèse.
- 29 Bien que l'analyse de données textuelles outillée ait montré toute sa puissance dans la saisie et la cartographie des phénomènes textuels de nature micro-structurale et macro-structurale, il apparaît que la visualisation des segments, et plus largement des faits phraséologiques, n'a jusqu'à présent été envisagée que de manière marginale. L'identification et la visualisation des segments se font essentiellement à partir de listes de concordances obtenues sur requêtes lexicales (forme ou lemme), grammaticales (code grammatical) ou mixtes, ou bien à partir de listes de segments répétés. Nous aimerions introduire ici trois autres modes d'extraction et/ou représentation : les graphes de cooccurents et polycooccurents, les arbres lexico-syntaxiques récurrents et la projection sur carte des sections d'un corpus.
- 30 Étienne Brunet mentionnait déjà il y a une dizaine d'années que les **graphes de cooccurents** générés par le logiciel *Hyperbase* rendaient possible la visualisation de phénomènes phraséologiques : « quand un mot a peu de liaisons, cela signifie qu'il est attaché au mot-pôle par un lien quasi exclusif qui le rend indifférent au reste et relève de la phraséologie » (2007 : 16). Ce cas était illustré par une analyse des cooccurents identifiés dans l'œuvre de Stendhal (Brunet 2007) ; la figure ci-dessous reproduit un cadran du graphe (coin nord-est) généré à partir de la forme-requête *mort*.

Figure 1. Extrait du graphe de cooccurrent de la forme graphique *mort* dans le corpus Stendhal (Brunet 2007)

- 31 Pour É. Brunet, le lien unique unissant le requêté *mort* à *félix* contraste avec le nombre multiple de liens faisant de *sentence* un nœud entre *mort*, *exécution*, *grégoire*, *terminer*... : ceci serait l'indice d'un usage de type phraséologique entre *félix* et *mort* dans le corpus - usage qui, en l'absence d'illustrations, doit être précisé par un retour au texte. À cet égard, les **graphes de poly-cooccurrents** tels que ceux obtenus au moyen des logiciels *Lexico3* et *Alceste* fournissent une lecture plus immédiate des segments phraséologiques. Par une schématisation similaire aux grammaires locales, ils rendent compte de l'environnement du requêté au-delà de la cooccurrence binaire, ainsi que de l'orientation de ses attractions lexicales potentiellement inter-reliées. Une application de cet outil peut se trouver dans Fleury, Lefeuve, Pirès (2012), où, pour la requête *quoi*, les auteurs obtiennent le graphe reproduit en figure 2 dont la lecture met notamment en évidence deux réalisations du type *quoi* → *mm* → *peu* → *un* → *vraiment* et *quoi* → *mm* → *enfin* → *vraiment*.

Figure 2. Graphe de poly-cooccurrents de la forme *quoi* dans le corpus du français parlé parisien des années 2000 (Fleury, Lefeuve, Pirès 2012)

- 32 Ces poly-cooccurrents sont à même de reconstruire les *patterns* lexico-pragmatiques dans lesquels *quoi* est employé – *quoi*+marque d'hésitation ou de reformulation (*mm*), *quoi*+modalisation (*vraiment*, *enfin*) – ainsi qu'un « patron syntaxique régulier en *c'est* +

vraiment + attribut + quoi » (Fleury, Lefeuvre, Pirès 2012 : 6-8) – *patterns* que nous nommerions pour notre part segments phraséologiques.

- 33 Les **arbres lexico-syntaxiques récurrents (ALR)** sont une troisième possibilité de saisie et représentation des faits phraséologiques. Ces arbres s'inscrivent dans la continuité des travaux portant sur la notion de cooccurrence syntaxique empruntée à Stefan Evert « caractéris[ant] une association statistique significative reliant deux mots par une relation syntaxique » (Tutin, Kraif 2016 : sp) ; les ALR rendent ainsi compte des dépendances hiérarchiques de nature syntaxique à l'œuvre dans un corpus (préalablement traité par un analyseur syntaxique idoine), identifiées dès lors que leur fréquence d'apparition dans le corpus atteint et/ou dépasse un seuil de significativité statistique. Générés par le *Lexicoscope*, outil d'exploration développé par Olivier Kraif et Sacha Diwersy (utilisé dans ce numéro), ces arbres se présentent sous la forme d'un graphe tel que celui reproduit ci-dessous. L'extraction itérative d'arbres récurrents menée ici à partir du pivot *constater* dans un corpus d'articles scientifiques aboutit à la saisie des diverses réalisations lexico-syntaxiques dans lesquelles le verbe apparaît ; la figure 3 représente l'une de ces réalisations, le segment phraséologique *il est frappant de constater*.

Figure 3. Arbre lexico-syntaxique récurrent du verbe pivot *constater* dans un corpus d'articles scientifiques (Tutin, Kraif 2016)

- 34 Enfin, on mentionnera le cas des cartographies de section proposées par le logiciel *Lexico 3* qui permettent, elles, de consulter l'existence/l'absence de segments requêtés au sein des sections du texte, ainsi que leur fréquence et leur répartition dans l'espace du texte. Cet espace est segmenté en sections représentées par des carrés : ceux-ci sont colorés ou non selon la présence ou la prégnance statistique (au choix) de l'item recherché en leur sein. Dans la figure produite ci-après, on observe la localisation des segments répétés de longueur ≥ 4 dans le texte du *Manifeste du Parti Communiste* de 1848

(Salem 2006 : 847), et plus particulièrement leur concentration dans des espaces – voire paragraphes – spécifiques.

Figure 4. Localisation des segments répétés de longueur ≥ 4 dans le texte du *Manifeste du Parti Communiste* de 1848 (Salem 2006)

3. Phraséologie et méthodes quantitatives

- 35 Bien que la recherche phraséologique ne soit pas récente (nous l'avons rappelé en début de cette introduction), et que les travaux l'abordant d'un point de vue quantitatif, statistique, fassent l'objet de publications depuis lors, on ne peut que constater à la suite de Gries (2015b) l'absence de consensus autour des questions méthodologiques que ces approches soulèvent : comment repérer et extraire les segments phraséologiques ? Quels sont les mesures, les algorithmes les plus pertinents, utiles, éprouvés, pour quantifier la force de l'association entre ses constituants, la flexibilité morphosyntaxique, la fréquence d'emploi... Quelle proportion de bruit (ou faux résultats) est générée par une extraction automatique ? Comment la limiter, quel(s) filtre(s) utiliser ?
- 36 Avant de commenter plus avant ces points et de présenter quelques-unes des mesures statistiques existantes, constatons tout d'abord que la question sémantique – c'est-à-dire le caractère compositionnel ou non de l'unité, et sa transparence ou non – est absolument absente des réflexions quantitatives autour de leur repérage automatisé. Ce n'est que le traitement manuel et systématique de listes, générées quant à elles automatiquement, qui permet de déterminer le sens d'un segment phraséologique, à des fins de description lexicographique ou de désambiguïsation (Stefanowitsch, Gries 2003).
- 37 L'essentiel des propositions algorithmiques se fonde sur la comparaison des tableaux de contingence d'items considérés 2 à 2 (Tableau 1 empruntée à Gries 2015b : 138), et cherchent à conclure, par l'application d'un test tenant compte des fréquences observées et théoriques, à l'existence d'une association plus ou moins marquée entre les deux items : le postulat sous-jacent est que la fréquence d'emploi est corrélée au figement d'une unité en discours (puis en langue), donc plus un signe ou une structure est employée, plus elle devient conventionnelle.

Tableau 1. Tableau de contingence pour deux items (ici, mots)

	mot ₁ présent	mot ₂ absent	Totaux
mot ₁ présent	a	b	a+b
mot ₁ absent	c	d	c+d
Totaux	a+c	b+d	a+b+c+d

- 38 Les tests les plus utilisés sont le *Fisher-Yates Exact test*, le calcul du *log-likelihood ratio* (également noté ll ou G^2), le chi-carré (χ^2), la mesure de *Mutual Information*²⁴. Nous rapportons ici les conclusions de Gries (2015b), Stefanowitsch, Gries (2003) autour de la mise en œuvre et de la pertinence de ces tests statistiques – à l’exception du χ^2 , largement diffusé dans les pratiques de la linguistique de corpus francophone – pour l’étude des segments phraséologiques, et mentionnons également les expériences et comparatifs récemment publiés dans la communauté francophone : Benigno, Kraif, Grossmann, Velez (2016), Kraif, Novakova, Sorba (2016), Legallois, Charnois, Poibeau (2016)²⁵.
- 39 Le *Fisher-Yates Exact test* est jugé comme la mesure la plus pertinente pour quantifier le degré d’association entre deux items (Whal, Gries à par.), mais aussi la plus lourde à mettre en œuvre, notamment dans le cas de traitements de corpus de type itératifs.
- 40 Le *log-likelihood ratio* est considéré comme une mesure utile pour mesurer le degré d’association entre les constituants d’un segment phraséologique parce qu’il permet d’extraire correctement les unités de fréquence élevée, bien qu’il puisse se révéler trop sensible aux associations de fréquence très élevée ou de fréquence faible (Benigno, Kraif, Grossmann, Velez 2016 : 137-8).
- 41 Wahl, Gries (à par.) et Legallois, Charnois, Poibeau (2016) mentionnent enfin que la mesure de *Mutual Information*, « probablement comme la mesure d’association la plus connue » (Whal, Gries à par. : 5)²⁶, s’avère intéressante pour extraire et étudier les unités de très basse fréquence où l’attraction entre les deux constituants (ou plus) est très forte (par exemple, dans le cas des expressions figées).
- 42 Notons que les études outillées en phraséologie tendent à adopter plusieurs de ces tests, de manière à composer avec leurs limites. C’est ce que l’on observe dans les travaux francophones déjà cités : Benigno, Kraif, Grossmann, Velez (2016), Kraif, Novakova, Sorba (2016), Legallois, Charnois, Poibeau (2016).
- 43 Ces diverses mesures statistiques, si elles « reflètent l’association mutuelle entre le mot₁ et le mot₂ [elles] ne font pas la différence entre le fait de savoir si le mot₁ permet de prédire le mot₂ ou inversement » (Gries 2015b : 139)²⁷. En outre, ces mesures ne permettent pas, ou de manière peu satisfaisante, de tenir compte des cinq paramètres listés ci-après, dont certains (1, 2, 4) sont pourtant constitutifs de la définition même des faits phraséologiques (Gries 2015b) :
1. existence d’unités de longueurs différentes, avec possibilité d’emboîtement et/ou de recoupement (des séquences phraséologiques de longueur n sont incluses dans des séquences de longueur $n+1$, parfois avec superposition d’un ou plusieurs éléments, par exemple segment [abc] et segment [cde] appartenant en fait au segment [abcde]) ;
  2. direction/orientation dans l’ordre des constituants de l’unité phraséologique ;
  3. ratio *type/token* entre l’unité phraséologique et ses éventuelles différentes réalisations morphosyntaxiques ;
  4. degré d’association entre constituants d’unité dont la longueur est supérieure à 2 ;
  5. possible non-contiguïté entre constituants d’une unité phraséologique.
- 44 À cet égard, d’autres mesures statistiques doivent être utilisées pour tenir compte de ces divers paramètres. Nous en mentionnerons certaines ici, bien que brièvement, car elles ne font pas l’objet d’applications dans le présent recueil ; pour chacune, nous renvoyons aux références indiquées permettant d’approfondir la lecture.

45 Concernant la question de la longueur du segment phraséologique et des possibilités d'inclusion ou de recoupement (point 1), du ratio *type/token* (point 3) et du degré d'association pour des unités de longueur supérieure à 2 (point 4), Gries (2015c) suggère de recourir à l'une des méthodes suivantes :

- mesure de **lexical gravity** : formulée dans Daudaravičius, Murcinkevičienė (2004) et remodelée dans Gries, Mukherjee (2010), cette mesure permet de calculer « le degré selon lequel les mots tendent à apparaître au sein de n-grams [...] plutôt que de manière indépendante » (Gries 2015a : 54)²⁸, en tenant compte de la fréquence des *types* et des *token* constitutifs d'un segment phraséologique. La mesure est présentée dans Spina, Tanganelli (2012 : 76-77 en particulier), article paru dans le numéro 11 de la revue *Corpus* consacré à la cooccurrence, ainsi que dans Gries (2015c : 97-98) et Wahl et Gries (à par. : 6) ;
- établissement d'une **adjusted frequency list** : formulée par Brook O'Donnell, cette mesure ajuste les fréquences brutes d'apparition de n-grams contenus dans des séquences plus longues (et tokenisées uniquement) ; son principal inconvénient pour l'étude des segments phraséologiques est qu'elle ne permet pas d'extraire des segments présentant une fréquence basse – à l'image du calcul de *Mutual Information*. Cette mesure est présentée chez Gries (2015a : 54) et Wahl, Gries (à par. : 7) ;
- recours à l'**algorithme MERGE** (*Multi-word Expression from the Recursive Grouping of Elements*), qu'Alexander Wahl a développé pendant sa thèse co-encadrée par Stephen Gries. Cet algorithme exposé dans Wahl, Gries (à par.) est conçu pour pouvoir extraire des segments discontinus (point 5) et tolérant les basses fréquences ; il repose sur une recherche récursive de segments faits de deux ou plus items (dans une fenêtre de taille paramétrée par l'utilisateur), tolérant ou non la non-contiguïté des items (également paramétrable). L'algorithme calcule la fréquence (*types* et *token*) de chaque séquence et de chaque mot ainsi extrait, et applique le calcul du *log-likelihood* pour déterminer quelle structure sous-jacente rend le mieux compte des réalisations lexico-grammaticales observées : l'algorithme attribue donc une représentation plus abstraite aux séquences extraites, lesquelles sont ensuite fusionnées. L'opération est menée de manière itérative jusqu'à épuisement du processus : plus aucune fusion de segments n'est envisageable. MERGE génère enfin un dictionnaire des formes et segments de toute longueur, discontinus ou non.

46 Concernant la question relative à la mesure de l'orientation définissant le lien entre les constituants d'un segment phraséologiques (point 2), nous renvoyons aux travaux de Luong, Brunet, Longrée, Mayaffre, Mellet, Poudat (2010) et Bonneau (2012). Ces études fondées sur différents corpus (littéraire, politique, scientifique) reposent sur l'établissement de matrices cooccurentielles permettant de définir l'énergie et la disponibilité cooccurentielles d'items, notions définies comme suit :

étant entendu que l'attrait du mot A pour le mot B n'est pas nécessairement symétrique à l'attrait de B pour A [...], nous proposerons de distinguer l'énergie cooccurentielle des mots (la part de ses propres occurrences qu'un mot consacre aux autres – ce que A donne à B, C, D, etc.) et la disponibilité cooccurentielle des mots (la part d'occurrences que chaque mot reçoit des autres – ce que A reçoit de B, C, D, etc.). (Luong, Brunet, Longrée, Mayaffre, Mellet, Poudat 2010 : 322)

47 Enfin, il convient de formuler deux dernières remarques, bien qu'elles dépassent le seul cadre des études phraséologiques : d'une part, les réponses apportées aux points 1 à 5 listés ci-dessus ne restent que des mesures imparfaites si elles ne tiennent pas compte de la dispersion des unités dans le corpus et ses différentes parties (voir à ce sujet Gries 2015b ; Stefanowitsch, Gries 2003). Tenir compte de la dispersion permet de pondérer les effets de fréquence, la force d'association entre items et de repérer les

comportements idiosyncrasiques (Gries 2009) : les travaux d'André Salem sur les séquences textuelles (ou segments répétés) dans le discours syndical est à cet égard exemplaire puisque « la présence dans chacun des volets du corpus » (2006 : 843) est intégrée au paramétrage précédant l'extraction des séquences. Plus qu'un algorithme, c'est donc ici un point de méthode sur lequel Stephen Gries met en garde les chercheurs ; rappelons tout de même que cela est depuis longtemps intégré dans la communauté francophone d'analyse de données textuelles, puisque dès les années 1980, on relève de tels commentaires chez Lafon (1984) ou Serant, Thoiron (1988) relativement au calcul de la richesse lexicale et de la prise en compte de la localisation topographique des formes répétées²⁹.

- 48 D'autre part, le recours à des tests d'indépendance, dont les tests précédemment mentionnés sont représentatifs, entre en contradiction avec la nature même des données langagières qu'elles proposent d'évaluer. En effet, les signes linguistiques (simples ou complexes) non seulement se définissent par des relations de nature syntagmatique et paradigmaticque, mais leur actualisation résulte également de facteurs imbriqués relatifs au canal (écrit/oral), au genre de discours, au registre, aux rapports de place entre interactants et instances impliqués... C'est ce qui pousse S. Gries à préconiser un rapprochement des méthodes statistiques utilisées en linguistique de corpus de celles employées en psycholinguistique (par ex. analyse de la variance ou tests ANOVA).

4. Corpus 17 : synthèse et présentation des contributions

- 49 Dans le cadre de la réflexion actuelle et féconde sur les unités phraséologiques, c'est donc une **perspective explicitement méthodologique, informatique (informatisée) et textuelle** que nous souhaitons développer à travers les cinq contributions rassemblées ici. Le recours à un appareillage informatique et/ou statistique est central : il est au fondement du repérage des unités phraséologiques étudiées – **segments répétés, jets textuels, schèmes lexico-grammaticaux, routines discursives, collocations, constructions syntaxiques spécifiques, locutions figées** ; c'est également lui qui permet de déterminer les emplacements textuels pertinents pour la saisie des segments ; enfin, les logiciels et la statistique accompagnent, guident l'exploration micro et macro-textuelle visant à mettre au jour la trame dans laquelle les segments s'insèrent et qu'ils contribuent à tisser. À cet égard, cette livraison de *Corpus* propose d'embrasser, à des degrés divers, l'ensemble des thématiques chères à la revue – à savoir la réflexion par/sur les corpus (transversale à tous les numéros depuis celui inaugural en 2002, et plus spécifiquement centrée sur le multilingue en 2014, l'oral et le multimodal en 2015), le questionnement des outils pour l'interrogation et l'enrichissement des données (la distance intertextuelle en 2003, l'annotation en 2007, la cooccurrence en 2012), l'adoption d'une posture disciplinaire pour questionner les corpus et évaluer leurs apports (la phonologie en 2004, la lexicométrie/logométrie en 2005, la stylistique en 2006, la syntaxe en 2010, la dialectologie en 2013, la linguistique de corpus et/ou contrastive en 2014).
- 50 Cette livraison de *Corpus*, par son centrage sur ce que nous avons appelé *segments phraséologiques*, met au cœur des analyses, des questions liées à l'usage des méthodes quantitatives et statistiques jugées encore confidentielles « dans les études sur l'analyse

de discours et la phraséologie en français » (Sitri, Tutin 2016 : 6)³⁰, bien que largement établies dans le paysage anglo-saxon issu de la *corpus linguistics* notamment³¹.

4.1 Présentation générale du recueil

- 51 Sont représentées dans ce recueil diverses **méthodes de repérage** éprouvées ailleurs, et notamment recensées dans le numéro 53 de la revue *Lidil* (Sitri, Tutin 2016 : 12-13). Ainsi, la mise au jour de segments phraséologiques peut-elle se faire par la saisie d'items via leur **co-apparition linéaire** (collocation, concordancier, segments répétés) **ou non** (cooccurents) ou leur appartenance à des schémas de **dépendance syntaxique** (arbres lexico-syntaxiques récurrents). Les contributions rassemblées tendent à articuler ces différentes méthodologies. Toutes font usage de **corpus monolingues** (français) représentatifs de genres discursifs distincts : contes du XVII^e siècle, rapports éducatifs, directives de l'Union européenne et transpositions françaises, textes scientifiques ; ces corpus sont **richement annotés**, *a minima* d'informations grammaticales et morphosyntaxiques (lemmes, catégories grammaticales, traits morphosyntaxiques, informations sémantiques), *a maxima* d'informations relatives aux chaînes de coréférence et aux fonctions rhétorico-discursives de certains items.
- 52 Le recueil s'inscrit dans le cadre d'études croisant, dans des proportions différentes, **approches inductives** (*corpus-driven*) et **déductives** (*corpus-based*). L'approche inductive est guidée par des calculs statistiques mettant au jour des listes de segments qu'il faut ensuite contrôler par une lecture *a posteriori* et par des retours au texte (permettant d'éliminer les faux résultats causés par une graphie erronée, un problème typographique ou lié à la casse, une mauvaise lemmatisation, la présence de signes de ponctuation, la présence d'entités nommées...), par effets de seuillage (sur la fréquence, la co-fréquence, la distance entre items, le nombre d'items, le nombre d'associations retenues, leur dispersion jugée trop homogène ou hétérogène – selon les objectifs poursuivis – dans les parties du corpus...), par calculs statistiques enfin³². L'approche déductive, elle, est conduite par la formulation de requêtes d'items préalablement identifiés comme intéressants, sur la base de la connaissance que le-la chercheur-e a de son corpus, des travaux déjà menés ailleurs, du recours à des indexes (lexicaux, morphosyntaxiques...) hiérarchiques de fréquence. Le repérage déductif de segments phraséologiques n'est cependant, on le sait, pas dépourvu d'une étape de filtrage par le-la chercheur-e.
- 53 Les corpus explorés sont l'occasion de **travailler la définition du segment phraséologique** soumis à l'analyse ou ayant émergé de celle-ci, et chemin faisant d'**interroger les méthodes et outils** servant leur repérage. Sont ainsi abordés les aspects suivants : quel sens donner à une fréquence d'emploi et quel lien poser avec la notion de représentativité (contribution de Cislaru, Olive) ? Comment saisir le caractère multiniveau de certains segments phraséologiques, comment tenir compte de la possible non-contiguïté des items, comment déterminer la taille de l'empan à extraire (contributions de Née, Sitri, Veniard, Fleury, et de Kraif, Tutin) ? Comment, dans des contextes d'emploi très similaires, mettre au jour des segments différents ou bien des segments formellement identiques mais aux fonctions discursives différentes (contributions de Gledhill, Patin, Zimina) ? Chaque contribution propose des moyens différents pour **visualiser les segments phraséologiques** (listes, tableaux, arbres de dépendance hiérarchique, surlignage dans le texte). Enfin, si toutes s'interrogent sur le

rôle des segments dans et pour la **cohésion du texte**, les contributions de Boré, Malrieu, de Gledhill, Patin, Zimina, et de Née, Sitri, Veniard, Fleury identifient plus spécifiquement leur fonction en lien avec des **séquences textuelles** : introduire un discours direct, décrire une situation, exprimer une recommandation, évaluer une situation.

4.2 Présentation détaillée des contributions

- 54 Ce recueil s'ouvre sur la contribution de **Georgeta Cislaru et Thierry Olive** qui présente les premiers résultats d'une recherche en cours portant sur « les unités langagières produites de manière spontanée entre deux pauses [et nommées] **jets textuels** » (Cislaru, Olive *ici même*) : ces jets, qui nous placent **en production** (au cours du processus d'écriture), sont confrontés à d'autres unités du texte, parmi lesquelles les **segments répétés** obtenus « par coupures dans la dimension syntagmatique d'un texte déjà existant » (Salem 1987 cité dans Cislaru, Olive *ibid.*), et les cadres collocationnels, qui nous situent tous deux **en réception**. L'étude s'appuie sur un corpus de dossiers de rapports éducatifs que les auteurs ont enregistré en temps réel, au moyen de l'outil de suivi de rédaction *Inputlog*, dans ses diverses phases d'écriture (évolution des états successifs et opérations de révision). Si la confrontation des jets aux segments ne permet pas de poser un parallèle entre eux (le jet ne serait pas sujet au degré de figement que connaissent connecteurs et clichés génériques), leur mise en lien avec les cadres collocationnels paraît plus fructueuse : ainsi émerge un **patron coordinatif** ({X et Y}, {X mais Y}) ainsi qu'une **routine scripturale** organisée autour de connecteurs ou d'adverbes de phrase. Le parcours des auteurs les amène à revenir sur la **démarche même de segmentation textuelle** et sur les implications théoriques et méthodologiques complexes qui la sous-tendent. Ainsi soulèvent-ils par exemple la question du lien entre *mémorisation* et *automatisme* (d'écriture), entre *figement* et *préconstruit*, entre *répétition* et *fréquence*. Ce sont ainsi quelques notions au fondement de la textualité et des faits phraséologiques qui sont envisagées.
- 55 La seconde contribution, de **Catherine Boré et Denise Malrieu**, s'intéresse aux **séquences de discours représentés dans les genres narratifs**, et en particulier aux empan au discours direct (DD) dans un corpus équilibré de **contes du XVII^e siècle** (*Le Pigeon et la Colombe* de Mme D'Aulnoy, et *Histoires ou Contes du temps passé* de Charles Perrault). La présentation autonymique de la parole « se caractéris[ant] par une rupture énonciative dont les propriétés – modalité énonciative, deixis, temps verbaux, entre autres – se différencient en fonction du type de discours narratorial » (Boré, Malrieu *ici même*), les auteures entendent, à travers une **description fine des scènes énonciatives** (primaire – discours narrateur/narrataire ; seconde – discours représentés), mettre au jour des « structures prototypiques du segment introducteur de DD » (*ibid.*) repérées à travers des configurations différentielles de traits morphosyntaxiques palliant l'absence de marques de ponctuation (selon les normes en usage du XVII^e siècle). Segments introducteurs de discours directs et incises sont également pris en considération. L'enrichissement du corpus allie **annotations manuelle** (discours narratorial, discours représentés, phénomènes d'enchâssement) et **automatique** (traits morphosyntaxiques injectés par l'étiqueteur *TreeTagger*) dans des balises conformes au langage TEI ; l'exploration du corpus est conduite au moyen du **logiciel TXM** et tire profit de la souplesse et de la finesse du langage CQL pour formuler

des requêtes précises permettant d'accéder à différents niveaux de profondeur de DR. Les résultats sont consultés dans le **concordancier**, lequel permet d'observer des récurrences et des tendances lourdes en lien avec les changements de modalité énonciative (allocutive, exclamative, interrogative...), les temps verbaux, les catégories sémantiques des verbes (parole, gestuelle), la ponctuation faible, la position de l'incise dans la séquence au DD, les types de propositions (causative, relative, conjonctive). **Un segment introducteur du DD se fait jour** par la combinaison relativement souple de ces items appartenant à différents niveaux de la description linguistique (lexicale, morphosyntaxique, énonciative), où la présence d'aucun item n'est strictement obligatoire.

- 56 La troisième contribution est proposée par **Christopher Gledhill, Stéphane Patin et Maria Zimina** ; elle inscrit l'étude des faits phraséologiques dans le contexte d'une linguistique de corpus appliquée aux langues de spécialité, et influencée par la grammaire systémique fonctionnelle de Michael A.K. Halliday. En particulier, les auteurs étudient les **schèmes lexico-grammaticaux** construits autour des verbes *procéder, effectuer et réaliser*, au sein d'un **corpus comparable de textes juridiques** – rassemblant les directives de l'Union européenne et leurs transpositions dans le droit français – réuni dans le cadre du projet international *Observatoire sur l'eurolecte*. L'enjeu est triple : (i) mettre au jour les divergences linguistiques, textuelles ou discursives des volets du corpus pourtant d'évidence très similaires ; (ii) appliquer la méthode textométrique à l'étude de segments dont la réalisation est intrinsèquement variable ; (iii) confirmer l'hypothèse selon laquelle un schème lexico-grammatical, même s'il est employé dans des contextes similaires, peut jouer des fonctions discursives et rhétoriques distinctes, notamment en lien avec sa localisation dans les sections du texte. Pour ce faire, les auteurs ont recours au **logiciel de textométrie Le Trameur**, et à un outil intégré, la **Lecture Textométrique Différentielle** qui fournit « un affichage synchrone des résultats de [l']analyse textométrique parallèle [des textes comparés] sous forme de surlignage au fil du texte » (Gledhill, Patin, Zimina, *ici même*) ; chemin faisant, ils croisent les résultats obtenus après **calculs des segments répétés, des spécificités, des cooccurrences spécifiques et des poly-cooccurrences** pour repérer et caractériser les schèmes étudiés, et proposent une **visualisation graphique de ces résultats** par, notamment, des **graphes de cooccurrents** en étoile et sur des **trames textuelles** parallèles surlignées. Enfin, les **fonctions discursives** de chaque schème (description de conditions, expression de recommandations légales ou techniques...) sont discutées en tenant compte de leur **localisation dans les sections des textes** (Préambule, Dispositifs, Annexes).
- 57 La contribution d'Émilie Née, Frédérique Sitri, Marie Veniard et Serge Fleury s'intéresse à la **routine discursive**, notion que les auteurs développent depuis quelques années et considèrent comme « un nouvel observable pour l'analyse du discours » (Née, Sitri, Vénard 2016). Moins que la définition ou la caractérisation de segments, c'est ici de la question de l'**association de patrons partiellement figés** et de la question de ses **déterminations textuelles et/ou discursives** qu'il s'agit. Autrement dit, les auteurs cherchent à mettre au jour des « enchaînements entre des routines, ou entre des routines et d'autres indices linguistiques » qui façonneraient la trame textuelle. Leur **démarche est inductive** et repose sur l'exploration d'un **corpus de rapports éducatifs** recueillis dans le domaine de la protection de l'enfance (faisant partie du projet ANR Écritures) ; ce **corpus est enrichi** d'informations

morphosyntaxiques et sémantiques (connecteurs argumentatifs, temporels, aspectuels ; marques de représentation du discours autre), dont l'annotation est opérée semi-automatiquement, en associant à *TreeTagger* un traitement manuel via l'interface personnalisée disponible dans le logiciel *Le Trameur*. Au terme d'un parcours articulant lecture de **concordanciers enrichis**, **calculs des cooccurents** et visualisation des résultats au moyen des fonctions disponibles (**Trame et carte des sections**), une configuration d'items se fait jour de sorte à mettre en évidence l'existence d'une **séquence évaluative**, « structurée autour d'un balancement entre généralisation et particularisation », « à mettre en relation avec l'activité et la pratique professionnelle d'évaluation qui est au cœur du métier d'éducateur spécialisé » (Née, Sitri, Veniard, Fleury *ici même*).

- 58 La dernière contribution de ce recueil proposée par **Agnès Tutin et Olivier Kraif**, procède à une comparaison quantitative et qualitative de deux méthodes d'identification des **motifs textuels**, la méthode des **segments répétés** et la méthode par **arbres lexico-syntaxiques** récurrents. Cette dernière permet de dépasser le critère de linéarité inhérent à la définition actuelle de motif (Longrée, Mellet (2013) ; Legallois, Charnois, Poibeau (2016)) et au mode de saisie traditionnel (concordancier, segments répétés...) : les arbres, en effet, reposent sur une annotation des données textuelles en dépendances hiérarchiques, qui permet de saisir des réalisations discontinues et non strictement ordonnées. Il importe alors de déterminer si le repérage des motifs s'en trouve amélioré. Pour ce faire, les auteurs soumettent à des traitements textométriques le **corpus Termith**, un **corpus arboré d'articles scientifiques** (réunis dans les projets Scientext et Termith) ; en particulier, les analyses portent sur les **verbes de communication**. Elles exploitent notamment les fonctionnalités du **Lexicoscope**, outil en ligne qui permet « d'extraire, pour un pivot (ou mot-pôle) donné, un tableau enregistrant, dans un espace de relation syntaxique déterminé [...] ses collocatifs syntaxiques les plus significatifs [ainsi que s]es cooccurents significatifs » (Tutin, Kraif *ici même*). Le **degré d'association** entre items est calculé au moyen du **log-likelihood ratio** et le lexicogramme (*i. e.*, nom donné au tableau associant le mot-pôle et ses collocatifs) rend compte de la **dispersion de la cooccurrence**. Pour chaque séquence lexico-grammaticale extraite du corpus, les auteurs déterminent à quel type de phraséologie elle appartient – **routines discursives**, **collocations**, **constructions syntaxiques spécifiques**, **locutions figées**, **items non pertinents** – et comparent la pertinence des résultats et la proportion de bruit. L'étude montre que la méthode des arbres paraît plus adaptée au repérage des routines et des collocations quand celle des segments répétés semble plus efficace pour les constructions et les locutions, ce qui permet aux auteurs de revenir sur les propriétés de certains de ces segments phraséologiques ainsi que sur les limites posées par l'annotation et les techniques logicielles utilisées.

BIBLIOGRAPHIE

- Adam J.-M. (2008). *La linguistique textuelle. Introduction à l'analyse textuelle des discours* (2^e édition). Paris : Armand Colin.
- Adam J.-M. (2011a). « Les consécutives intensives : un schéma syntaxique commun à plusieurs genres de discours », *Linx*, 64-65, 115-131.
- Adam J.-M. (dir.) (2015). *Faire texte. Frontières textuelles et opérations de textualisation*. Besançon : Annales Littéraires de l'Université de Franche-Comté.
- Anscombre J.-C. (2011). « Figement, idiomatité et matrices lexicales », in J.-C. Anscombre, S. Mejri (éd.), *Le figement linguistique : la parole entravée*, Paris : Ed. Champion, 17-40.
- Bécue M., Peiro R. (1993). « Les quasi-segments pour une classification automatique des réponses ouvertes », *JADT 1993, Montpellier*. Paris, École nationale supérieure des télécommunications.
- Bindinelli M. (2014). « Détection et analyse de séquences argumentatives. Modaux et semi-modaux dans les débats présidentiels américains », séminaire invitée du laboratoire FoReLL, Université de Poitiers.
- Benigno V., Kraif O., Grossmann F., Velez A. (2016). « La notion de collocation fondamentale : une étude de corpus », *Cahiers de lexicologie*, 108, 125-146.
- Biber D. (1988). *Variation across speech and writing*, Cambridge : Cambridge University Press.
- Biber D. (2006). *University Language : a Corpus-based Study of Spoken and Written Registers*. Amsterdam/Philadelphia : John Benjamins.
- Biber D., Johansson S., Leech G., Conrad S., Finegan E. (1999). *The Longman grammar of spoken and written English*. Harlow : Pearson Education Limited.
- Biber D., Conrad S., Cortes V. (2004). « If you look at...: Lexical Bundles in University Teaching and Textbooks », *Applied Linguistics*, 25(3), 371-405.
- Bonneau J. (2012). « La cooccurrence asymétrique : propriétés quantitatives en disponibilité et énergie », in A. Dister, D. Longrée, G. Purnelle (éd.), *Actes des 11^e Journées internationales d'Analyse statistique des Données Textuelles*. Liège : Université de Liège, 189-202.
- Bouveret M., Legallois D. (2012). *Constructions in French*. Amsterdam, Philadelphia : John Benjamin.
- Brin-Henry F., Knittel M.-L. (2016). « Étude lexicosémantique du nom difficulté(s) dans les comptes rendus de bilan orthophonique : apports structuraux et conceptuels », *Lidil*, 53, consulté sans pagination.
- Brunet É. (2006). « Le corpus conçu comme une boule », in C. Duteil-Mougel et B. Foulquié, *Corpus en Lettres et Sciences sociales – Des documents numériques à l'interprétation*, Actes issu du Colloque international et école d'été d'Albi, 10-14 juillet 2006, 69-78. Consultable en ligne : <http://www.revue-texto.net/Parutions/Livres-E/Albi-2006/Brunet.pdf>.
- Brunet É. (2007). « Fréquences et séquences. Mise en œuvre dans Hyperbase », *Lexicometrica* [en ligne], 7, « Topographie et topologie textuelles », <http://lexicometrica.univ-paris3.fr/numspeciaux/special9/brunet.pdf>.
- Brunet É. (2014). « La répétition dans la phrase. Étude statistique », *Semen*, 38, 15-33.

- Carter-Thomas S. (2014). « Valeurs et fonctions des éléments initiaux commentaires. Analyse contrastive d'un corpus d'articles de recherche en économie », *Corpus* [en ligne], 13, <https://corpus.revues.org/2455>.
- Carter-Thomas S., Prévost S. (2014). *Éléments initiaux dans la phrase : approches inter-genres et inter-langues*, *Corpus* [en ligne], 13, <https://corpus.revues.org/2424>.
- Cheng W. (2007). « Concragramming : A Corpus-Driven Approach to Learning the Phraseology of Discipline-Specific Texts », *CORELL : Computer Resources for Language Learning*, 1, 22-35.
- Cislaru G., Sitri F., Pugnière-Saavedra F. (2013). « Figement et configuration textuelle : les segments de discours répétés dans les rapports éducatifs », in C. Bolly, L. Degand (éd.), *Across the Line of Speech and Writing Variation. Corpora and Language in Use*, Louvain-la-Neuve : Presses universitaires de Louvain, 165-183.
- Fillmore C. (1988). « The mechanisms of construction grammar », *Berkeley Linguistic Society*, 14, 35-55.
- Firth J. R. (1957). *Papers in Linguistics (1934-1951)*, Oxford : Oxford University Press.
- Firth J. R. (1959 [1968]). « The treatment of language in general linguistics' », in F. R. Palmer (éd.), *Selected papers of J.R. Firth (1952-59)*. London, Bloomington : Longman and Indiana University Press, 206-209.
- Fleury S., Lefevre F., Pirès M. (2012). « Étude syntaxique, discursive, lexicométrique et sociolinguistique du mot *quoi* dans le Corpus du français parlé parisien des années 2000 », in Y. Grinschpun et J. Nyee-Doggen (éd.), *Regards croisés sur la langue française : usages, pratiques, histoire. Mélanges en l'honneur de Sonia Branca-Rosoff*, p. 97-112, 2012. Consulté en ligne : <https://halshs.archives-ouvertes.fr/halshs-01143334v2/document>.
- Fontini F., Boukhaled A., Ganascia J.-G. (2015). « Analyse et extraction des motifs linguistiques dans un corpus théâtral », Communication aux Journées d'étude CONSCILA « Pour une grammaire topologique des genres et des styles » du 16 janvier 2015.
- Gledhill C. (2000). « The Discourse Functions of Collocation in Research Article Introduction », *English for Specific Purposes*, 19, 115-135.
- Gries S. Th., (2008). « Phraseology and linguistic theory: a brief survey », in S. Granger, F. Meunier (éd.), *Phraseology : an interdisciplinary perspective*. Amsterdam, Philadelphia : John Benjamins, 3-25. Consultable en ligne : http://www.linguistics.ucsb.edu/faculty/stgries/research/2008_STG_PhraseologyLingTheory_Phraseology.pdf.
- Gries S. Th., (2015a). « Quantitative designs and statistical techniques », in D. Biber, R. Reppen (éd.), *The Cambridge Handbook of English Corpus Linguistics*. Cambridge : Cambridge University Press, 50-71. Consultable en ligne : http://www.linguistics.ucsb.edu/faculty/stgries/research/2015_STG_QuantDesAndMethCorpLing_CUPHb.pdf.
- Gries S. Th., (2015b). « 50-something years of work on collocations : what is or should be next... », in S. Hoffmann, B. Fischer-Starcke, A. Sand (éd.), *Current issues in phraseology*. Amsterdam & Philadelphia : John Benjamins, 135-164, [réédition d'un article paru en 2013]. Consultable en ligne : http://www.linguistics.ucsb.edu/faculty/stgries/research/2015_STG_DeltaP&H_CurrIssInPhraseol.pdf.
- Gries S. Th., (2015c). « Some current quantitative problems in corpus linguistics and a sketch of some solutions », *Language and Linguistics* 16(1). 93-117. Consultable en ligne : http://www.linguistics.ucsb.edu/faculty/stgries/research/2015_STG_QuantProblCorpLing&Solutions_LL.pdf.

- Halliday M. A. K. (1961). « Categories of the theory of grammar », *Word*, 17(3), 241-92.
- Hausmann F. J., Blumenthal P. (2006). « Présentation : collocations, corpus, Dictionnaires », *Langue française*, 150, 7-13.
- Ho-Dac L.-M., Péry-Woodley M.-P. (2008). « Méthodologie exploratoire outillée pour l'étude de l'organisation du discours », in J. Durand, B. Habert, B. Laks (éd.), *Actes du Congrès Mondial de Linguistique française*. Paris : Institut de Langue Française, 737-748. Consultable en ligne : <http://dx.doi.org/10.1051/cmlf08155>.
- Hoey M. (2005). *Lexical priming : A new theory of words and language*. Londres : Routledge.
- Hunston S., Francis G. (2000). *Pattern Grammar : a Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam, Philadelphia : John Benjamin.
- Kraif O. (2016). « Le lexicoscope : un outil d'extraction des séquences phraséologiques basé sur des corpus annotés », *Cahiers de lexicologie*, 108, 91-106.
- Kraif O., Tutin A. (2016). *Phraséologie et linguistique appliquée*, *Cahiers de lexicologie*, 108. Paris : Classiques Garnier.
- Kraif O., Novakova I., Sorba J. (2016). « Constructions lexico-syntaxiques spécifiques dans le roman policier et la science-fiction », *Lidil*, 53, consulté sans pagination.
- Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Genève, Paris : Slatkine Champion.
- Lamalle C., Salem A. (2002). « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels », *Actes des 6^e Journées internationales d'Analyse statistique des Données Textuelles*. Saint Malo : Université de Rennes 1, 13-15 mars, 403-411.
- Leblanc J.-M. (2016). « Phraséologie et formules rituelles dans le discours politique, l'expérimentation en lexicométrie », *Lidil*, 53, consulté sans pagination.
- Legallois D. (s. d). « La grammaire d'une langue peut-elle s'apprendre à partir de ses unités phraséologiques ? », Consultable en ligne : http://www.crisco.unicaen.fr/IMG/pdf/Legallois_la_grammaire_d_une_langue.pdf.
- Legallois D., Tutin A. (coord.) (2013a). *Vers une extension du domaine de la phraséologie*, *Langages*, 189. Paris : Larousse.
- Legallois D., Tutin A. (2013b). « Présentation : Vers une extension du domaine de la phraséologie », *Langages*, 189, 3-25.
- Legallois D., Charnois T., Poibeau T. (2016). « Repérer les clichés dans les romans sentimentaux grâce à la méthode des "motifs" », *Lidil*, 53, consulté sans pagination.
- Léon J. (2004). « Lexies, synapsies, synthèmes : le renouveau des études lexicales en France au début des années 1960 », in G. Hassler (éd.), *History of Linguistics in Texts and Concepts - Geschichte der Sprachwissenschaft in Texten und Konzeptionen*, Münster : Nodus Publikationen, 405-418.
- Léon J. (2007). « From Linguistic Events and Restricted Languages to Registers. Firthian legacy and Corpus Linguistics », *The Henry Sweet Society Bulletin*, 49, 5-26.
- Longrée D., Mellet S. (2013). « Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours », *Langages*, 189, 65-79.
- Longrée D., Mellet S. (2015). « Dépasser l'opposition « paradigmatic » / « syntagmatic » : pour une grammaire topologique des genres et des styles. Application à la littérature latine classique », Communication aux Journées d'étude CONSCILA « Pour une grammaire topologique des genres et des styles » du 16 janvier 2015.

- Luong X., Brunet É., Longrée D., Mayaffre D., Mellet S., Poudat C. (2010). « La cooccurrence, une relation asy-métrique ? », in S. Bolasco, I. Chiari, L. Giuliano (éd.). *Actes des 10^e Journées internationales d'Analyse statistique des Données Textuelles*. Rome : Università la Sapienza, 321-331.
- Martinez W. (2012). « Au-delà de la cooccurrence binaire... Poly-cooccurrences et trames de cooccurrence », *Corpus* [en ligne], 11. Consulté le 13 novembre 2016, <http://corpus.revues.org/2262>.
- Mayaffre D. (2007). « L'analyse des données textuelles aujourd'hui : du corpus comme une urne au corpus comme un plan. Retour sur les travaux actuels de topographie/topologie textuelle », *Lexicometrica* 7 « Topographie et topologie textuelles ». Consultable en ligne : <http://lexicometrica.univ-paris3.fr/numspeciaux/special9/mayaffre.pdf>.
- Mayaffre D., Viprey J.-M. (2012a). *La cooccurrence, du fait statistique au fait textuel*, *Corpus*, 11. Nice : Université Nice Sophia-Antipolis, UMR Bases, Corpus, Langage.
- Mayaffre D., Viprey J.-M. (2012b). « Présentation », *Corpus*, 11, 7-19.
- Mellet S. (coord.) (2007). « Topologie et topographie textuelles », *Lexicometrica* 7. Numéro consultable en ligne : <http://lexicometrica.univ-paris3.fr/numspeciaux/special9.htm>.
- Mellet S., Barthélémy J.-P. (2007). « La topologie textuelle : légitimation d'une notion émergente », *Lexicometrica* n° spécial. Consultable en ligne : <http://lexicometrica.univ-paris3.fr/numspeciaux/special9/mellet.pdf>.
- Née É., Sitri F., Veniard M. (2014). « Pour une approche des routines discursives dans les écrits professionnels », Communication donnée lors du Congrès Mondial de Linguistique Française, Juillet 2014, Berlin, Allemagne. Consultable en ligne : <http://dx.doi.org/10.1051/shsconf/20140801195>.
- Née É., Sitri F., Veniard M. (2016). « Les routines, une catégorie pour l'analyse des discours : le cas des rapports éducatifs », *Lidil*, 53, consulté sans pagination.
- O'Donnell M. B., Scott M., Mahlberg M., Hoey M. (2012). « Exploring text-initial words, clusters and congrams in a newspaper corpus », *Corpus Linguistics and Linguistic Theory*, 8(1), 73-101.
- Philippe G., Adam J.-M. (2015). « Continuité et textualité », in J.-M. Adam, *Faire texte. Frontières textuelles et opérations de textualisation*. Besançon : Annales littéraires de l'Université de Franche-Comté, 35-80.
- Quiniou S., Cellier P., Charnois T., Legallois D. (2012). « What About Sequential Data Mining Techniques to Identify Linguistic Patterns for stylistics ? », Communication à *International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'12)*. New Delhi, India. 166-177. Consultable en ligne : <https://hal.archives-ouvertes.fr/hal-00675578>.
- Rastier F. (1996). « Pour une sémantique des textes : questions d'épistémologie », extrait de F. Rastier (éd.). *Sens et Textes*. Paris : Didier, 9-35. Version disponible dans *Texte ! Textes & Cultures* [en ligne], http://www.revue-texto.net/Inedits/Rastier/Rastier_PourSdT.html.
- Rastier F. (2001). « Éléments de théorie des genres », *Texte ! Textes & Cultures* [en ligne], <http://www.revue-texto.net/index.php?id=555>.
- Rastier F. (2005). « Enjeux épistémologiques de la linguistique de corpus », in G. Williams (dir.). *La linguistique de Corpus*. Rennes : Presses Universitaires de Rennes, 31-46. Consultable en ligne : http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html.
- Römer U. (2010). « Establishing the Phraseological Profile of a Text Type : The Construction of Meaning in Academic Book Reviews », *English Text Construction*, 3(1), 95-119.

- Römer U., O'Donnell M. B. (2009). « Positional variation of phrase-frames in a new corpus of proficient student writing », Communication lors de la journée AACL Edmonton, Canada, 9 octobre. Support consultable en ligne : <https://sites.ualberta.ca/~aACL2009/PDFs/RoemerODonnell2009AACL.pdf>.
- Rouveyrol L. (2014). « Recherche et pratiques pédagogiques en langues de spécialité », *Cahiers de l'Aplut*, XXXIII(3), 53-75.
- Pražuch W. (2016). « La phraséologie du discours politique sur la crise : analyse d'allocutions d'hommes politiques français (2008-2012) », *Cahiers de lexicologie*, 59-74.
- Salem A. (1987). *Pratique des segments répétés. Essai de statistique textuelle*, Paris, Klincksieck.
- Salem A. (2006). « Proximités segmentales », Publication dans les actes des 8^e Journées internationales d'Analyse statistique des Données Textuelles (JADT 2006). Besançon : Université de Franche-Comté, 19-21 avril, 839-849.
- Schmale G. (2013). « Qu'est-ce qui est préfabriqué dans la langue ? - Réflexions au sujet d'une définition élargie de la préformation langagière », *Langages*, 189, 27-45.
- Serant D., Thoirion P. (1988). « Topographie des formes répétées », *Revue Informatique et Statistique dans les Sciences humaines*, XXIV, 1-4, 333-343.
- Siepmann D. (2016). « Lexicologie et phraséologie du roman contemporain : quelques pistes pour le français et l'anglais », *Cahiers de lexicologie*, 108, 21-42.
- Sinclair J. (1991). *Corpus, Concordance, Collocation : Describing English language*. Oxford : Oxford University Press.
- Sitri F., Tutin A. (2016a). *Phraséologie et genres du discours : patrons, motifs, routines*, *Lidil*, 53. Grenoble : Éditions littéraires et linguistiques de l'Université de Grenoble (ellug).
- Sitri F., Tutin A. (2016b). « Introduction », *Lidil*, 53, consulté sans pagination.
- Spina S., Tanganelli E. (2012). « Les collocations comme indice pour distinguer les genres textuels », *Corpus* [en ligne], 11. <http://corpus.revues.org/2219>.
- Stefanowitsch A., Gries S. Th. (2003). « Collostructions : Investigating the interaction of words and constructions », *International Journal of Corpus Linguistics*, 8(2), 209-243.
- Stefanowitsch A., Gries S. Th. (2009). « Corpora and grammar », in L. Anke, K. Merja (éd.). *Corpus linguistics : an international handbook*, vol. 2, 933-951. Berlin & New York : Mouton de Gruyter. Consultable en ligne : http://www.linguistics.ucsb.edu/faculty/stgries/research/2009_AS-STG_CorporaGrammar_HSK.pdf.
- Stubbs M. (2002). « Two quantitative methods of studying phraseology in English », *International journal of Corpus Linguistics*, 7(2), 215-44.
- Tutin A., Kraif O. (2016). « Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines : l'apport des arbres lexico-syntaxiques récurrents », *Lidil*, 53, consulté sans pagination.
- Viprey J.-M. (2006). « Structure non-séquentielle des textes », *Langages*, 163, 71-85.
- Wahl A., Gries S. Th. (à par.). « Multi-word expressions : A novel computational approach to their bottom-up statistical extraction », in P. Cantos (éd.). *Lexical collocation analysis : advances and applications*, Berlin, New York : Springer. Consultable en ligne : http://www.linguistics.ucsb.edu/faculty/stgries/research/Subm_AW_STG_MWEs-MERGE&AFL.pdf.
- Williams G. (2006). « La linguistique et le corpus : une affaire prépositionnelle », in C. Duteil-Mougel et B. Foulquié, *Corpus en Lettres et Sciences sociales – Des documents numériques à*

l'interprétation, Actes issu du Colloque international et école d'été d'Albi, 10-14 juillet 2006, 163-170. Consultable en ligne : <http://www.revue-texto.net/Parutions/Livres-E/Albi-2006/Williams.pdf>.

Wray A. (2002). *Formulaic Language and the Lexicon*, Cambridge : Cambridge University Press.

Zimina M. (2005). « Exploration textométrique de corpus de traduction », *Meta : journal des traducteurs* [en ligne], 50(4), <http://id.erudit.org/iderudit/019925ar>.

Zimina M. (2014). « Bi-text : a systematic functional approach and textometric analysis », Communication au 25^e *European Systemic Functional Linguistics Conference*, Paris : Université Paris-Diderot, 10-12 juillet.

Logiciels et plateformes cités

Hyperbase - une version portail est accessible à l'adresse suivante : <http://hyperbase.unice.fr/> ; la version pour poste fixe est payante (informations à l'adresse suivante : <http://bcl.cnrs.fr/article69>) mais une version d'essai peut être téléchargée sur le site <http://ancilla.unice.fr/>.

Le Trameur - peut être téléchargé librement (pour Windows et MacOSX) sur le site officiel <http://www.tal.univ-paris3.fr/trameur/> (documentation également disponible).

Lexico3 - peut être téléchargé librement (pour Windows) sur le site officiel <http://lexi-co.com/> (documentation également disponible).

Lexicoscope - le portail est accessible à l'adresse suivante : <http://phraseotext.u-grenoble3.fr/lexicoscope/>.

TXM - peut être téléchargé librement (pour version Windos, Mac, Linux - 32 et 64bits) sur le site officiel : <http://textometrie.ens-lyon.fr/> (documentation également disponible) ; une version « portail » est également accessible ici : <http://portal.textometrie.org/>.

ANNEXES

Annexe 1

Grille d'analyse synthétique de quelques faits phraséologiques

Tableau 2. Tableau synthétique des critères définitoires appliqués à certaines catégories d'unités phraséologiques

<i>Faits phraséologiques</i> <i>Critères définitoires¹</i>	CCOL	CLG	COL	CON	CP	FE	LB	MT	MWE	NG	PF	PT	POL	QSR	RD	SR ST
le fait phraséologique est constitué de deux unités uniquement (✓) ou de deux unités et plus (+)	+	+	✓	+	+	✓	+	✓	+	+	+	+	+	+	+	+
les items appartiennent à différents niveaux linguistiques	✓	✓	✗	✓	?	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗
le fait phraséologique tolère la variation morphosyntaxique en son sein	✓	✓	✗	✓	✗	✗	✓	✓	✗	✗	✓	✓	✗	✓	✓	✗
le fait phraséologique peut tolérer l'existence de trous en son sein	✓	✗	✗	✓	✗	✗	✗	✓	✓	✗	✓	?	✗	✓	✓	✗
la définition du fait phraséologique repose sur un critère de fréquence	✓	✓	✓	+	✗	✗	✓	✓	✓	✓		✓	✓	✓	✓	✓

¹ Il a été tenté d'insérer à ce tableau les critères relatifs à la compositionnalité et/ou à la transparence des faits phraséologiques, ainsi qu'aux mécanismes langagiers, discursifs ou hybrides les rendant possibles : il s'est avéré difficile de synthétiser en une case cette information, un même terme pouvant recouvrir des unités analysables de points de vue différents selon, souvent, que le regard soit porté en synchronie ou en diachronie, et selon l'ancrage théorique du-de la chercheur-e faisant usage de la catégorie.

Légende

CCOL = cadres collocationnels ; **CLG** = colligations ; **COL** = collocations ; **CON** = constructions ; **CP** = constructions préformées ; **FE** = *formulaic expressions* ; **LB** = *lexical bundles* ; **MT** = motifs ; **MWE** = *multi-word expressions* ; **NG** = n-grammes ; **PF** = *phrase-frame* ; **PT** = *patterns* ; **POL** = poly-cooccurents ; **QSR** = quasi-segments répétés ; **RD** = routine discursive ; **SR/ST** = segments répétés / séquence textuelle

En l'absence de mentions explicites, quand le critère définitoire est validé, la case indique une encoche ; quand il ne l'est pas, une croix. Le point d'interrogation ? signifie que le trait ne peut être appliqué de manière systématique à tous les faits phraséologiques désignés.

Annexe 2

La recherche sur les faits phraséologiques : quelques initiatives recensées (2015-2017)

Journées d'étude, colloques, ateliers, écoles d'été

2015 : JE « Les constructions comme unités fondamentales de la langue » organisées par Dominique Legallois organisé à Caen ; JE « Pour une grammaire topologique des genres et des styles », organisé à Paris par le CONSCILA ; Col. « Marqueur et structures, quelles articulations dans la (re)construction du sens ? » organisé à Paris Ouest Nanterre la

Défense ; Col. « PHRASÉOLOGIE : descriptions linguistiques et traitement informatique » organisé à Puebla (Mexique) ;

2016 : Col : « Les séquences figées : des propriétés linguistiques à l'enseignement en FLE » organisé par Marko Vidak à Brest ; EE : 2^e école d'été l'action COST PARSEME (PARSing and Multi-word Expressions), organisée à La Rochelle ;

2017 : Session thématique « Phrases préfabriquées : sémantique, syntaxe, pragmatique et prosodie » du colloque « Représentations du sens linguistique VII » organisé à Sherbrook (Canada) ; Atel. « The 13th Workshop on Multiword Expressions (MWE 2017) » dans le cadre du colloque de *European Chapter of the Association for Computational Linguistics*, organisé à Valencia (Espagne) ; Col : « La phraséologie française : débats théoriques et dimensions appliquées (didactique, traduction et traitement informatique) », organisé à l'Université d'Artois.

Publications

2015 : Gries (2015b), « 50-something years of work on collocations : what is or should be next... », in S. Hoffmann, B. Fischer-Starcke, A. Sand (éd.). *Current issues in phraseology*. Amsterdam & Philadelphia : John Benjamins, 135-164.

2016 : Sitri, Tutin (coord.), *Phraséologie et genres de discours : patrons, motifs, routines, Lidil*, 53 ; Kraif, Tutin, *Phraséologie et linguistique appliquée, Cahiers de lexicologie*, 108.

2017 : Legallois (coord.), *Les constructions comme unités fondamentales de la langue : illustrations, évaluations, critiques* ; Muryn, Novakova, *De la phraséologie aux genres textuels : état des recherches et perspectives méthodologiques, Synergies Pologne*, 14 ; Wahl, Gries, « Multi-word expressions : A novel computational approach to their bottom-up statistical extraction », in P. Cantos (éd.), *Lexical collocation analysis : advances and applications*. Berlin & New York : Springer.

NOTES

1. Nous tenons à vivement remercier Virginie Lethier et Sylvie Mellet pour leur relecture attentive.
2. Nous proposons en annexe 1 une comparaison-synthèse des propriétés définitoires de quinze de ces faits phraséologiques : cadres collocationnels, colligations, collocations, constructions, *formulaic expressions*, *lexical bundles*, motifs, *multi-word expressions*, n-grams, *phrase-frame*, *patterns*, poly-cooccurents, quasi-segments répétés, routines discursives, segments répétés.
3. « $p(\text{word}_1|\text{word}_2)$ is not the same as $p(\text{word}_2|\text{word}_1)$, just compare $p(\text{offin spite})$ to $p(\text{in spite of})$ ». Cette formulation est avancée par S. Gries pour interroger la fiabilité des mesures d'association utilisées en linguistique de corpus.
4. Ces mesures statistiques, que nous citerons plus avant, permettent de définir (sans intervention du chercheur, bien qu'à partir d'un filtre ou d'un algorithme décidé par lui) la longueur et la fréquence minimale et maximale servant à repérer les unités phraséologiques, ainsi que l'orientation unissant les constituants.
5. Nous inversons ainsi le regard porté ailleurs sur la notion de cooccurrence - ainsi Damon Mayaffre et Jean-Marie Viprey mentionnent-ils tenir compte des études phraséologiques « pour l'éclairage qu'elles apportent aux questions spécifiques de la cooccurrence » (2012 : 10), facteur de la textualité - et rejoignons la perspective décrite dans Née, Sitri, Veniard (2016) qui voient

dans le repérage des segments « une voie d'entrée dans le corpus, une mise en avant de points saillants du discours » (2016 : sp) et donc des textes.

6. Le continuum rend ainsi bien compte de l'existence de deux pôles où la phraséologie est envisagée à travers une conception restreinte ou élargie : voir à ce sujet les introductions des numéros 189 de *Langages* (Legallois, Tutin 2013b) et 53 de *Lidil* (Sitri, Tutin 2016).

7. « Conventionalized, memorized, multi-word sequences represent an important component in modern language sciences research (...). Being able to identify them automatically, using the explosion of corpus resources that are ever more available, is an increasingly important goal for researchers in various disciplines ».

8. Cette courte synthèse s'appuie sur les références suivantes : Léon (2004), Hausmann, Blumenthal (2006), Williams (2006), Legallois, Tutin (2013b), Gries (2008).

9. « Unité fondamentale » est ainsi à prendre en ces deux sens : les faits phraséologiques sont au fondement de la description des productions et en constituent un élément essentiel (quantitativement prégnant). Sur ce dernier point, Alexander Wahl et Stephen Th. Gries font état de travaux estimant que les faits phraséologiques représentent entre un tiers et la moitié des schèmes lexicaux-grammaticaux relevés, ce qui tend à montrer que la proportion d'unités préformées et mémorisées est conséquente (Wahl, Gries à par. : 2).

10. « [t]he normal carrier of meaning is the phrase and not the word ; the word is the limiting case of the phrase, and has no other status in the description of meaning [...]. One of the great strengths of a phraseological approach is the preservation of the integrity of text much longer than alternative approaches to description, and in turn this entails the preservation of meaning ». On retrouve ici une conception traditionnelle des unités en analyse de données textuelles (lexicométrie, logométrie ou textométrie) : Damon Mayaffre rappelle que « le mot est une unité graphique, le plus souvent ambiguë, sans sens explicite, pas même dotée de signification. Le segment [répété], lui, devient une unité linguistique porteuse de sens. La forme « classe » n'a pas de sens, « classe ouvrière » en est doté » (2007 : 9).

11. La notion de *restricted language* théorise le fait que toute situation de communication porte avec elle un ensemble de facteurs extralinguistiques pesant sur les choix langagiers (lexicaux et syntaxiques essentiellement) disponibles aux locuteurs engagés dans l'échange. Cf. Léon (2007).

12. Ils avaient déjà montré leur pertinence pour l'attribution d'auteur ; voir à ce sujet Brunet (2006).

13. Voir par exemple à ce sujet Adam (2008 : chapitre 5) et Viprey (2006).

14. « Phraseology is the ideal point of contact between a corpus and a description, because it accepts surface phenomena, and this, initially, is what a corpus provides; no pre-processing is required, no abstractions, no information such as parts of speech added ».

15. Un même *pattern* lexico-grammatical peut se réaliser par diverses collocations : on citera l'exemple de Legallois, Charnois, Poibeau (2016 : sp) au sujet des motifs du roman sentimental : « prit une profonde respiration, poussa un profond soupir, peuvent être considérés comme des réalisations différentes d'un même schéma ».

16. Par exemple, dans ce recueil : pour un étiquetage morphosyntaxique, *TreeTagger* (disponible pour de très nombreuses langues) ; pour un étiquetage des dépendances syntaxiques (inspiré de la syntaxe de Lucien Tesnière), le parseur *XIP*.

17. C'est là une évolution majeure de ces dix dernières années, Ute Römer déclarant en 2010 : « As Stubbs (2007 : 181) rightly notes, "there is no purely automatic way of identifying phrasal units of meaning" » (Römer 2010 : 100).

18. Agrawal Rakesh, Srikant Ramakrishnan, (1995), « Mining sequential patterns », *Proc. of ICDE'95*, 3-14.

19. « The theory of Lexical Priming (Hoey 2004, 2005, 2010) introduces textual position as yet another associative feature. Hoey (2005) labels this phenomenon 'textual colligation', suggesting that "every word is primed to occur in, or avoid, certain positions within the discourse" (p. 13) ».

20. Elle est également envisagée comme perspective de recherche pour les motifs dans Legallois, Charnois, Poibeau (2016).
21. Ainsi, dans son corpus de comptes-rendus d'ouvrages scientifiques, U. Römer distingue quatre portions équilibrées (nombre de mots similaire) jouant un rôle dans l'organisation textuelle (section 1 - introduction, sections 2 et 3 - résumé, section 4 - évaluation) ; O'Donnell, Scott, Mahlberg, Hoey (2012), dans un corpus d'articles de presse, annotent spécifiquement chaque phrase selon qu'elle soit en position non-initiale ou à l'initial d'un titre ou d'un sous-titre, de paragraphes suivant ces titre et sous-titre, de paragraphes constitués de deux phrases ou moins.
22. Ou « regroupements d'unités opérés sur la base de leur identité ou de leurs ressemblances » (Lamalle, Salem 2002 : 404).
23. Ou « divisions [...] en pages, en tranches de 100 mots, en paragraphes, en phrases ou encore en fenêtres de longueur fixe se limitant à quelques mots à droite et à gauche d'un pôle retenu. La division en phrases, peut-être réalisée approximativement par la sélection d'un certains nombre de caractères que l'on dotera du statut de délimiteurs de phrases. Cependant, le paragraphe que l'on peut souvent identifier au caractère retour-chariot dans les textes encodés sur support lisible par un ordinateur constitue une division privilégiée du texte, à la fois relativement fiable [...] et présent[ant] dans de nombreux cas une certaine régularité du point de vue de la taille des ensembles qu'elle découpe » (*ibid.* : 405-406).
24. Gries (2015b) mentionne également d'autres tests, à notre sens moins connus ou cités : le *Minimum Sensitivity* et le calcul dit *simple conditional probability*. Le premier ne permet pas de déterminer l'orientation du rapport entre les deux items constitutifs de l'unité phraséologique, le deuxième s'avère pertinent dans le cas des seules unités phraséologiques asymétriques, bien qu'il génère cependant un taux élevé d'erreurs (cf. Gries 2015b : 139-140).
25. Dans la communauté francophone, la tendance grandissante consistant à recourir à des tests statistiques différents du χ^2 , dont l'usage est largement répandue dans la communauté francophone (en particulier en analyse de données textuelles et linguistique de corpus), à s'essayer ce faisant à des mesures peut-être plus confidentielles, se fait sous l'impulsion de chercheurs issus du traitement automatique des langues : Thierry Charnois (lab. LIPN), Lydia-Mai Ho-Dac (lab. CLLE-ERSS), Olivier Kraif (Lab. Lidilem), Thierry Poibeau (lab. LaTTiCe).
26. « Of the measures that have been developed, some have emerged as more popular than others. For example, mutual information (MI) is probably the most well-known association measure ».
27. « Nearly all [association measures] reflect the mutual association of word₁ and word₂ to each other (...). However, what all these measures do *not* distinguish is whether word₁ is more predictive of word₂ or the other way around ».
28. « One approach towards the same goal is Gries and Mukherjee's (2010 : Section 2.2) implementation of lexical gravity G, which also leads to the notion of lexical stickiness – the degree to which words like to occur in n-grams (cf. Sinclair's 1991 idiom principle) rather than on their own (cf. Sinclair's 1991 open-choice principle) ».
29. Ainsi, Daniel Serant et Philippe Thoiron de dire : « [...] on peut se demander si les informations concernant le seul plan quantitatif de la répétitivité sont suffisantes. Ne doit-on pas discriminer entre deux textes ayant la même distribution de fréquences [...] mais dont les items répétés seraient distribués différemment au long du texte ? » (1988 : 1-2).
30. Nous partageons ce constat bien que des tentatives pour favoriser le recours aux corpus et aux outils d'exploration existent depuis longtemps : Franz Josef Hausmann et Peter Blumenthal ne disaient-il pas dans l'introduction du numéro 150 de la revue *Langue française* que, bien que « [t]ous les contributeurs ne comprennent pas le terme de « collocation », point de mire de [leurs] réflexions, exactement de la même façon, [...] tous sont convaincus de l'extrême utilité des

corpus et des concordances pour donner une description réaliste de la dimension combinatoire des mots et de la langue » (2006 : 11).

31. Cette affirmation doit toutefois être nuancée au regard du constat de Stephen Th. Gries tiré notamment des travaux de Michael Stubbs : « For example, Stubbs (2002) and Stubbs and Barth (2003) largely ignore the immensely interesting work that has been done concerning the automatic or semi-automatic identification of multi-word units (...) » (2008 : 15-16) ; sa remarque s'étend à la plupart des travaux en linguistique de corpus et plus encore en linguistique cognitive. En retour, on pensera à pondérer ce positionnement : l'article de Biber, Conrad, Cortes (2004) ne mentionne-t-il pas les travaux novateurs d'André Salem sur l'exploration des unités poly-lexicales guidée par leur fréquence d'apparition (p. 373).

32. À titre d'illustration en dehors de ce volume, et en raison de sa méthodologie présentée de manière très didactique, on pourra se reporter à Benigno, Kraif, Grossmann, Velez (2016), ou avec une approche plus anglo-saxonne, Römer (2010).

AUTEUR

MARION BENDINELLI

M.C.F. en sciences du langage

Université Bourgogne Franche-Comté

E.A. ELLIADD/Discours, Texte, Espace Public, Société